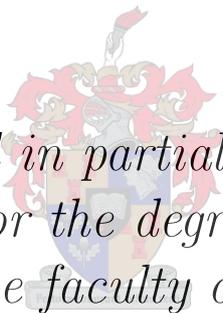


PCA and CVA biplots: A study of their underlying theory and quality measures

by

Hilmarié Brand



*Thesis presented in partial fulfillment of the
requirements for the degree of Master of
Commerce in the faculty of Economic and
Management Sciences*

at

Stellenbosch University

Supervisor: Prof. N.J. Le Roux

Co-supervisor: Prof. S Lubbe

Date: March 2013

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2013

Abstract

The main topics of study in this thesis are the Principal Component Analysis (PCA) and Canonical Variate Analysis (CVA) biplots, with the primary focus falling on the quality measures associated with these biplots. A detailed study of different routes along which PCA and CVA can be derived precedes the study of the PCA biplot and CVA biplot respectively. Different perspectives on PCA and CVA highlight different aspects of the theory that underlie PCA and CVA biplots respectively and so contribute to a more solid understanding of these biplots and their interpretation. PCA is studied via the routes followed by Pearson (1901) and Hotelling (1933). CVA is studied from the perspectives of Linear Discriminant Analysis, Canonical Correlation Analysis as well as a two-step approach introduced in Gower *et al.* (2011). The close relationship between CVA and Multivariate Analysis of Variance (MANOVA) also receives some attention.

An explanation of the construction of the PCA biplot is provided subsequent to the study of PCA. Thereafter follows an in depth investigation of quality measures of the PCA biplot as well as the relationships between these quality measures. Specific attention is given to the effect of standardisation on the PCA biplot and its quality measures.

Following the study of CVA is an explanation of the construction of the weighted CVA biplot as well as two different unweighted CVA biplots based on the two-step approach to CVA. Specific attention is given to the effect of accounting for group sizes in the construction of the CVA biplot on the representation of the group structure underlying a data set. It was found that larger groups tend to be better separated from other groups in the weighted CVA biplot than in the corresponding unweighted CVA biplots. Similarly it was found that smaller groups tend to be separated to a greater extent from other groups in the unweighted CVA biplots than in the corresponding weighted CVA biplot.

A detailed investigation of previously defined quality measures of the CVA biplot follows the study of the CVA biplot. It was found that the accuracy with which the group centroids of larger groups are approximated in the weighted CVA biplot is usually higher than that in the corresponding unweighted CVA biplots. Three new quality measures that assess that accuracy of the Pythagorean distances in the CVA biplot are also defined. These quality measures assess the accuracy of the Pythagorean distances between the group centroids, the Pythagorean distances between the individual samples and the Pythagorean distances between the individual samples and group centroids in the CVA biplot respectively.

Opsomming

Die hoofonderwerpe van studie in hierdie tesis is die Hoofkomponent Analise (HKA) bistipping asook die Kanoniese Veranderlike Analise (KVA) bistipping met die primêre fokus op die kwaliteitsmaatstawwe wat daarmee geassosieer word. 'n Gedetailleerde studie van verskillende roetes waarlangs HKA en KVA afgelei kan word, gaan die studie van die HKA en KVA bistippings respektiewelik vooraf. Verskillende perspektiewe op HKA en KVA belig verskillende aspekte van die teorie wat onderliggend is tot die HKA en KVA bistippings respektiewelik en dra sodoende by tot 'n meer breedvoerige begrip van hierdie bistippings en hulle interpretasies. HKA word bestudeer volgens die roetes wat gevolg is deur Pearson (1901) en Hotelling (1933). KVA word bestudeer vanuit die perspektiewe van Linieêre Diskriminantanalise, Kanoniese Korrelasie-analise sowel as 'n twee-stap-benadering soos voorgestel in Gower *et al.* (2011). Die noue verwantskap tussen KVA en Meerveranderlike Analise van Variansie (MANOVA) kry ook aandag.

'n Verduideliking van die konstruksie van die HKA bistipping word voorsien na afloop van die studie van HKA. Daarna volg 'n indiepte-ondersoek van die HKA bistipping kwaliteitsmaatstawwe sowel as die onderlinge verhoudings tussen hierdie kwaliteitsmaatstawwe. Spesifieke aandag word gegee aan die effek van die standaardisasie op die HKA bistipping en sy kwaliteitsmaatstawwe.

Opvolgend op die studie van KVA is 'n verduideliking van die konstruksie van die geweegde KVA bistipping sowel as twee verskillende ongeweegde KVA bistippings gebaseer op die twee-stap-benadering tot KVA. Spesifieke aandag word gegee aan die effek wat die inagneming van die groepsgroottes in die konstruksie van die KVA bistipping op die voorstelling van die groepstruktuur onderliggend aan 'n datastel het. Daar is gevind dat groter groepe beter geskei is van ander groepe in die geweegde KVA bistipping as in die oorstemmende ongeweegde KVA bistipping. Soortgelyk daaraan is gevind dat kleiner groepe tot 'n groter mate geskei is van ander groepe in die ongeweegde KVA bistipping as in die oorstemmende geweegde KVA bistipping.

'n Gedetailleerde ondersoek van voorheen gedefinieerde kwaliteitsmaatstawwe van die KVA bistipping volg op die studie van die KVA bistipping. Daar is gevind dat die akkuraatheid waarmee die groeps-gemiddeldes van groter groepe benader word in die geweegde KVA bistipping, gewoonlik hoër is as in die ooreenstemmende ongeweegde KVA bistippings. Drie nuwe kwaliteitsmaatstawwe wat die akkuraatheid van die Pythagoras-afstande in die KVA bistipping meet, word gedefinieer. Hierdie kwaliteitsmaatstawwe beskryf onderskeidelik die akkuraatheid van die voorstelling van die Pythagoras-afstande tussen die groeps-gemiddeldes, die Pythagoras-afstande tussen die individuele observasies en die Pythagoras-afstande tussen die individuele observasies en groeps-gemiddeldes in die KVA bistipping.

Acknowledgements

I wish to express my gratitude to my promoter, Prof. N.J. Le Roux, for his guidance, patience and encouragement throughout this study.

I wish to thank my co-supervisor, Prof. S Lubbe, for her support throughout this study.

I wish to thank the National Research Foundation without whose financial support I would not have been able to complete this study.

I wish to thank my husband, A Beelders, my parents, P.J. and E Brand, and my dear friends V Williams and W Cloete, without whose love, support and encouragement I would not have been able to complete this study.

I wish to thank everybody at SACEMA (South African Centre for Epidemiological Modelling and Analysis) who has supported me throughout this study, in particular Prof. A Welte, Prof. J Hargrove and Dr. A.G. Hitchcock.

SOLI DEO GLORIA

Contents

Contents	i
List of Figures	vi
List of Tables	ix
1 Introduction	1
1.1 Objectives	3
1.2 The scope of this thesis	4
1.3 Notation	6
Scalars, vectors and matrices	6
1.3.1 Scalars, vectors and matrices	6
1.3.2 Vector spaces	9
1.4 Definitions and terminology	9
1.5 Abbreviations	10
1.6 Some needed linear algebra results	10
1.6.1 The spectral decomposition (eigen-decomposition) of a symmetric matrix	11
1.6.2 The square root matrix of a positive definite (p.d) matrix	12
1.6.3 Singular values and singular vectors	13
1.6.4 The singular value decomposition (svd) of a matrix	14
1.6.5 Expressing a matrix of a given rank as the inner product of two matrices of the same rank	17
1.6.6 Generalised inverses	18
1.6.7 Projection	19
1.6.7.1 Projection onto an affine subspace	23
1.6.8 The principal axis theorem	24
1.6.9 Huygens' principle	25
1.6.10 The Eckart-Young theorem	27
1.6.11 The best fitting r -dimensional affine subspace to a configuration of points in higher dimensional space	29
1.6.12 The Two-Sided Eigenvalue Problem	31
1.6.13 The generalised svd (The svd in a metric other than \mathbf{I})	36
1.6.14 The generalised Eckart-Young theorem (The Eckart-Young theorem in a metric other than \mathbf{I})	37

2	PCA and the PCA biplot	40
2.1	Introduction	40
2.2	Deriving PCA	40
2.2.1	Pearson's approach to PCA	41
2.2.2	Hotelling's approach to PCA	47
2.3	Principal components with zero and/or equal variances	60
2.4	Interpretation of the coefficients of the principal components	61
2.5	The number of principal components to retain	63
2.6	The traditional (classical) biplot	64
2.7	The biplot proposed by Gower and Hand (1996)	79
2.7.1	The construction of the PCA biplot	81
2.7.1.1	Interpolation and the interpolative biplot	81
2.7.1.2	Prediction and the predictive PCA biplot	84
2.7.1.3	The relationship between prediction and multivariate regression analysis	89
2.8	Data structured into groups	90
2.9	Summary	93
3	PCA biplot quality measures	95
3.1	Orthogonality properties underlying a PCA biplot	95
3.2	The overall quality of the PCA biplot	99
3.3	Adequacies	106
3.3.1	Definition and properties	106
3.3.2	Visual representation	110
3.4	Predictivities	112
3.4.1	Axis predictivities	112
3.4.1.1	Definition and properties	112
3.4.1.2	The relationship between the axis predictivity and adequacy of a biplot axis	115
3.4.1.3	The relationship between the axis predictivities and the overall quality	123
3.4.1.4	The relationship of the axis predictivities with the overall quality when the PCA biplot is constructed from the standardised measurements	126
3.4.1.5	Axis predictivities and the interpretation of the PCA biplot	129
3.4.1.6	The scale dependence of the PCA biplot, overall quality, axis predictivities and adequacies: an illustrative example	132
3.4.1.7	Changing the PCA biplot scaffolding axes	135
3.4.2	Sample predictivities	138
3.4.2.1	Definition and properties	138
3.4.2.2	Using sample predictivities to detect outliers	141
3.4.2.3	The relationship between sample predictivities and the overall quality	143
3.5	Summary	146

4	CVA and the CVA biplot	148
4.1	Introduction	148
4.2	CVA is equivalent to LDA for the multi-group case	149
4.2.1	Weighted CVA	149
4.2.1.1	Discrimination using weighted CVA	149
4.2.1.2	Classification using weighted CVA	165
4.2.2	Unweighted CVA	174
4.2.2.1	Discrimination using unweighted CVA	174
4.2.2.2	Classification using unweighted CVA	178
4.2.3	The connection between weighted and unweighted CVA	179
4.2.4	The scale invariance of CVA	180
4.2.5	Important hypotheses to test prior to performing CVA	183
4.3	Deriving CVA as a special case of Canonical Correlation Analysis (CCA)	194
4.3.1	Canonical Correlation Analysis (CCA)	194
4.3.2	CVA as a special case of CCA	202
4.4	CVA as a two-step procedure	205
4.5	The CVA biplot	225
4.5.1	Interpolation	228
4.5.2	Prediction	229
4.6	The scale invariance of the CVA biplot	236
4.7	A comparison between a CVA biplot and a PCA biplot	240
4.8	The effect of accounting for the group sizes in the CVA biplot	243
4.9	Summary	249
4.10	Appendix	251
4.10.1	The derivation of the result in Section 4.2	251
4.10.2	The derivation of the result in Section 4.4	253
5	Quality of the CVA biplot	257
5.1	Orthogonality properties underlying a CVA biplot	257
5.2	The overall quality of the CVA biplot	259
5.2.1	The overall quality of the CVA biplot with respect to the canonical variables	259
5.2.1.1	Definition and properties	259
5.2.1.2	Scale invariance	261
5.2.2	The overall quality of the CVA biplot with respect to the original variables	262
5.2.2.1	Definition and properties	262
5.2.2.2	Scale dependence	264
5.3	Adequacies	265
5.3.1	Definition and properties	265
5.3.2	Scale invariance	268
5.4	Axis predictivities	268
5.4.1	Definition and properties	268
5.4.2	The relationship of the axis predictivities with the overall quality with respect to the original variables	270

5.4.3	Scale invariance	271
5.5	Group predictivities	272
5.5.1	Definition and properties	272
5.5.2	Group predictivities and the accuracy of distances represented in the CVA biplot	279
5.5.3	The effect of accounting for the group sizes in the construction of the CVA biplot on the group predictivities	280
5.5.4	The relationship between group predictivities and the overall quality with respect to the canonical variables	283
5.5.5	Scale invariance	284
5.6	Group contrast predictivities	285
5.6.1	Definition and Properties	285
5.6.2	Scale invariance	289
5.7	Axis predictivities, group predictivities and group contrast predictiv- ities: an illustrative example	289
5.8	Within-group sample predictivities	294
5.8.1	Definition and properties	294
5.8.2	Within-group sample predictivities and the accuracy of dis- tances represented in the CVA biplot.	299
5.8.3	Scale invariance	301
5.8.4	Within-group sample predictivities of ‘new’ samples	301
5.9	The overall within-group sample predictivity associated with a group	303
5.9.1	Definition and properties	303
5.9.2	Scale Invariance	304
5.10	Mixed contrast predictivities	305
5.10.1	Definition and Properties	305
5.10.2	Scale invariance	307
5.11	Sample predictivities	308
5.11.1	Definition and properties	308
5.11.2	Sample predictivities and the accuracy of distances represented in the CVA biplot	314
5.11.3	Scale invariance	315
5.11.4	The overall sample predictivity associated with a group	315
5.11.4.1	Definition and properties	315
5.11.4.2	Scale Invariance	319
5.11.5	The total sample predictivity associated with a data set	320
5.11.5.1	Definition and properties	320
5.11.5.2	Scale Invariance	321
5.11.6	Sample predictivities measures of ‘new’ samples	321
5.12	Sample contrast predictivities	323
5.12.1	Definition and Properties	323
5.12.2	Scale invariance	325
5.13	Within-group axis predictivities	326
5.13.1	Definition and properties	326
5.13.2	Scale invariance	331

5.13.3	The relationship between axis predictivities and within-group axis predictivities	332
5.14	Changing the CVA biplot scaffolding axes	335
5.15	Summary	336
6	Conclusion	339
6.1	What has been achieved in this thesis?	339
6.2	The way forward	340
6.2.1	The robust PCA biplot	341
6.2.2	Variable Selection	341
6.3	To conclude...	342
	Bibliography	343

List of Figures

2.1	<i>The top ten percentages (Top10) and graduation rates (Grad) (empty circles) of the 25 universities of the University data set along with the best fitting straight line (solid line) and the approximated data points (solid circles)). The two dashed lines illustrates the orthogonal projection of the data points corresponding to the 17th and 25th universities of the University data set onto the best fitting straight line to the two-dimensional configuration of points.</i>	45
2.2	<i>The two-dimensional traditional PCA biplot (i.e. $\alpha = 1$) constructed from the standardised measurements of the University data set.</i>	73
2.3	<i>The two-dimensional traditional biplot constructed from the standardised measurements of the University data set with $\alpha = 0$.</i>	78
2.4	<i>The two-dimensional predictive PCA biplot constructed from the standardised measurements of the University data set.</i>	80
2.5	<i>The two-dimensional predictive correlation biplot constructed from the standardised measurements of the University data set.</i>	81
2.6	<i>The two-dimensional interpolative biplot constructed from the standardised measurements of the University data set, illustrating the vector-sum approach for Purdue University.</i>	84
2.7	<i>The two-dimensional predictive PCA biplot constructed from the standardised measurements of the University data set.</i>	88
2.8	<i>(a) The two-dimensional predictive PCA biplot of the Ocotea data set with 95% bags constructed for <i>O. bullata</i> and <i>O. porosa</i> and a convex hull constructed for <i>O. kenyensis</i>; (b) The two-dimensional predictive PCA biplot of the Ocotea data set with 50% bags constructed for <i>O. bullata</i> and <i>O. porosa</i> and a convex hull constructed for <i>O. kenyensis</i>.</i>	91
3.1	<i>Left: An orthogonal decomposition of a vector; Right: A non-orthogonal decomposition of a vector.</i>	96
3.2	<i>The scree plot corresponding to the (standardised) University data set. . .</i>	103
3.3	<i>The overall quality of the PCA biplot of the University data set, constructed from the standardised data, corresponding to each possible dimensionality of the PCA biplot.</i>	104
3.4	<i>A unit circle in the two-dimensional PCA biplot space of the standardised University data set that is centred at the origin together with the projections of the six-dimensional unit vectors, $\{\mathbf{e}_k\}$, onto the biplot space.</i>	110

3.5	(a) The two-dimensional interpolative PCA biplot of the University data set with thick lines the relative lengths of which represents the relative magnitudes of the adequacies of the measured variables; (b) A small section of the interpolative PCA biplot in (a).	111
3.6	The overall quality and axis predictivities of the PCA biplot constructed from the standardised measurements of the University data set.	128
3.7	The two-dimensional predictive PCA biplot constructed from the standardised measurements of the National Track data set.	131
3.8	(a) The two-dimensional PCA biplot constructed from the first two principal components of the standardised simulated data set; (b) The two-dimensional PCA biplot constructed from the first and third principal components of the standardised simulated data set.	137
3.9	The two-dimensional PCA biplot constructed from the last two principal components of the standardised National Track data set.	143
3.10	The two-dimensional PCA biplot constructed from the standardised measurements of the University data set.	146
4.1	The two-dimensional CVA display of the simulated data set.	165
4.2	The two-dimensional unweighted CVA display of the simulated data set.	177
4.3	The two-dimensional unweighted CVA display of the simulated data set constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	224
4.4	(a) The two-dimensional predictive unweighted CVA biplot of the simulated data set constructed with $\mathbf{C} = \mathbf{I}$; (b) The two-dimensional predictive unweighted CVA biplot of the simulated data set constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$	235
4.5	The two-dimensional predictive unweighted CVA biplot of the simulated data set constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$	236
4.6	(a) and (c) The two-dimensional predictive PCA biplot constructed from the standardised measurements of the Ocotea data set; (b) and (d) The two-dimensional predictive CVA biplot of the Ocotea data set. In (a) and (b) 95% bags are superimposed for the species <i>O. bullata</i> and <i>O. porosa</i> while a convex hull is constructed for the specie <i>O. kenyensis</i> . In (c) and (d) 50% bags are superimposed for the species <i>O. bullata</i> and <i>O. porosa</i> while a convex hull is constructed for the specie <i>O. kenyensis</i>	242
4.7	(a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the first simulated data set; (b) The two-dimensional weighted CVA biplot of the first simulated data set.	244
4.8	(a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the second simulated data set; (b) The two-dimensional weighted CVA biplot of the second simulated data set.	245
4.9	(a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the third simulated data set; (b) The two-dimensional weighted CVA biplot of the third simulated data set.	245
4.10	(a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the fourth simulated data set; (b) The two-dimensional weighted CVA biplot of the fourth simulated data set.	246

4.11	(a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the fifth simulated data set; (b) The two-dimensional weighted CVA biplot of the fifth simulated data set.	247
4.12	(a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the sixth simulated data set; (b) The two-dimensional weighted CVA biplot of the sixth simulated data set.	248
4.13	(a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the seventh simulated data set; (b) The two-dimensional weighted CVA biplot of the seventh simulated data set.	248
5.1	The two-dimensional (predictive) unweighted (with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$) CVA biplot of the seventh simulated data set showing the group centroid (asterisk) and 50% bag for each of the four groups.	290
5.2	Two-dimensional weighted predictive CVA biplot of the race data set.	330

List of Tables

2.1	<i>The standard deviations of the measured variables of the University data set.</i>	73
2.2	<i>The sample correlation matrix associated with the University data set. . .</i>	78
2.3	<i>The predictions of the measurements of the University of California, Berkeley (UCBerkeley), and Purdue University (Purdue) produced by the two-dimensional predictive PCA biplot constructed from the standardised measurements of the University data set.</i>	88
2.4	<i>The standard deviations of the measured variables of the Ocotea data set.</i>	91
3.1	<i>The overall quality of the PCA biplot constructed from the standardised measurements of the University data set corresponding to each possible dimensionality of the PCA biplot.</i>	102
3.2	<i>The adequacies of the biplot axes of the two-dimensional PCA biplot constructed from the standardised measurements of the University data set. .</i>	109
3.3	<i>The adequacies and predictivities of the biplot axes representing the six measured variables of the University data set corresponding to all possible dimensionalities of the PCA biplot constructed from the standardised measurements.</i>	121
3.4	<i>The sample correlation matrix associated with the National Track data set.</i>	131
3.5	<i>The axis predictivities corresponding to the two-dimensional PCA biplot constructed from the standardised measurements of the National Track data set.</i>	132
3.6	<i>The standard deviations of the eight measured variables of the National Track data set.</i>	132
3.7	<i>The axis predictivities corresponding to the one-dimensional PCA biplot constructed from the unstandardised measurements of the National Track data set.</i>	132
3.8	<i>The coefficients of the first principal component of the unstandardised national track data set.</i>	133
3.9	<i>The adequacies of the eight biplot axes in the one-dimensional PCA biplot constructed from the unstandardised measurements of the National track data set.</i>	133
3.10	<i>The weights of the axis predictivities in the expression of the overall quality of the PCA biplot constructed from the unstandardised measurements of the National Track data set.</i>	133

3.11	<i>The overall qualities corresponding to the one-dimensional PCA biplots constructed from the unstandardised and standardised measurements of the National Track data respectively.</i>	134
3.12	<i>The coefficients of the first principal component of the standardised national track data set.</i>	134
3.13	<i>The adequacies of the eight biplot axes of the one-dimensional PCA biplot constructed from the standardised measurements of the National Track data set.</i>	134
3.14	<i>The axis predictivities of the eight biplot axes of the one-dimensional PCA biplot constructed from the standardised measurements of the National Track data set.</i>	134
3.15	<i>The sample correlation matrix corresponding to the simulated data set.</i>	136
3.16	<i>The contributions of the principal components (PCs) to the sample variances of the standardised variables of the simulated data set.</i>	136
3.17	<i>(a) The axis predictivities corresponding to the two-dimensional PCA biplot constructed from the first two principal components of the standardised measurements of the simulated data set; (b) The axis predictivities corresponding to the two-dimensional PCA biplot constructed from the first and third principal components of the standardised measurements of the simulated data set.</i>	137
3.18	<i>The individual contributions of the eight principal components to the sample predictivity associated with Greece corresponding to the PCA biplot constructed from the standardised measurements of the National Track data set.</i>	142
3.19	<i>The sample predictivities of Yale University (Yale) University of Chicago (UChicago), University of California, Berkeley (UCBerkeley) and Purdue University (Purdue) corresponding to the PCA biplot of the University data set constructed from the standardised measurements.</i>	145
3.20	<i>The overall qualities of the PCA biplot of the University data set constructed from the standardised measurements.</i>	145
4.1	<i>The (population) group means of the four groups of the simulated data set.</i>	164
4.2	<i>The (population) correlation matrix associated with each of the four five-variate normal distributions from which the samples of the simulated data set were drawn.</i>	164
5.1	<i>The group predictivities of the one-dimensional weighted and unweighted CVA biplots of the third simulated data set.</i>	282
5.2	<i>The group predictivities corresponding to the one-dimensional CVA biplots of the fifth simulated data set.</i>	282
5.3	<i>The group predictivities of the one-dimensional weighted and unweighted CVA biplots of the seventh simulated data set.</i>	283
5.4	<i>The Pythagorean distances between the points representing the group centroids of the seventh simulated data set in the two-dimensional CVA biplot space.</i>	291
5.5	<i>The group predictivities of the two-dimensional unweighted CVA biplot (with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$) of the seventh simulated data set.</i>	291

5.6	<i>The group contrast predictivities corresponding to the two-dimensional unweighted CVA biplot of the seventh simulated data set.</i>	291
5.7	<i>The Pythagorean distances between the points representing the group centroids of the seventh simulated data set in the canonical space. (These distances are proportional to the Mahalanobis distances between the group centroids in the measurement space.)</i>	292
5.8	<i>The axis predictivities of the two-dimensional unweighted CVA biplot (with $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$) of the seventh simulated data set.</i>	292
5.9	<i>The observed group centroids of the seventh simulated data set.</i>	293
5.10	<i>The within-group sample predictivity, sample predictivity and group predictivity of the corresponding group centroid of the fifth, 22nd, 351st and 366th sample of the fourth simulated data set, corresponding to the two-dimensional weighted CVA biplot.</i>	313
5.11	<i>The overall sample predictivities corresponding to the one-dimensional CVA biplots of the fourth simulated data set.</i>	318
5.12	<i>The group predictivities corresponding to the one-dimensional CVA biplots of the fourth simulated data set.</i>	318
5.13	<i>The overall within-group sample predictivities of the one-dimensional CVA biplots of the fourth simulated data set.</i>	318
5.14	<i>The overall sample predictivities corresponding to the one-dimensional CVA biplots of the fifth simulated data set.</i>	319
5.15	<i>The group predictivities corresponding to the one-dimensional CVA biplots of the fifth simulated data set.</i>	319
5.16	<i>The overall within-group sample predictivities of the one-dimensional CVA biplots of the fifth simulated data set.</i>	319
5.17	<i>The total sample predictivity of the third simulated data set associated with the unweighted and weighted CVA biplots constructed from the fourth simulated data set.</i>	323
5.18	<i>The sample within-group correlation matrix associated with the race data set.</i>	331
5.19	<i>The within-group axis predictivities associated with the two-dimensional weighted CVA biplot of the race data set.</i>	331

Chapter 1 - Introduction

Lehmann (1988) defines Statistics as “the enterprise dealing with the collection of data sets, and extraction and presentation of the information they contain”. In the light of this definition it is clear that graphical presentations of a data set form an integral part of any statistical analysis - graphical displays not only present the information contained in the data but can also be used to extract information that is difficult or even impossible to extract by means of traditional parametric multivariate analyses. In the words of Everitt (1994) “there are many patterns and relationships that are easier to discern in graphical displays than by any other data analysis method”. According to Chambers *et al.* (1983) “there is no single statistical tool that is as powerful as a well-chosen graph”.

In most fields of application data is typically multivariate and hence the task of investigating and analysing multivariate data is often faced in practice. The fact that humans can only visualise objects which are at most three dimensional presents the need to reduce the dimensionality of the observed data in some way whenever the dimensionality of the data is greater than three. Unfortunately dimension reduction is always accompanied by loss of information. That is, an observed data set can only be approximated in a space that is of lower dimensionality than the data set.

In order to represent the observed data set as accurately as possible, the lower dimensional display space should be chosen such that the loss of information resulting from the dimension reduction is as small as possible. If the dissimilarity between two measurement vectors is measured by some distance metric, then in order to minimise the loss of information, the lower dimensional display space should be chosen such that it represents the set of distances between the measurement vectors as accurately as possible according to some criterion. Metric multidimensional scaling (MDS) methods are designed for this exact purpose - each metric MDS method is designed to minimise some measure of discrepancy between a set of distances in the full measurement space and the corresponding approximated distances in the lower dimensional display space. The main difference between different metric MDS techniques lie in the distance metric that is used. The distance metric that is used depends on the type of data at hand as well as the specific aspect of the data set that is to be represented as well as possible in the lower dimensional display space. Two metric MDS techniques will be studied in this thesis, namely principal component analysis (PCA) and canonical variate analysis (CVA). In PCA and CVA dissimilarities between measurement vectors are measured using the Pythagorean distance metric and Mahalanobis distance metric respectively.

Even though MDS configurations are optimal in the sense that they represent the set of distances of interest as well as possible according to some criterion, they

lack information regarding the original measured variables. This problem can be addressed by applying biplot methodology to MDS configurations. Biplots were introduced by Gabriel (1971), who also coined the name. A biplot is a joint map of the samples and variables of a data set. Applying biplot methodology to any MDS configuration therefore enhances the informativeness of the lower-dimensional graphical display by adding information regarding the measured variables. The ‘bi’ in ‘biplot’ refers to the fact that two modes, namely samples and variables, are represented simultaneously and not to the dimension of the display space. The biplot proposed by Gabriel is known as the traditional (or classical) biplot. In the traditional biplot each row (sample) and column (variable) of the data matrix under consideration is represented by a vector emanating from the origin. These vectors are such that the inner product of a vector representing a row and a vector representing a column approximates the corresponding element of the data matrix. Gabriel proposed that the rows of the data matrix be represented only by the endpoints of the corresponding vectors so that samples and variables can be easily differentiated in the biplot.

The main weakness of the traditional biplot is that inner products are difficult to visualise. Gower and Hand (1996) addressed this problem by proposing that the (continuous) variables be represented by axes, called biplot axes, which are calibrated such that the approximations to the elements of the data matrix of interest can be read off from the biplot axes by means of orthogonal projection onto the calibrated axes, as is done in the case of ordinary scatter plots. Biplots constructed in this manner can therefore be regarded as multivariate analogues of ordinary scatter plots (Gower and Hand, 1996) and can thus easily be interpreted by both statisticians and non-statisticians. This modern approach to biplots will be followed throughout this thesis. The biplot proposed by Gower and Hand (1996) also allows for the representation of categorical variables - these are represented by simplexes consisting of points called category level points (CLP’s) (Gower and Hand, 1996). Only biplots of data sets consisting of samples measured on continuous variables only will however be discussed in this thesis.

Biplot methodology extends “the mere representation of data to an exploratory analysis in itself by the application of several novel ideas” (Gardner and Le Roux, 2003). Examples of such novel ideas are the addition of alpha-bags (Gardner, 2001) and classification regions to biplots aimed at the optimal discrimination of groups and the classification of samples of unknown origin, like the CVA biplot.

After a data set has been graphically represented by means of a biplot, a natural question to ask is, ‘how accurately does the biplot represent the original higher-dimensional data set?’ as the answer to this question will determine to what extent the relationships and predictions suggested by the biplot are representative of reality. This presents the need for measures of the quality of the different aspects of a biplot.

The main topics that will be studied in this thesis are those of the PCA and CVA biplots, with the primary focus falling on the quality measures of these biplots. Over the last few years much work has been done on PCA and CVA biplots and even more so on the quality measures associated with these biplots (Gardner-Lubbe *et al.* (2008); Gower *et al.* (2011)). Most of the measures that will be discussed in this thesis were proposed in Gardner-Lubbe *et al.* (2008). These quality measures then

received more attention and were extended in Gower *et al.* (2011). In this thesis the existing PCA biplot and CVA biplot quality measures and the relationships between them will be studied in more depth. New quality measures will also be defined for some important aspects of the CVA biplot for which no quality measures have been proposed to date. Furthermore, taking forth the work of Gower *et al.* (2011) on weighted and unweighted CVA biplots, the effect of accounting for group sizes in the construction of the CVA biplot on (1) the representation of the group structure underlying a data set and (2) the quality measures of the CVA biplot, will be investigated in more depth.

A limitation of the currently available literature on PCA and CVA biplots is the little attention paid to the different perspectives from which PCA and CVA can be viewed. Different perspectives on PCA and CVA highlight different aspects of the analyses that underlie PCA and CVA biplots respectively and so contribute to a more solid understanding of these biplots and their interpretation. For this reason a detailed discussion of different routes along which PCA and CVA can be derived will forego the study of the PCA biplot and CVA biplot respectively.

1.1 Objectives

The primary aims of this thesis are to:

1. Study different routes along which PCA can be derived;
2. Investigate the quality measures associated with PCA biplots as well as the relationships between these quality measures;
3. Study different perspectives from which CVA can be viewed;
4. Study the previously defined quality measures associated with CVA biplots as well as the relationships between these quality measures;
5. Investigate the effect of accounting for the (possibly) different group sizes in the construction of the CVA biplot on (a) the representation of the group structure underlying a data set and (b) the quality measures of the CVA biplot using simulated data sets;
6. Define quality measures for aspects of the CVA biplot for which no quality measures have been proposed to date.

The secondary objectives of this thesis are to:

1. Illustrate the differences and similarities between the traditional PCA biplot proposed by Gabriel (1971) and the PCA biplot proposed by Gower and Hand (1996);
2. Demonstrate the effect of standardisation on the PCA biplot and its quality measures;

3. Illustrate the differences and similarities between ordinary MDS CVA displays and the CVA biplot proposed by Gower and Hand (1996);
4. Illustrate the differences between PCA and CVA biplots using existing data sets;

The remainder of this chapter is devoted to an outline of the scope of this thesis which is provided in Section 1.2, a description of the adopted notation, terminology and abbreviations in Sections 1.3 - 1.5 and the discussion of a number of results from linear algebra that will be utilised throughout this study, provided in Section 1.6.

1.2 The scope of this thesis

Chapters 2 and 3 are devoted to Principal Component Analysis (PCA), the PCA biplot and the quality measures of the PCA biplot. The PCA biplot is, as its name indicates, closely related to PCA itself. A solid understanding of PCA is therefore required to understand the construction and interpretation of the PCA biplot. For this reason Chapter 2 commences with a detailed discussion of two of the most well known routes via which PCA can be derived, namely those followed by Pearson (1901) and Hotelling (1933). Pearson focused on the approximation of the data matrix of interest in a lower dimensional affine subspace of the measurement space. More specifically, he searched for the straight line or hyperplane which is best fitting to the higher-dimensional configuration of points with coordinate vectors given by the row vectors of the data matrix in terms of least squares. Hotelling on the other hand derived PCA by searching for uncorrelated linear combinations of the measured variables that account for as much of the total variability associated with the measured vector variable as possible. The remainder of Chapter 2 is devoted to the PCA biplot. Since it is the quality of the PCA biplot as approximation of the data matrix at hand which is of interest in this thesis, PCA will be viewed from Pearson's perspective when discussing the construction and interpretation of the PCA biplot. The traditional PCA biplot proposed by Gabriel (1971) and the PCA biplot proposed by Gower and Hand (1996) are discussed in detail. The study of the PCA biplot is set forth in Chapter 3 which focuses on measures of the quality of different aspects of the PCA biplot.

Chapters 4 and 5 are devoted to Canonical Variate Analysis (CVA) and the CVA biplot. As the name indicates, the CVA biplot is based on the statistical analysis CVA, a dimension reduction technique that is used to analyse data sets consisting of samples structured into a number of predefined distinct groups. CVA aims to (1) optimally discriminate amongst groups and (2) classify objects of unknown origin as accurately as possible. Chapter 4 commences with a detailed study of three different ways in which CVA can be defined, namely as (1) the equivalent to Linear Discriminant Analysis (LDA) for the multi-group case (2) a special case of Canonical Correlation Analysis (CCA) and (3) a two-step approach consisting of a transformation of the measurement vectors and a least squares approximation in the transformed space. Depending on whether the sizes of the groups are taken into account in the analysis or not, CVA is referred to as being weighted or un-

weighted respectively. Accordingly the CVA biplot can also be either weighted or unweighted. The construction of the weighted and two different types of unweighted CVA biplots will be discussed in this chapter. Specific attention will be paid to the effect of taking the group sizes into account in the construction of the CVA biplot on the representation of the group structure in the biplot. The construction of the CVA biplot will be explained from the perspective of the two-step approach to CVA since this approach naturally allows for the construction to be performed very similarly to that of the PCA biplot. Various quality measures associated with CVA biplots are discussed in Chapter 5 - these include quality measures which were defined for the PCA biplot, adjusted so as to make them appropriate for the CVA biplot, as well as a number of 'new' quality measures.

Chapter 6 consists of an outline of what has been achieved in this study and suggestions regarding possible future work.

The figures in this thesis have been constructed, and the reported quality measures calculated, using the programming language R (R Core Team, 2012). Existing functions as well as newly developed functions were utilised. Most of the functions can be found in the R package 'UBbipl', which can be downloaded from the website: www.wiley.com/legacy/wileychi/gower/material. The functions in 'UBbipl' were extended for the calculation of the new CVA biplot quality measures that are defined in Chapter 5.

1.3 Notation

1.3.1 Scalars, vectors and matrices

n	The total number of samples.
J	The number of groups.
n_j	The number of samples in the j th group: $\sum_{j=1}^J n_j = n$.
p	The number of variables.
$\mathbf{a} (k \times 1)$	A general column vector of length k with i th element equal to a_i .
\mathbf{a}'	The transpose of $\mathbf{a} (k \times 1)$. If the dimension of the vector is omitted from the notation, the dimension will be evident from the context.
\mathbf{e}_k	The column vector of which all elements are zero except for the k th element which is equal to one.
$\mathbf{1}$	The column vector all elements of which are equal to one.
$\cos(\theta_{\mathbf{a},\mathbf{b}})$	The cosine of the angle between the vectors \mathbf{a} and \mathbf{b} .
$\cos(\theta_{\mathbf{a}_i,\mathbf{a}_j})$	The cosine of the angle between the vectors \mathbf{a}_i and \mathbf{a}_j .
\underline{x}	A stochastic variable
$\underline{\mathbf{x}} (p \times 1)$	A stochastic $p \times 1$ vector variable. If the dimension of the vector is omitted from the notation, the dimension will be evident from the context.
$\mathbf{A} (m \times k)$	A general matrix with m rows and k columns.
$[\mathbf{A}]_{ik}$	The ik th element of the matrix, \mathbf{A} .
\mathbf{a}'_i	The i th row vector of the matrix, \mathbf{A} .
$\mathbf{a}_{(j)}$	The j th column vector of the matrix, \mathbf{A} .
$\bar{\mathbf{a}}$	The mean vector of the matrix $\mathbf{A} (m \times k)$ i.e. $\bar{\mathbf{a}} = \frac{1}{m} \mathbf{A} \mathbf{1}$.
$\text{diag}(\mathbf{A})$	The vector with i th element equal to $[\mathbf{A}]_{ii}$.

$ \mathbf{A} $	The determinant of the square matrix, \mathbf{A} .
$\text{adj}(\mathbf{A})$	The adjoint of the matrix \mathbf{A} i.e. $\text{adj}(\mathbf{A}) = \mathbf{A}^{-1} \mathbf{A} $.
$\text{tr}(\mathbf{A})$	The trace of the square matrix \mathbf{A} .
$\ \mathbf{A} (m \times k)\ ^2$	The sum of the squared elements of the matrix $\mathbf{A} (m \times k)$ i.e. $\text{tr}(\mathbf{A}\mathbf{A}')$.
$\ \mathbf{a} (k \times 1)\ ^2$	The squared length of the vector \mathbf{a} of length k i.e. $\text{tr}(\mathbf{a}\mathbf{a}') = \mathbf{a}'\mathbf{a}$.
\mathbf{A}_r	If \mathbf{A} is a general $m \times n$ matrix, then \mathbf{A}_r is the submatrix of \mathbf{A} consisting of the first r columns of \mathbf{A} . If \mathbf{A} is a diagonal matrix or a rectangular matrix with only non-zero elements on its main diagonal, then \mathbf{A}_r is the $r \times r$ diagonal submatrix of \mathbf{A} which consists of the first r rows and columns of \mathbf{A} .
\mathbf{I}_p	The $p \times p$ identity matrix.
$\mathbf{A}_{(r)}$	If \mathbf{A} is a general $m \times n$ matrix, then $\mathbf{A}_{(r)}$ is the submatrix of \mathbf{A} consisting of the last r columns of \mathbf{A} .
\mathbf{A}^r	Assuming that \mathbf{A} is an invertible matrix, \mathbf{A}^r is that submatrix of \mathbf{A}^{-1} consisting of the first r rows of \mathbf{A}^{-1} .
$\mathbf{A}^{(r)}$	Assuming that \mathbf{A} is an invertible matrix, $\mathbf{A}^{(r)}$ is that submatrix of $\mathbf{A}^{(r)}$ consisting of the last r rows of \mathbf{A}^{-1} .
$\widehat{\mathbf{A}}_r$	A rank r approximation of \mathbf{A} .
d_{ij}	The Pythagorean distance between the i th and j th sample in the full p -dimensional measurement space.
δ_{ij}	The Pythagorean distance between the i th and j th sample in the lower dimensional display space.
$\mathbf{G} (n \times J)$	An indicator matrix indicating the group membership of n samples, each belonging to one of J groups. The element $[G]_{ij}$ equals one if the i th sample belongs to the j th group and zero otherwise, $i \in [1 : n]$, $j \in [1 : J]$.
$\mathbf{N} (J \times J)$	The diagonal matrix with j th diagonal element given by the size of the j th group, n_j i.e. $\mathbf{N} = \mathbf{G}'\mathbf{G} = \text{diag} \{n_1, n_2, \dots, n_J\}$.

$\bar{\mathbf{X}}$	Given an $n \times p$ matrix \mathbf{X} with i th row vector giving the centred measurements of the i th sample on p measured variables, $i \in [1:n]$, that is \mathbf{X} is centred such that $\mathbf{1}'\mathbf{X} = \mathbf{0}'$, then $\bar{\mathbf{X}}$ is the matrix of group means corresponding to \mathbf{X} i.e. $\bar{\mathbf{X}} = \mathbf{N}^{-1}\mathbf{G}'\mathbf{X}$.
$\bar{\mathbf{x}}^j$	The j th group mean i.e. $\bar{\mathbf{x}}^{j'}$ is the j th row vector of $\bar{\mathbf{X}}$.
\mathbf{B}	The matrix of between-groups sums of squares and cross products.
\mathbf{W}	The matrix of within-group sums of squares and cross products.
Σ	The population covariance matrix.
$\hat{\Sigma}$	The estimated (or sample) covariance matrix.
$\Sigma_{\mathbf{B}}$	The population between-groups covariance matrix.
$\hat{\Sigma}_{\mathbf{B}}$	The estimated (or sample) between-groups covariance matrix.
$\Sigma_{\mathbf{W}}$	The population within-group covariance matrix.
$\hat{\Sigma}_{\mathbf{W}}$	The estimated (or sample) within-group covariance matrix.
$\Sigma_{\mathbf{W}}^j$	The population within-group covariance matrix of the j th group.
$\hat{\Sigma}_{\mathbf{W}}^j$	The estimated (or sample) within-group covariance matrix of the j th group.
$\operatorname{argmax}_{\mathbf{a}} \{f(\mathbf{a})\}$	The argument (vector in this case) which maximises $\{f(\mathbf{a})\}$ over all possible choices of \mathbf{a} .
$i \in [1:n]$	The scalar, i , can take on the value of any integer between 1 and n , including the values 1 and n i.e. $i \in [1:n]$ will be assumed to mean $i \in [1:n]$, $i \in \mathbb{Z}$.

1.3.2 Vector spaces

$\mathcal{V}(\cdot)$	The column space of the matrix argument i.e. the vector space generated by the column vectors of the matrix argument
$\mathcal{V}^\perp(\cdot)$	The orthogonal complement of the column space of the matrix argument.
\mathbb{R}^p	The vector space containing all p - dimensional real vectors.
\mathcal{L}	The lower dimensional display space.
\mathcal{L}^\perp	The orthogonal complement of \mathcal{L} .

1.4 Definitions and terminology

Euclidean distance	Any of the Euclidean embeddable distances (see Gower and Hand (1996) p. 246).
Pythagorean distance between two points \mathbf{x}_i and \mathbf{x}_j	A special case of Euclidean embeddable dis- tance, given by Pythagoras' theorem, namely $\left\{ \sum_{k=1}^p (x_{ik} - x_{jk})^2 \right\}^{\frac{1}{2}} = \left\{ (\mathbf{x}_i - \mathbf{x}_j)' (\mathbf{x}_i - \mathbf{x}_j) \right\}^{\frac{1}{2}}$.
Mahalanobis distance between two points \mathbf{x}_i and \mathbf{x}_j	$\left\{ (\mathbf{x}_i - \mathbf{x}_j)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right\}^{\frac{1}{2}}$ where $\boldsymbol{\Sigma}$ is the popula- tion covariance matrix associated with the stochas- tic vector variable $\underline{\mathbf{x}}$.
Sample Mahalanobis distance between two points \mathbf{x}_i and \mathbf{x}_j	$\left\{ (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right\}^{\frac{1}{2}}$ where \mathbf{S} is the sample covariance matrix associated with the stochastic vector variable $\underline{\mathbf{x}}$ corresponding to the set of sam- ples that \mathbf{x}_i and \mathbf{x}_j form part of.
Vector hyperplane	A vector hyperplane in a p -dimensional vector space \mathcal{V} is a $(p - 1)$ -dimensional subspace of \mathcal{V} .
Affine subspace	Given a p -dimensional vector space \mathcal{V} with $\mathbf{v} \in \mathcal{V}$, and a subspace \mathcal{S} of \mathcal{V} , the set $\{\mathbf{s} + \mathbf{v} : \mathbf{s} \in \mathcal{S}\}$ is called an affine subspace of (or flat in) \mathcal{V} . An affine subspace of \mathcal{V} does therefore not necessarily con- tain the null vector.
Affine hyperplane	An affine hyperplane in a p -dimensional vector space \mathcal{V} is a $(p - 1)$ -dimensional affine subspace of \mathcal{V} . An affine subspace that does not pass through the origin can be obtained by performing a trans- lation transformation on a vector hyperplane. In the remainder of this thesis, affine hyperplanes will be referred to simply as hyperplanes.

1.5 Abbreviations

Principal Component Analysis	PCA
Principal Component	PC
Canonical Variate Analysis	CVA
Analysis Of Distance	AOD
Canonical Correlation Analysis	CCA
Correspondence Analysis	CA
Singular Value Decomposition	svd
positive semi-definite	p.s.d.
positive definite	p.d
Sum of Squared Residuals	SSR

1.6 Some needed linear algebra results

A number of basic linear algebra results will be used in this thesis, some of which are discussed below. Before discussing any of these results, consider two definitions which will be encountered frequently in this thesis, namely that of an orthogonal matrix and that of an orthonormal matrix.

A square matrix, \mathbf{U} is an orthogonal matrix if and only if

$$\mathbf{U}'\mathbf{U} = \mathbf{I} \text{ and } \mathbf{U}\mathbf{U}' = \mathbf{I}$$

or equivalently, if and only if

$$\mathbf{U}^{-1} = \mathbf{U}' .$$

It is clear that the set of row vectors and the set of column vectors of an orthogonal matrix are both orthonormal sets. This means that each row vector has unit length and is orthogonal to each of the other row vectors. Similarly, each column vector has unit length and is orthogonal to each of the other column vectors. A rectangular matrix, \mathbf{B} , is an orthonormal matrix if and only if $\mathbf{B}'\mathbf{B} = \mathbf{I}$. When \mathbf{B} is orthonormal, $\mathbf{B}\mathbf{B}'$ is not equal to the identity matrix. The column vectors of an orthonormal matrix therefore form an orthonormal set but the row vectors do not. The row vectors are not orthogonal and each has length smaller or equal to one. An orthonormal matrix is just a submatrix of an orthogonal matrix. Any $n \times p$, where $p \leq n$, rectangular matrix the column vectors of which are p distinct column vectors of an $n \times n$ orthogonal matrix is an orthonormal matrix. For example, if \mathbf{U} is an $n \times n$ orthogonal matrix and $p < n$, then \mathbf{U}_p is an orthonormal matrix. The fact that

the length of each row vector of an orthonormal matrix is less than or equal to one, is shown below:

$$\begin{aligned}
 \mathbf{U}\mathbf{U}' &= \mathbf{I} \\
 \longrightarrow \mathbf{u}'_i\mathbf{u}_i &= \sum_{j=1}^n u_{ij}^2 \quad \forall i \in [1:n] \\
 &= 1 \\
 \longrightarrow \sum_{j=1}^p u_{ij}^2 &= 1 - \sum_{j=p+1}^n u_{ij}^2 \quad \forall i \in [1:n] \\
 \sum_{j=p+1}^n u_{ij}^2 \geq 0 &\longrightarrow \sum_{j=1}^p u_{ij}^2 \leq 1 \quad \forall i \in [1:n]
 \end{aligned}$$

It is evident from the above that each row vector of an orthonormal matrix has a squared length of less than or equal to one and hence also a length of less than or equal to one. Note that it is possible for some of the row vectors of an orthonormal matrix to have lengths equal to one, but it is impossible for all of the row vectors to have lengths equal to one. In order for all the rows to have lengths equal to one, the last $n-p$ columns of \mathbf{U} can contain only zeros, in which case $\mathbf{U}'\mathbf{U} \neq \mathbf{I}$ and hence \mathbf{U} is not an orthogonal matrix. It is also possible for some of the row vectors of an orthonormal matrix to be orthogonal, but again this cannot be true for all row vectors.

1.6.1 The spectral decomposition (eigen-decomposition) of a symmetric matrix

The spectral decomposition of an $n \times n$ symmetric matrix \mathbf{A} of rank $q \leq n$ is given by

$$\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}' = \sum_{i=1}^n d_i \mathbf{v}_i \mathbf{v}'_i$$

where \mathbf{V} is an $n \times n$ orthogonal matrix, the column vectors of which are the normalised orthogonal eigenvectors of \mathbf{A} and \mathbf{D} is a $n \times n$ diagonal matrix, the diagonal elements of which are the eigenvalues of \mathbf{A} :

$$\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}' \longrightarrow \mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{D}.$$

The orthogonal matrix, \mathbf{V} , is said to orthogonally diagonalise \mathbf{A} since $\mathbf{V}'\mathbf{A}\mathbf{V} = \mathbf{D}$. The ordering of the diagonal elements of \mathbf{D} is arbitrary - for every possible ordering

of the diagonal elements of \mathbf{D} together with the corresponding ordering of the column vectors of \mathbf{V} , $\mathbf{A} = \mathbf{VDV}'$ is true. Also, the i th element of any diagonal matrix \mathbf{D} will from this point be denoted with a single subscript: $[D]_{ii} = d_i$. It is assumed in this thesis that the diagonal elements of \mathbf{D} are ordered to be non-increasing i.e. that $d_1 \geq d_2 \geq \dots \geq d_p$ and that the column vectors of \mathbf{V} are ordered accordingly. Since the rank of \mathbf{A} is $q \leq n$, only the first q diagonal elements of \mathbf{D} are non-zero. When \mathbf{A} is positive definite, all n diagonal elements of \mathbf{D} (i.e. all n eigenvalues of \mathbf{A}) are positive while when \mathbf{A} is positive semi-definite, only the first q diagonal elements of \mathbf{D} are positive and the last $n - q$ diagonal elements all equal 0. Similarly, when \mathbf{A} is negative definite, all n diagonal elements of \mathbf{D} (i.e. all n eigenvalues of \mathbf{A}) are negative while when \mathbf{A} is negative semi-definite, only the first q diagonal elements of \mathbf{D} are negative and the last $n - q$ diagonal elements all equal 0.

1.6.2 The square root matrix of a positive definite (p.d) matrix

The square root matrix of a positive definite symmetric matrix \mathbf{A} , is given by the $n \times n$ matrix \mathbf{B} if and only if

$$\mathbf{BB}' = \mathbf{A} .$$

The square root matrix of a positive definite symmetric matrix \mathbf{A} is not unique. Two types of square root matrices exist, namely the symmetric square root matrix and the square root matrix obtained from the Cholesky decomposition (Harville, 1997) of the positive definite symmetric matrix. For a given positive definite symmetric matrix, each of these two types of square root matrices is unique. The square root matrix produced by the Cholesky decomposition is an upper triangular matrix. In the remainder of this thesis only the symmetric square root matrix will be considered. The term ‘square root matrix’ will therefore be used exclusively to refer to the symmetric square root matrix.

The symmetric square root matrix of a positive definite symmetric matrix \mathbf{A} , is given by the $n \times n$ matrix \mathbf{B} if and only if

$$\mathbf{BB} = \mathbf{A} .$$

The symmetric square root matrix of the positive definite symmetric matrix, \mathbf{A} , is denoted by $\mathbf{A}^{1/2}$. If the spectral decomposition of \mathbf{A} is given by

$$\begin{aligned} \mathbf{A} &= \mathbf{VD}^2\mathbf{V}' \\ &= \sum_{i=1}^n d_i^2 \mathbf{v}_i \mathbf{v}_i' \end{aligned}$$

then the symmetric square root matrix of \mathbf{A} , is given by

$$\begin{aligned}\mathbf{A}^{1/2} &= \mathbf{VDV}' \\ &= \sum_{i=1}^n d_i \mathbf{v}_i \mathbf{v}_i' .\end{aligned}$$

The symmetric square root matrix of a positive definite symmetric matrix \mathbf{A} , is also positive definite:

$$d_i^2 > 0 \longrightarrow \sqrt{d_i^2} = |d_i| > 0 \quad i \in [1 : n] .$$

1.6.3 Singular values and singular vectors

Let \mathbf{X} be an $n \times p$ matrix and $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^p$. The pair of vectors (\mathbf{u}, \mathbf{v}) is called a singular vector pair of the matrix \mathbf{X} associated with the singular value λ if the following two conditions are satisfied:

$$\mathbf{X}\mathbf{v} = \lambda\mathbf{u} \tag{1.6.1}$$

$$\mathbf{X}'\mathbf{u} = \lambda\mathbf{v} . \tag{1.6.2}$$

The vectors \mathbf{u} and \mathbf{v} are respectively called the left and right singular vectors of \mathbf{X} associated with the singular value λ . When $n \geq p$, the matrix \mathbf{X} has p pairs of singular vectors and accordingly p singular values. Singular values are defined to be non-negative. The reason for this is that, if λ were to be negative, then multiplying λ as well as one of the singular vectors associated with λ by -1 , results in the conditions in (1.6.1) and (1.6.2) still being satisfied while the singular value is redefined to be the non-negative value, $-\lambda$:

$$\mathbf{X}\mathbf{v} = (-\lambda)(-\mathbf{u}) \longrightarrow \mathbf{X}\mathbf{v} = \lambda\mathbf{u}$$

$$\mathbf{X}'(-\mathbf{u}) = (-\lambda)\mathbf{v} \longrightarrow \mathbf{X}'\mathbf{u} = \lambda\mathbf{v}$$

$$\mathbf{X}(-\mathbf{v}) = (-\lambda)\mathbf{u} \longrightarrow \mathbf{X}\mathbf{v} = \lambda\mathbf{u}$$

$$\mathbf{X}'\mathbf{u} = (-\lambda)(-\mathbf{v}) \longrightarrow \mathbf{X}'\mathbf{u} = \lambda\mathbf{v} .$$

The singular vectors and singular values of the rectangular matrix \mathbf{X} and the eigenvectors and eigenvalues of the symmetric matrices $\mathbf{X}\mathbf{X}'$ and $\mathbf{X}'\mathbf{X}$ are closely related. If \mathbf{u} and \mathbf{v} are respectively the left and right singular vector of \mathbf{X} associated

with the singular value λ , that is if

$$\begin{aligned}\mathbf{X}\mathbf{v} &= \lambda\mathbf{u} \\ \text{and } \mathbf{X}'\mathbf{u} &= \lambda\mathbf{v}\end{aligned}$$

then

$$\begin{aligned}\mathbf{X}\mathbf{v} = \lambda\mathbf{u} &\longrightarrow \mathbf{X}\left(\frac{1}{\lambda}\mathbf{X}'\mathbf{u}\right) = \lambda\mathbf{u} \\ &\longrightarrow \mathbf{X}\mathbf{X}'\mathbf{u} = \lambda^2\mathbf{u} \\ \text{and } \mathbf{X}'\mathbf{u} = \lambda\mathbf{v} &\longrightarrow \mathbf{X}'\left(\frac{1}{\lambda}\mathbf{X}\mathbf{v}\right) = \lambda\mathbf{v} \\ &\longrightarrow \mathbf{X}'\mathbf{X}\mathbf{v} = \lambda^2\mathbf{v}.\end{aligned}$$

It is evident that the left singular vectors of the rectangular matrix \mathbf{X} are eigenvectors of the symmetric matrix $\mathbf{X}\mathbf{X}'$ while the right singular vectors of \mathbf{X} are eigenvectors of the symmetric matrix $\mathbf{X}'\mathbf{X}$ and the squared non-zero singular values of \mathbf{X} are the non-zero eigenvalues of both $\mathbf{X}\mathbf{X}'$ and $\mathbf{X}'\mathbf{X}$.

When the p singular values are distinct, the associated singular vectors are uniquely defined up to multiplication by a scalar. If the singular vectors are normalised to have unit lengths, as will be assumed from this point onwards, then p distinct singular values implies that the associated singular vectors are uniquely defined up to multiplication by -1 . On the other hand, if two singular values, λ_1 and λ_2 , are equal, then the left and right singular vectors associated with λ_1 and λ_2 are not uniquely defined. The two left singular vectors are defined to be any two orthogonal vectors generating the two-dimensional eigenspace of $\mathbf{X}\mathbf{X}'$ associated with the eigenvalue $\lambda_1^2 = \lambda_2^2$ while the two right singular vectors are defined to be any two orthogonal vectors generating the two-dimensional eigenspace of $\mathbf{X}'\mathbf{X}$ associated with the eigenvalue $\lambda_1^2 = \lambda_2^2$.

1.6.4 The singular value decomposition (svd) of a matrix

The singular value decomposition of a matrix factorises the matrix into three matrices - one containing all the left singular vectors, one containing all the singular values and one containing all the right singular vectors.

The singular value decomposition of an $n \times p$ matrix \mathbf{X} of rank q where $q \leq p \leq n$ is given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' \tag{1.6.3}$$

where \mathbf{U} is an $n \times n$ orthogonal matrix, \mathbf{V} is a $p \times p$ orthogonal matrix and \mathbf{D} is an $n \times p$ matrix with q non-zero elements on its main diagonal and zero elements everywhere else. Note that in much of the available literature, equation (1.6.3) is referred to as the complete (or full) svd of the matrix \mathbf{X} . Equation (1.6.3) will however be referred to as the svd of \mathbf{X} in the remainder of this thesis.

It is important to note that the elements on the main diagonal of the matrix \mathbf{D} can be arranged to appear in any order, as long as the column vectors of the matrices, \mathbf{U} and \mathbf{V} , are ordered accordingly. Let

$$[\mathbf{D}]_{ii} = d_i \quad \forall i \in [1 : p] .$$

In the remainder of this thesis it will be assumed that the elements on the main diagonal of \mathbf{D} are arranged in descending order, that is,

$$d_1 \geq d_2 \geq \dots \geq d_p$$

and that the column vectors of \mathbf{U} and \mathbf{V} are ordered accordingly.

Since the rank of \mathbf{X} is equal to q , \mathbf{X} has only q non-zero singular values. It follows that the first q values on the main diagonal of \mathbf{D} are non-zero while the last $p - q$ values on the main diagonal are all equal to zero, that is:

$$d_1 \geq d_2 \geq \dots \geq d_q > d_{q+1} = \dots = d_p = 0 .$$

This implies that the matrix, \mathbf{D} , has the following structure:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where \mathbf{D}_q is the $q \times q$ diagonal matrix, the i th diagonal element of which is equal to d_i , $i \in [1 : q]$. The svd of \mathbf{X} can therefore be expressed in the following reduced form:

$$\mathbf{X} = \mathbf{U}_q \mathbf{D}_q \mathbf{V}'_q \tag{1.6.4}$$

where \mathbf{U}_q is the $n \times q$ orthonormal matrix the i th column vector of which is given by the i th column vector of \mathbf{U} and \mathbf{V}_q is the $p \times q$ orthonormal matrix the i th column vector of which is given by the i th column vector of \mathbf{V} . It is important to note

that in much of the available literature on the svd of a matrix, equation (1.6.4) is referred to as the svd of the matrix, \mathbf{X} . In this thesis however, equation (1.6.4) will be referred to as the reduced form of the svd of the matrix \mathbf{X} .

It follows from $\mathbf{X} = \mathbf{UDV}'$ and the fact that \mathbf{U} and \mathbf{V} are orthogonal matrices that

$$\begin{aligned} \mathbf{XV} &= \mathbf{UD} \\ \text{and } \mathbf{X}'\mathbf{U} &= \mathbf{VD} . \end{aligned}$$

and hence that

$$\begin{aligned} \mathbf{X}\mathbf{v}_{(i)} &= d_i\mathbf{u}_{(i)} \\ \text{and } \mathbf{X}'\mathbf{u}_{(i)} &= d_i\mathbf{v}_{(i)} \end{aligned}$$

for $i \in [1 : p]$. It follows that $\mathbf{u}_{(i)}$ and $\mathbf{v}_{(i)}$ are respectively the left and right singular vectors of \mathbf{X} associated with the singular value d_i . Since $d_1 \geq d_2 \geq \dots \geq d_p$, d_i is the i th largest singular value of \mathbf{X} , hence the i th column vectors of \mathbf{U} and \mathbf{V} are respectively the left and right singular vectors of \mathbf{X} corresponding to the i th largest singular value of \mathbf{X} , $i \in [1 : p]$. For convenience the i th largest singular value of \mathbf{X} henceforth be referred to as the i th singular value and similarly, the left and right singular vectors associated with the i th largest singular value will be referred to as the i th left and right singular vector of \mathbf{X} respectively.

Recall from Section 1.6.3 that the singular vectors and singular values of the rectangular matrix, \mathbf{X} , and the eigenvectors and eigenvalues of the symmetric matrices \mathbf{XX}' and $\mathbf{X}'\mathbf{X}$ are closely related. Similarly, the svd of an $n \times p$ rectangular matrix, \mathbf{X} , and the spectral decompositions of the square symmetric matrices, $\mathbf{X}'\mathbf{X}$ and \mathbf{XX}' are closely related. If the svd of an $n \times p$ rectangular matrix \mathbf{X} of rank q is given by $\mathbf{X} = \mathbf{UDV}$, then the spectral decompositions of \mathbf{XX}' and $\mathbf{X}'\mathbf{X}$ are given by $\mathbf{U}_q\mathbf{D}_q^2\mathbf{U}'_q$ and $\mathbf{V}_q\mathbf{D}_q^2\mathbf{V}'_q$ respectively:

$$\begin{aligned} \mathbf{XX}' &= \{\mathbf{UDV}'\}'\{\mathbf{UDV}'\}' \\ &= \mathbf{UDV}'\mathbf{VD}'\mathbf{U}' \\ &= \mathbf{UDD}'\mathbf{U}' \text{ since } \mathbf{V}'\mathbf{V} = \mathbf{I} \\ \longrightarrow \mathbf{XX}' &= \mathbf{U}_q\mathbf{D}_q^2\mathbf{U}'_q \text{ since } d_i = 0 \ \forall i \in [q + 1 : p] \\ \mathbf{X}'\mathbf{X} &= \{\mathbf{UDV}'\}'\{\mathbf{UDV}'\}' \\ &= \mathbf{VD}'\mathbf{U}'\mathbf{UDV}' \\ &= \mathbf{VD}'\mathbf{DV}' \text{ since } \mathbf{U}'\mathbf{U} = \mathbf{I} \\ \longrightarrow \mathbf{X}'\mathbf{X} &= \mathbf{V}_q\mathbf{D}_q^2\mathbf{V}'_q \text{ since } d_i = 0 \ \forall i \in [q + 1 : p] . \end{aligned}$$

It is evident that the q squared non-zero singular values of \mathbf{X} are identical to the q non-zero eigenvalues of both the square matrices, $\mathbf{X}\mathbf{X}'$ and $\mathbf{X}'\mathbf{X}$. It is also evident that the left singular vector of \mathbf{X} associated with the i th largest singular value of \mathbf{X} , that is d_i , is equal to the eigenvector of $\mathbf{X}\mathbf{X}'$ which is associated with the i th largest eigenvalue of $\mathbf{X}\mathbf{X}'$, that is d_i^2 , $i \in [1 : p]$. Similarly, the right singular vector of \mathbf{X} associated with the i th largest singular value of \mathbf{X} , that is d_i , is equal to the eigenvector of $\mathbf{X}'\mathbf{X}$ which is associated with the i th largest eigenvalue, of $\mathbf{X}'\mathbf{X}$, that is d_i^2 , $i \in [1 : p]$. It is important to note that the last $n - p$ column vectors of the matrix \mathbf{U} are not left singular vectors of the $n \times p$ matrix \mathbf{X} - there exist only p singular values and p singular vector pairs for the matrix \mathbf{X} since $n \geq p$. The last $n - p$ column vectors of \mathbf{U} are elements of the column space of \mathbf{X} or equivalently, elements of the orthogonal complement of the null space of \mathbf{X}' . Note that the svd of a symmetric matrix corresponds exactly with the spectral decomposition of the matrix.

1.6.5 Expressing a matrix of a given rank as the inner product of two matrices of the same rank

Any matrix of rank q can be expressed as the inner product of two rank q matrices (Rao, 1965). Consider an $n \times p$, where $p \leq n$, rank q matrix \mathbf{X} with svd given by $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$. Using the reduced form of the svd of \mathbf{X} , namely $\mathbf{X} = \mathbf{U}_q\mathbf{D}_q\mathbf{V}'_q$ and partitioning the matrix \mathbf{D}_q into \mathbf{D}_q^α and $\mathbf{D}_q^{1-\alpha}$, where $0 \leq \alpha \leq 1$, the matrix \mathbf{X} can be expressed as the inner product of two rank q matrices:

$$\begin{aligned} \mathbf{X} &= \mathbf{U}\mathbf{D}\mathbf{V}' \\ &= \mathbf{U}_q\mathbf{D}_q\mathbf{V}'_q \\ &= \mathbf{U}_q\mathbf{D}_q^\alpha\mathbf{D}_q^{1-\alpha}\mathbf{V}'_q \\ \longrightarrow \mathbf{X} &= \mathbf{E}\mathbf{F} \text{ where } \mathbf{E} = \mathbf{U}_q\mathbf{D}_q^\alpha \text{ and } \mathbf{F} = \mathbf{D}_q^{1-\alpha}\mathbf{V}'_q. \end{aligned}$$

Since the column vectors of the matrix $\mathbf{E} = \mathbf{U}_q\mathbf{D}_q^\alpha$ are just scalar multiples of the column vectors of \mathbf{U}_q , which are known to be orthogonal, the rank of $\mathbf{E} = \mathbf{U}_q\mathbf{D}_q^\alpha$ follows as q . Similarly, since the row vectors of $\mathbf{F} = \mathbf{D}_q^{1-\alpha}\mathbf{V}'_q$ are just scalar multiples of the row vectors of \mathbf{V}'_q , which are known to be orthogonal, the rank of $\mathbf{F} = \mathbf{D}_q^{1-\alpha}\mathbf{V}'_q$ follows as q . This shows that any matrix of rank q can be written as the inner product of two rank q matrices. Geometrically, this means that any $n \times p$ matrix of rank q can be perfectly represented by $n + p$ vectors in q -dimensional space, meaning that the exact elements of the original matrix can be retrieved from the q -dimensional configuration of $n + p$ vectors.

It is clear from the above that the factorization of \mathbf{X} into two rank q matrices, \mathbf{E} and \mathbf{F} , is not unique - every possible value of the scaling parameter, α , where

$$0 \leq \alpha \leq 1$$

results in the inner-product matrix, \mathbf{EF} , being equal to \mathbf{X} . Varying the value of α from 0 to 1 shifts the emphasis from the representation of the objects to the representation of the variables. Not only can the scaling parameter, α be changed without changing the inner product \mathbf{EF} , the configuration depicted by the row vectors of \mathbf{E} and the column vectors of \mathbf{F} can be rotated or reflected about any of the q Cartesian axes without changing the inner product, \mathbf{EF} . This is shown below:

$$\begin{aligned}\mathbf{X} &= \mathbf{U}_q \mathbf{D}_q^\alpha \mathbf{D}_q^{1-\alpha} \mathbf{V}'_q \\ &= \mathbf{U}_q \mathbf{D}_q^\alpha \mathbf{Q}' \mathbf{Q} \mathbf{D}_q^{1-\alpha} \mathbf{V}'_q \text{ given that } \mathbf{Q} \text{ is an orthogonal matrix} \\ &= (\mathbf{U}_q \mathbf{D}_q^\alpha \mathbf{Q}') (\mathbf{Q} \mathbf{D}_q^{1-\alpha} \mathbf{V}'_q) \\ &= \mathbf{EF} \text{ where } \mathbf{E} = \mathbf{U}_q \mathbf{D}_q^\alpha \mathbf{Q}' \text{ and } \mathbf{F} = \mathbf{Q} \mathbf{D}_q^{1-\alpha} \mathbf{V}'_q.\end{aligned}$$

Since \mathbf{Q} is an orthogonal matrix, multiplication by \mathbf{Q} performs a reflection and/or a rotation. If $|\mathbf{Q}| = 1$, multiplication by \mathbf{Q} performs a rotation while if $|\mathbf{Q}| = -1$, multiplication by \mathbf{Q} performs either a reflection only or a reflection and a rotation. It follows from the above that if the row vectors of $\mathbf{U}_q \mathbf{D}_q^\alpha$ and the column vectors of $\mathbf{D}_q^{1-\alpha} \mathbf{V}'_q$ are reflected and/or rotated in exactly the same way (i.e. their coordinates are multiplied by the same orthogonal matrix), then the inner product \mathbf{EF} is left unchanged.

1.6.6 Generalised inverses

The generalised inverse of an $n \times p$ matrix, \mathbf{A} , is any $n \times p$ matrix, \mathbf{A}^- , which satisfies the equation,

$$\mathbf{A} \mathbf{A}^- \mathbf{A} = \mathbf{A} . \quad (1.6.5)$$

The generalised inverse of a singular matrix is not unique, while a non-singular matrix only has one generalised inverse, namely its inverse. If in addition to equation (1.6.5), the matrix \mathbf{A}^- also satisfies equations 1.6.6, 1.6.7 and 1.6.8, then \mathbf{A}^- is the unique Moore-Penrose pseudoinverse (or Moore-Penrose inverse for short) (Harville, 1997) of the matrix, \mathbf{A} :

$$\mathbf{A}^- \mathbf{A} \mathbf{A}^- = \mathbf{A}^- \quad (1.6.6)$$

$$(\mathbf{A} \mathbf{A}^-)' = \mathbf{A} \mathbf{A}^- \quad (1.6.7)$$

$$(\mathbf{A}^- \mathbf{A})' = \mathbf{A}^- \mathbf{A} . \quad (1.6.8)$$

It is important to note that in some literature, the term, 'generalised inverse', is used as a synonym for the Moore-Penrose pseudoinverse. In this thesis however,

‘the generalised inverse of the matrix’ \mathbf{A} will always refer to any matrix \mathbf{A}^- which satisfies equation (1.6.5).

1.6.7 Projection

All projections are defined in terms of inner products. Therefore, before discussing projections, the definition and characteristics of an inner product must be considered. All inner products can be expressed as a bilinear form. A bilinear form on the other hand only qualifies as an inner product if the matrix of the bilinear form is symmetric and positive definite. Consider two vectors, \mathbf{a} and \mathbf{b} , in \mathbb{R}^p . The function $\mathbf{a}'\mathbf{M}\mathbf{b}$ is called a bilinear form in \mathbf{a} and \mathbf{b} and the matrix \mathbf{M} is called the matrix of the bilinear form. Only when the matrix \mathbf{M} is symmetric and positive definite does the bilinear form $\mathbf{a}'\mathbf{M}\mathbf{b}$ qualify as an inner product for \mathbb{R}^p , because only then are the following four conditions satisfied:

$$\begin{aligned} \mathbf{a}'\mathbf{M}\mathbf{b} &= \mathbf{b}'\mathbf{M}\mathbf{a} \\ \mathbf{a}'\mathbf{M}\mathbf{a} &> 0 \text{ for } \mathbf{a} \neq \mathbf{0} \\ (k\mathbf{a})'\mathbf{M}\mathbf{b} &= k(\mathbf{a}'\mathbf{M}\mathbf{b}) \\ \text{and } (\mathbf{a} + \mathbf{g})'\mathbf{M}\mathbf{b} &= \mathbf{a}'\mathbf{M}\mathbf{b} + \mathbf{g}'\mathbf{M}\mathbf{b} . \end{aligned}$$

Note that in this thesis all positive definite matrices are considered to be symmetric. Hence, the requirement for a bilinear form to qualify as an inner product for \mathbb{R}^p , is that the matrix of the bilinear form must be positive definite. The inner product $\mathbf{a}'\mathbf{M}\mathbf{b}$, is said to be with respect to \mathbf{M} or in the metric \mathbf{M} . Let the inner product in the metric \mathbf{M} be denoted by $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{M}}$, that is

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{M}} = \mathbf{a}'\mathbf{M}\mathbf{b} .$$

The Euclidean inner product, often referred to as the usual inner product, is the inner product given by the bilinear form where the matrix of the bilinear form is the identity matrix, \mathbf{I} . That is, the Euclidean inner product between two vectors, \mathbf{a} and \mathbf{b} , is given by

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{I}} = \mathbf{a}'\mathbf{b} .$$

It will be assumed that when the subscript is omitted from the inner product notation, the inner product being referred to is the inner product in the metric \mathbf{I} , i.e.

the usual (Euclidean) inner product, that is

$$\langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{I}} = \mathbf{a}'\mathbf{b}.$$

A vector space in which the inner product is defined by the Euclidean inner product, is called a Euclidean inner product vector space.

When \mathbf{M} is positive definite and the inner product in the metric \mathbf{M} is chosen to be the inner product for \mathbb{R}^p , then two vectors, \mathbf{a} and \mathbf{b} , are orthogonal if and only if $\mathbf{a}'\mathbf{M}\mathbf{b} = 0$. When $\mathbf{a}'\mathbf{M}\mathbf{b} = 0$, \mathbf{a} and \mathbf{b} are said to be orthogonal with respect to \mathbf{M} or orthogonal in the metric \mathbf{M} . Let the orthogonality of \mathbf{a} and \mathbf{b} in the metric \mathbf{M} be denoted by $\mathbf{a} \perp_{\mathbf{M}} \mathbf{b}$, that is

$$\mathbf{a} \perp_{\mathbf{M}} \mathbf{b} \equiv \mathbf{a}'\mathbf{M}\mathbf{b} = 0.$$

Consider a $p \times q$ matrix, \mathbf{L} , which is such that $\mathbf{M} = \mathbf{L}\mathbf{L}'$. It is shown below that two vectors \mathbf{a} and \mathbf{b} in \mathbb{R}^p are orthogonal in the metric \mathbf{M} if and only if the vectors $\mathbf{L}'\mathbf{a}$ and $\mathbf{L}'\mathbf{b}$ are orthogonal in the metric \mathbf{I} i.e. orthogonal with respect to the usual inner product (Harville, 1997):

$$\begin{aligned} \mathbf{a} \perp_{\mathbf{M}} \mathbf{b} &\equiv \mathbf{a}'\mathbf{M}\mathbf{b} = 0 \\ &\longrightarrow \mathbf{a}'\mathbf{L}\mathbf{L}'\mathbf{b} = 0 \\ &\longrightarrow (\mathbf{L}'\mathbf{a})'(\mathbf{L}'\mathbf{b}) = 0 \\ &\longrightarrow (\mathbf{L}'\mathbf{a}) \perp_{\mathbf{I}} (\mathbf{L}'\mathbf{b}). \end{aligned}$$

When the inner product on \mathbb{R}^p is defined to be the inner product in the metric \mathbf{M} , the projection of a vector \mathbf{a} in \mathbb{R}^p onto another vector, \mathbf{b} , in \mathbb{R}^p is given by

$$\frac{\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{M}}}{\langle \mathbf{b}, \mathbf{b} \rangle_{\mathbf{M}}} \mathbf{b} = \frac{\mathbf{a}'\mathbf{M}\mathbf{b}}{\mathbf{b}'\mathbf{M}\mathbf{b}} \mathbf{b}. \quad (1.6.9)$$

When \mathbf{a} and \mathbf{b} are two vectors in a p -dimensional Euclidean inner product vector space, then the projection of \mathbf{a} onto \mathbf{b} is given by

$$\frac{\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{I}}}{\langle \mathbf{b}, \mathbf{b} \rangle_{\mathbf{I}}} \mathbf{b} = \frac{\mathbf{a}'\mathbf{b}}{\mathbf{b}'\mathbf{b}} \mathbf{b}. \quad (1.6.10)$$

Let \mathbf{a} and \mathbf{b} be elements of an p -dimensional inner product vector space \mathcal{W} , in which the inner product between \mathbf{a} and \mathbf{b} is defined in the metric \mathbf{M} and $\mathcal{V}(\mathbf{B})$ be a subspace of \mathcal{W} . The projection of \mathbf{a} onto $\mathcal{V}(\mathbf{B})$ in the metric \mathbf{M} is given by $\mathbf{z} = \mathbf{B}\mathbf{y}$ where \mathbf{y} is any solution of the linear system

$$\mathbf{B}'\mathbf{M}\mathbf{B}\mathbf{y} = \mathbf{B}'\mathbf{M}\mathbf{a}. \quad (1.6.11)$$

The linear system in equation (1.6.11) is always consistent (Harville, 1997). Every solution of the linear system in (1.6.11) is of the form,

$$\mathbf{y} = (\mathbf{B}'\mathbf{M}\mathbf{B})^{-} \mathbf{B}'\mathbf{M}'\mathbf{a}$$

where $(\mathbf{B}'\mathbf{M}\mathbf{B})^{-}$ is a generalised inverse of $\mathbf{B}'\mathbf{M}\mathbf{B}$. When \mathbf{B} is non-singular, the matrix $\mathbf{B}'\mathbf{M}\mathbf{B}$ is also non-singular and hence the linear system in (1.6.11) has a unique solution,

$$\mathbf{y} = (\mathbf{B}'\mathbf{M}\mathbf{B})^{-1} \mathbf{B}'\mathbf{M}'\mathbf{a}.$$

When \mathbf{B} is singular, the projection of \mathbf{a} onto the $\mathcal{V}(\mathbf{B})$ in the metric \mathbf{M} is therefore given by

$$\mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-} \mathbf{B}'\mathbf{M}'\mathbf{a}.$$

The matrix $\mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-1} \mathbf{B}'\mathbf{M}'$ is called the projection matrix for projection onto the column space of \mathbf{B} in the metric \mathbf{M} . The projection matrix, $\mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-} \mathbf{B}'\mathbf{M}'$, is invariant to the specific generalised inverse of the matrix $(\mathbf{B}'\mathbf{M}\mathbf{B})$ that is used (Harville, 1997). When \mathbf{B} is non-singular, the projection of \mathbf{a} onto $\mathcal{V}(\mathbf{B})$ in the metric \mathbf{M} is given by

$$\mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-1} \mathbf{B}'\mathbf{M}'\mathbf{a}. \quad (1.6.12)$$

Note that if the matrix \mathbf{B} is reduced to consist of one column vector only, then equation (1.6.12) simplifies to equation (1.6.9).

It can be shown that the weighted sums of squares,

$$(\mathbf{a} - \mathbf{B}\mathbf{y})' \mathbf{M} (\mathbf{a} - \mathbf{B}\mathbf{y})$$

is minimised when \mathbf{y} is a solution of the linear system of equations in (1.6.11) and hence when $\mathbf{B}\mathbf{y}$ is the projection of \mathbf{a} onto the column space of \mathbf{B} in the metric \mathbf{M} (Jolliffe, 2002).

Consider again the $p \times q$ matrix \mathbf{L} which is such that

$$\mathbf{L}\mathbf{L}' = \mathbf{M} .$$

If \mathbf{z} is the projection of \mathbf{a} onto the column space of \mathbf{B} in the metric \mathbf{M} , then $\mathbf{L}'\mathbf{z}$ is the projection of $\mathbf{L}'\mathbf{a}$ onto the column space of $\mathbf{L}'\mathbf{B}$ in the metric \mathbf{I} . This is evident upon substituting $\mathbf{L}\mathbf{L}'$ for \mathbf{M} in the linear system of equations in (1.6.11):

$$\begin{aligned} \mathbf{B}'\mathbf{M}\mathbf{B}\mathbf{y} &= \mathbf{B}'\mathbf{M}\mathbf{a} \\ \longrightarrow \mathbf{B}'\mathbf{L}\mathbf{L}'\mathbf{B}\mathbf{y} &= \mathbf{B}'\mathbf{L}\mathbf{L}'\mathbf{a} \\ \longrightarrow (\mathbf{L}'\mathbf{B})' (\mathbf{L}'\mathbf{B}) \mathbf{y} &= (\mathbf{L}'\mathbf{B})' (\mathbf{L}'\mathbf{a}) . \end{aligned} \quad (1.6.13)$$

It is evident that if \mathbf{y} is a solution of the linear system in (1.6.13), then $\mathbf{L}'\mathbf{B}\mathbf{y}$ will equal the projection of $\mathbf{L}'\mathbf{a}$ onto the column space of $\mathbf{L}'\mathbf{B}$ in the metric \mathbf{I} .

If \mathbf{a} and \mathbf{b} are elements of a p -dimensional Euclidean inner product vector space, \mathcal{W} , and $\mathcal{V}(\mathbf{B})$ is a subspace of \mathcal{W} , then the projection of \mathbf{a} onto $\mathcal{V}(\mathbf{B})$ is given by $\mathbf{z} = \mathbf{B}\mathbf{y}$ where \mathbf{y} is any solution of the linear system

$$\mathbf{B}'\mathbf{B}\mathbf{y} = \mathbf{B}'\mathbf{a} . \quad (1.6.14)$$

Every solution, \mathbf{y} , of the linear system in equation (1.6.14) has the form

$$\mathbf{y} = (\mathbf{B}'\mathbf{B})^- \mathbf{B}'\mathbf{a}$$

where $(\mathbf{B}'\mathbf{B})^-$ is a generalised inverse of $\mathbf{B}'\mathbf{B}$. When \mathbf{B} is non-singular, the matrix, $\mathbf{B}'\mathbf{B}$, is also non-singular and hence the linear system in equation (1.6.14) has a

unique solution given by

$$\mathbf{y} = (\mathbf{B}'\mathbf{B})^{-1} \mathbf{B}'\mathbf{a}.$$

When \mathbf{B} is singular, the projection of \mathbf{a} onto the subspace $\mathcal{V}(\mathbf{B})$, is therefore given by

$$\mathbf{B}(\mathbf{B}'\mathbf{B})^{-} \mathbf{B}'\mathbf{a}.$$

The projection matrix, $\mathbf{B}(\mathbf{B}'\mathbf{B})^{-} \mathbf{B}'$, is invariant to the choice of generalised inverse, $(\mathbf{B}'\mathbf{B})^{-}$, of $(\mathbf{B}'\mathbf{B})$ (Harville, 1997). When \mathbf{B} is non-singular, the projection of \mathbf{a} onto the subspace, $\mathcal{V}(\mathbf{B})$, is given by

$$\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1} \mathbf{B}'\mathbf{a}. \quad (1.6.15)$$

The matrix $\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1} \mathbf{B}'$ is called the projection matrix for projection onto the column space of \mathbf{B} in the metric \mathbf{I} , or simply the projection matrix for projection onto the column space of \mathbf{B} . Note that, if the matrix, \mathbf{B} , is reduced to consist of one column vector only, then equation (1.6.15) simplifies to equation (1.6.10).

It can be shown that the sums of squares

$$(\mathbf{a} - \mathbf{B}\mathbf{y})'(\mathbf{a} - \mathbf{B}\mathbf{y}) \quad (1.6.16)$$

is minimised when \mathbf{y} is a solution of the linear system in 1.6.14 and hence when $\mathbf{B}\mathbf{y}$ is the projection of \mathbf{a} onto the column space of \mathbf{B} in the metric \mathbf{I} .

1.6.7.1 Projection onto an affine subspace

Let \mathbf{x} be some vector in \mathbb{R}^p and let \mathcal{W} denote a linear subspace of \mathbb{R}^p spanned by the column vectors of the matrix \mathbf{V} . Each point in \mathcal{W} can therefore be expressed in the form, $\boldsymbol{\alpha}'\mathbf{V}'$, where $\boldsymbol{\alpha} \in \mathbb{R}$. Let \mathcal{L} denote the linear affine subspace obtained by offsetting \mathcal{W} by a point \mathbf{p} . Each point \mathbf{y} that lies in \mathcal{L} can therefore be expressed in the form,

$$\mathbf{y} = \mathbf{p} + \boldsymbol{\alpha}'\mathbf{V}'.$$

The point, $\mathbf{y} \in \mathcal{L}$ which is closest to \mathbf{x} in terms of Pythagorean distance is the orthogonal projection of \mathbf{x} onto \mathcal{L} . Let $\hat{\mathbf{x}}$ denote the orthogonal projection of \mathbf{x} onto \mathcal{L} . That is,

$$\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{y} \in \mathcal{L}} \{ \|\mathbf{x} - \mathbf{y}\|^2 \} .$$

Since $\hat{\mathbf{x}}$ lies in \mathcal{L} , $\hat{\mathbf{x}}$ can be expressed in the form,

$$\hat{\mathbf{x}} = \mathbf{p} + \hat{\boldsymbol{\alpha}}' \mathbf{V}' .$$

It follows that $\hat{\mathbf{x}} = \mathbf{p} + \hat{\boldsymbol{\alpha}}' \mathbf{V}'$, where

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \operatorname{argmin}_{\boldsymbol{\alpha}} \{ \|\mathbf{x} - (\mathbf{p} + \boldsymbol{\alpha}' \mathbf{V}')\|^2 \} \\ \longrightarrow \hat{\boldsymbol{\alpha}} &= \operatorname{argmin}_{\boldsymbol{\alpha}} \{ \|(\mathbf{x} - \mathbf{p}) - \boldsymbol{\alpha}' \mathbf{V}'\|^2 \} . \end{aligned}$$

The point $\boldsymbol{\alpha}' \mathbf{V}'$ in \mathcal{W} which is closest to the point, $\mathbf{x} - \mathbf{p}$ in \mathbb{R}^p in terms of Pythagorean distance, that is the point which minimises $\|(\mathbf{x} - \mathbf{p}) - \boldsymbol{\alpha}' \mathbf{V}'\|^2$ over all $\boldsymbol{\alpha}$, is given by the orthogonal projection of $\mathbf{x} - \mathbf{p}$ onto \mathcal{W} , that is

$$\boldsymbol{\alpha}' \mathbf{V}' = (\mathbf{x} - \mathbf{p})' \mathbf{V} \mathbf{V}' .$$

It follows that the point in \mathcal{L} which minimises $\|\mathbf{x} - \mathbf{y}\|^2$ over all $\mathbf{y} \in \mathcal{L}$, that is the orthogonal projection of \mathbf{x} onto \mathcal{L} is given by

$$\hat{\mathbf{x}} = \mathbf{p} + (\mathbf{x} - \mathbf{p})' \mathbf{V} \mathbf{V}' .$$

1.6.8 The principal axis theorem

Let $Q : \mathbb{R}^p \rightarrow \mathbb{R}$ be a quadratic form given by $Q = \mathbf{x}' \mathbf{A} \mathbf{x}$, where \mathbf{A} is some $p \times p$ symmetric matrix. According to the principal axis theorem there exist an orthogonal basis for \mathbb{R}^p which is such that

$$Q = \mathbf{y}' \mathbf{D} \mathbf{y}$$

where \mathbf{D} is a $p \times p$ diagonal matrix and \mathbf{y} gives the coordinates of \mathbf{x} with respect to those orthogonal basis vectors.

Given that \mathbf{A} is symmetric, it is orthogonally diagonalisable by the matrix the column vectors of which are the orthogonal eigenvectors of \mathbf{A} . That is, if \mathbf{V} denotes the $p \times p$ matrix the column vectors of which are the orthogonal eigenvectors of \mathbf{A} , then

$$\mathbf{VAV}' = \mathbf{D}$$

where \mathbf{D} is the $p \times p$ diagonal matrix with diagonal elements given by the eigenvalues of \mathbf{A} . Since the coordinates of $\mathbf{x} \in \mathbb{R}^p$ with respect to the orthogonal basis of \mathbb{R}^p given by the column vectors of \mathbf{V} , are given by the elements of $\mathbf{y} = \mathbf{V}'\mathbf{x}$, it follows that

$$Q = \mathbf{y}'\mathbf{D}\mathbf{y} = \sum_{i=1}^p d_i y_i^2.$$

The column vectors of the transformation matrix \mathbf{V} are called the principal axes of the conic, $\mathbf{x}'\mathbf{A}\mathbf{x} = c$, where c is an arbitrary non-negative constant. Note that the half-lengths of the principal axes of

$$\mathbf{x}'\mathbf{A}\mathbf{x} = c$$

are given by $\left\{ \frac{c}{\sqrt{d_i}} \right\}$.

1.6.9 Huygens' principle

According to Huygens' principle, the best fitting r -dimensional affine subspace to a configuration of points in p -dimensional (where $p \geq r$) space, passes through the centroid of those points. A derivation of Huygens' principle is given below (Greenacre, 1984).

Consider a configuration of n points in p -dimensional space. Let the coordinates of the i th point be denoted by \mathbf{x}_i and \mathbf{X} be the $n \times p$ matrix the i th row vector of which is equal to \mathbf{x}_i . Let the coordinates of the centroid of the n points be denoted by $\bar{\mathbf{x}}$, that is:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i.$$

Consider an arbitrary r -dimensional affine subspace of the p -dimensional measurement space and let this affine subspace be denoted by \mathcal{S}^* . Let \mathbf{x}_i^* denote the point in \mathcal{S}^* which is closest to \mathbf{x}_i in terms of Pythagorean distance. Let $\bar{\mathbf{x}}^*$ denote the centroid of the n points, $\{\mathbf{x}_i^*\}$, that is

$$\bar{\mathbf{x}}^* = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^*.$$

It follows that $\bar{\mathbf{x}}^*$ is the point in \mathcal{S}^* which is closest to $\bar{\mathbf{x}}$ in terms of Pythagorean distance. Let

$$\mathbf{t} = \bar{\mathbf{x}} - \bar{\mathbf{x}}^*.$$

Let the r -dimensional affine subspace which contains the n points, $\{\mathbf{x}_i^* + \mathbf{t}\}$, be denoted by \mathcal{S} . Let the point in \mathcal{S} which is closest to \mathbf{x}_i in terms of Pythagorean distance be denoted by $\hat{\mathbf{x}}_i$, that is:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i^* + \mathbf{t}.$$

It will now be shown that if the closeness of an affine subspace to a configuration of points is measured by the sum of the squared Pythagorean distances between the points and their representations in the affine subspace (i.e. the sum of the squared residuals), then the affine subspace which is closest to the configuration of points necessarily contains the centroid of those points. The sum of squared Pythagorean distances between the n points, $\{\mathbf{x}_i\}$, and the corresponding points in \mathcal{S}^* , that is $\{\mathbf{x}_i^*\}$, is given by:

$$\begin{aligned} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{x}_i^*)' (\mathbf{x}_i - \mathbf{x}_i^*) &= \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i + \hat{\mathbf{x}}_i - \mathbf{x}_i^*)' (\mathbf{x}_i - \hat{\mathbf{x}}_i + \hat{\mathbf{x}}_i - \mathbf{x}_i^*) \\ &= \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)' (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \sum_{i=1}^n (\hat{\mathbf{x}}_i - \mathbf{x}_i^*)' (\hat{\mathbf{x}}_i - \mathbf{x}_i^*) \\ &\quad + 2 \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)' (\hat{\mathbf{x}}_i - \mathbf{x}_i^*) \\ &= \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)' (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \sum_{i=1}^n (\hat{\mathbf{x}}_i - \mathbf{x}_i^*)' (\hat{\mathbf{x}}_i - \mathbf{x}_i^*) \\ \longrightarrow \sum_{i=1}^n (\mathbf{x}_i - \mathbf{x}_i^*)' (\mathbf{x}_i - \mathbf{x}_i^*) &= \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)' (\mathbf{x}_i - \hat{\mathbf{x}}_i) + n\mathbf{t}'\mathbf{t} \end{aligned}$$

since

$$\begin{aligned}
\sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)' (\hat{\mathbf{x}}_i - \mathbf{x}_i^*) &= \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)' \mathbf{t} \\
&= \left(n\bar{\mathbf{x}} - \sum_{i=1}^n \hat{\mathbf{x}}_i \right)' \mathbf{t} \\
&= \left(n\bar{\mathbf{x}} - \sum_{i=1}^n (\mathbf{x}_i^* + \mathbf{t}) \right)' \mathbf{t} \\
&= (n\bar{\mathbf{x}} - (n\bar{\mathbf{x}}^* + n\mathbf{t}))' \mathbf{t} \\
&= (n\mathbf{t} - n\mathbf{t})' \mathbf{t} \\
\longrightarrow \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)' (\hat{\mathbf{x}}_i - \mathbf{x}_i^*) &= 0.
\end{aligned}$$

It is evident that the sum of the squared Pythagorean distances between the n points, $\{\mathbf{x}_i\}$, and the corresponding points in \mathcal{S}^* , that is $\{\mathbf{x}_i^*\}$, is greater than the sum of the squared Pythagorean distances between the n points, $\{\mathbf{x}_i\}$, and the corresponding points in \mathcal{S} , that is $\{\hat{\mathbf{x}}_i\}$, by $n\mathbf{t}'\mathbf{t}$. This implies that the r -dimensional affine subspace which is closest to the p -dimensional configuration of n points, $\{\mathbf{x}_i\}$, in terms of the least squares criterion, necessarily contains the centroid of the n points, that is $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$, $1 \leq r \leq p$.

1.6.10 The Eckart-Young theorem

Consider an $n \times p$ matrix \mathbf{X} of rank q , where $q \leq p \leq n$, with svd given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' = \sum_{i=1}^q d_i \mathbf{u}_i \mathbf{v}_i'.$$

The Eckart-Young theorem (Eckart and Young, 1936) states that for any r , where $0 \leq r \leq q$,

$$\min_{\mathbf{B}: \text{rank}(\mathbf{B}) \leq r} \left\{ \text{tr} \{ (\mathbf{X} - \mathbf{B})(\mathbf{X} - \mathbf{B})' \} \right\} = \text{tr} \left\{ (\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})' \right\} \quad (1.6.17)$$

where

$$\hat{\mathbf{X}} = \mathbf{U}_r \mathbf{D}_r \mathbf{V}_r' = \sum_{i=1}^r d_i \mathbf{u}_i \mathbf{v}_i'.$$

That is, the best rank r approximation of \mathbf{X} (in terms of least squares) is given by

$$\widehat{\mathbf{X}}_r = \mathbf{U}_r \mathbf{D}_r \mathbf{V}'_r = \sum_{i=1}^r d_i \mathbf{u}_i \mathbf{v}'_i.$$

If $r < k \leq q$, then the best rank k approximation of \mathbf{X} is given by

$$\begin{aligned} \widehat{\mathbf{X}}_k &= \mathbf{U}_k \mathbf{D}_k \mathbf{V}'_k = \sum_{i=1}^k d_i \mathbf{u}_i \mathbf{v}'_i \\ \longrightarrow \widehat{\mathbf{X}}_k &= \sum_{i=1}^r d_i \mathbf{u}_i \mathbf{v}'_i + \sum_{i=r+1}^k d_i \mathbf{u}_i \mathbf{v}'_i. \end{aligned}$$

It is evident that if $k > r$, the best rank r approximation of \mathbf{X} is contained within the best rank k approximation of \mathbf{X} , that is, the solution provided by the Eckart-Young theorem is a nested solution.

The best rank r approximation of \mathbf{X} can be expressed as an orthogonal projection:

$$\widehat{\mathbf{X}}_r = \mathbf{U}_r \mathbf{D}_r \mathbf{V}'_r = \mathbf{X} \mathbf{V}_r \mathbf{V}'_r.$$

The geometric interpretation of the Eckart-Young theorem can therefore be stated as: the most accurate representation of a q -dimensional ($q \geq r$) configuration of points, with coordinate vectors given by the rows of \mathbf{X} , in a linear subspace of the measurement space that is of dimension r or less, is given by the orthogonal projection of these points onto the r -dimensional subspace that is spanned by the first r right singular vectors of \mathbf{X} , $\mathcal{V}(\mathbf{V}_r)$. For convenience, the orthogonal projection of the configuration of points with coordinate vectors given by the row vectors of \mathbf{X} , onto a subspace, say \mathcal{W} , will henceforth be referred to as the orthogonal projection of \mathbf{X} onto \mathcal{W} .

Consider the trace in (1.6.17) again. Since any $n \times p$ matrix of rank q can be expressed as $\mathbf{Y} \mathbf{Q}_r \mathbf{Q}'_r$ where \mathbf{Y} is an $n \times p$ matrix and \mathbf{Q}_r is a $p \times r$ orthonormal matrix, the trace in (1.6.17) can be expressed as:

$$\text{tr} \{ (\mathbf{X} - \mathbf{B}) (\mathbf{X} - \mathbf{B})' \} = \text{tr} \{ (\mathbf{X} - \mathbf{Y} \mathbf{Q}_r \mathbf{Q}'_r) (\mathbf{X} - \mathbf{Y} \mathbf{Q}_r \mathbf{Q}'_r)' \} \quad (1.6.18)$$

$$\begin{aligned} &= \text{tr} \{ (\mathbf{X} - \mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r + \mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r - \mathbf{Y} \mathbf{Q}_r \mathbf{Q}'_r) (\mathbf{X} \\ &\quad - \mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r + \mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r - \mathbf{Y} \mathbf{Q}_r \mathbf{Q}'_r)' \} \\ &= \text{tr} \{ (\mathbf{X} - \mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r) (\mathbf{X} - \mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r)' \} \\ &\quad + \text{tr} \{ (\mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r - \mathbf{Y} \mathbf{Q}_r \mathbf{Q}'_r) (\mathbf{X} \mathbf{Q}_r \mathbf{Q}'_r - \mathbf{Y} \mathbf{Q}_r \mathbf{Q}'_r)' \}. \quad (1.6.19) \end{aligned}$$

Equation (1.6.19) follows from equation (1.6.18) since

$$\text{tr} \{ (\mathbf{X} - \mathbf{X}\mathbf{Q}_r\mathbf{Q}_r') (\mathbf{X}\mathbf{Q}_r\mathbf{Q}_r' - \mathbf{Y}\mathbf{Q}_r\mathbf{Q}_r')' \} = 0.$$

According to the Eckart-Young theorem the trace in (1.6.18) will be minimised across all $n \times p$ matrices of rank r , $\mathbf{Y}\mathbf{Q}_r\mathbf{Q}_r'$, if $\mathbf{Y} = \mathbf{X}$ and $\mathbf{Q}_r = \mathbf{V}_r$. It is evident that the matrix \mathbf{V}_r minimises the trace $\text{tr} \{ (\mathbf{X} - \mathbf{X}\mathbf{Q}_r\mathbf{Q}_r') (\mathbf{X} - \mathbf{X}\mathbf{Q}_r\mathbf{Q}_r')' \}$, that is the column space $\mathcal{V}(\mathbf{V}_r)$ is the r -dimensional subspace which yields the smallest possible value of the sum of the squared Pythagorean distances between the points $\{\mathbf{x}_i\}$ and their orthogonal projections onto an r -dimensional subspace, that is, $\mathcal{V}(\mathbf{V}_r)$ is the closest subspace to the set of points $\{\mathbf{x}_i\}$ in terms of least squares.

1.6.11 The best fitting r -dimensional affine subspace to a configuration of points in higher dimensional space

Consider a configuration of n points in p -dimensional space. Let the p -component coordinate vector of the i th point be denoted by \mathbf{x}_i and let \mathbf{X} denote the $n \times p$ matrix with i th row vector equal to \mathbf{x}_i' . Furthermore, let the centroid of the n points be denoted by $\bar{\mathbf{x}}$, that is

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i.$$

The r -dimensional affine subspace that yields the smallest sum of squared Pythagorean distances between the n points in the p -dimensional measurement space and their orthogonal projections onto the affine subspace is referred to as the best fitting r -dimensional affine subspace in terms of least squares. In this thesis, the phrase 'best fitting' will only be used to refer to best fitting in terms of the least squares criterion, unless stated otherwise.

Consider the r -dimensional affine subspace that is spanned by the r p -component vectors, $\mathbf{a}_1, \dots, \mathbf{a}_r$ and passes through the point \mathbf{c} . The orthogonal projection of an arbitrary point \mathbf{x} in the p -dimensional measurement space onto this hyperplane can be expressed in the following form:

$$\mathbf{A}\mathbf{b} + \mathbf{c}$$

where \mathbf{A} is the $p \times r$ matrix with j th column vector \mathbf{a}_j , $j \in [1:r]$. In order to find the best fitting r -dimensional affine subspace to the configuration of n points in the p -dimensional measurement space, the vectors, $\mathbf{a}_1, \dots, \mathbf{a}_r$, and the point \mathbf{c} which

minimises the summation

$$\sum_{i=1}^n \{\mathbf{x}_i - (\mathbf{A}\mathbf{b}_i + \mathbf{c})\}' \{\mathbf{x}_i - (\mathbf{A}\mathbf{b}_i + \mathbf{c})\} \quad (1.6.20)$$

has to be found. According to Huygens's principal (Greenacre, 1984), the best fitting r -dimensional affine subspace to a configuration of points in p -dimensional space always contains the centroid of the data points, $r \in [1 : p]$. This means that in order for the summation in (1.6.20) to be a minimum, \mathbf{c} must equal the sample mean vector, $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$.

Let \mathbf{B} be the $n \times r$ matrix with i th row vector given by \mathbf{b}'_i , $i \in [1 : n]$. According to the Eckart-Young theorem, the summation

$$\sum_{i=1}^n \{\mathbf{x}_i - \bar{\mathbf{x}} - \mathbf{A}\mathbf{b}_i\}' \{\mathbf{x}_i - \bar{\mathbf{x}} - \mathbf{A}\mathbf{b}_i\} = \text{tr} \left\{ \left(\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X} - \mathbf{D} \right) \left(\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X} - \mathbf{D} \right)' \right\}$$

where $\mathbf{D} = \mathbf{B}\mathbf{A}'$, will be minimised over all $n \times p$ matrices of rank r , \mathbf{D} , if

$$\mathbf{D} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X}\mathbf{V}_r\mathbf{V}_r'$$

that is if

$$\mathbf{A} = \mathbf{V}_r \text{ and } \mathbf{b}_i = \mathbf{V}_r'(\mathbf{x}_i - \bar{\mathbf{x}})$$

where \mathbf{V}_r is the matrix, the columns of which are the right singular vectors of the centred data matrix, $\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X}$, corresponding to the r largest singular values of $\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X}$. It follows that the best fitting r -dimensional affine subspace to a configuration of points with coordinate vectors given by the row vectors of \mathbf{X} , passes through the centroid of the points, $\bar{\mathbf{x}}$, and is spanned by the first r right singular vectors of $\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X}$, or equivalently the first r eigenvectors of $\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X}$. The point which represents the point \mathbf{x}_i in the best fitting r -dimensional affine subspace is given by the orthogonal projection of \mathbf{x}_i onto this hyperplane, that is

$$\bar{\mathbf{x}} + \mathbf{e}'_i \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \right) \mathbf{X}\mathbf{V}_r\mathbf{V}_r'.$$

Note that it will be assumed throughout this thesis that, unless stated otherwise,

the eigenvalues of $\mathbf{X}'(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}$ are distinct and non-zero. The implications of some of the eigenvalues being equal or some being zero will be discussed briefly in Chapter 2. Note that the lower dimensional space in which the data is to be represented will be referred to as the display space and denoted by \mathcal{L} in the remainder of this thesis.

It can be shown that when \mathbf{X} is centred such that $\mathbf{1}'\mathbf{X} = \mathbf{0}'$, the following result holds (Gower and Hand, 1996):

$$n \sum_{i=1}^n r_i^2 = \sum_{i < j} (d_{ij}^2 - \delta_{ij}^2). \quad (1.6.21)$$

In equation (1.6.21) d_{ij} is the Pythagorean distance between the two points representing the i th and j th samples in the full p -dimensional measurement space, \mathbb{R}^p , δ_{ij} is the Pythagorean distance between the two points representing the i th and j th samples in the lower dimensional display space, \mathcal{L} , where the two points in \mathcal{L} are found by orthogonally projecting the points representing the i th and j th samples in \mathbb{R}^p onto \mathcal{L} and r_i^2 is the squared Pythagorean distance between the point representing the i th sample in \mathbb{R}^p and the point representing the i th sample in \mathcal{L} . This means that when a p -dimensional configuration of points, $\{\mathbf{x}_i\}$, is centred around the origin (in that $\sum_i \mathbf{x}_i = \mathbf{0}$), the best fitting r -dimensional subspace to the configuration, \mathcal{L} , not only represents the points in the p -dimensional configuration as well as possible in that it yields the smallest possible sum of squared residuals between those points and the corresponding points in an r -dimensional subspace to the p -dimensional measurement space, it also represents the Pythagorean distances between the points in the p -dimensional configuration as well as possible in that it yields the smallest possible value of $\sum_{i < j} (d_{ij}^2 - \delta_{ij}^2)$.

It will henceforth be assumed that, unless stated otherwise, if it is stated that a set of Pythagorean distances in a higher dimensional space is optimally approximated (or represented) by a set of Pythagorean distances in a lower dimensional space, it means that the sum of the differences between the squared Pythagorean distances in the higher dimensional space and the corresponding squared Pythagorean distances in the lower dimensional space, is minimised.

1.6.12 The Two-Sided Eigenvalue Problem

The two-sided eigenvalue problem is the problem of finding non-trivial solutions to the system of equations

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w} \quad (1.6.22)$$

(the trivial solution being $\mathbf{w} = \mathbf{0}$) where \mathbf{A} is an $n \times n$ symmetric matrix and \mathbf{B} an $n \times n$ positive definite symmetric matrix. A non-zero vector, \mathbf{w} , satisfying equa-

tion (1.6.22) for a given constant, λ , is called an eigenvector of the two-sided eigenvalue problem associated with the eigenvalue, λ , of the two-sided eigenvalue problem. It is clear that when $\mathbf{B} = \mathbf{I}$, the two-sided eigenvalue problem reduces to the usual known (one-sided) eigenvalue problem, that is the problem of finding non-trivial solutions to the system of equations in (1.6.22).

In order for a solution \mathbf{w} of equation (1.6.22) to be a non-zero vector, λ must be a root of the characteristic equation

$$|\mathbf{A} - \lambda\mathbf{B}| = 0.$$

This characteristic equation is a polynomial of degree n and therefore has n roots (not necessarily all distinct), each of which is associated with a non-trivial solution of equation (1.6.22). Note that, similar to the one-sided eigenvalue problem, the eigenvectors of the two-sided eigenvalue problem are only uniquely defined up to a scalar multiple, that is, if \mathbf{w} is an eigenvector of the two-sided eigenvalue problem associated with the eigenvalue, λ , then any scalar multiple of \mathbf{w} is also an eigenvector of the two-sided eigenvalue problem associated with the eigenvalue, λ . This is shown below:

$$\begin{aligned} \mathbf{A}\mathbf{w} &= \lambda\mathbf{B}\mathbf{w} \\ \longrightarrow \mathbf{A}(c\mathbf{w}) &= \lambda\mathbf{B}(c\mathbf{w}). \end{aligned}$$

In the remainder of this thesis, it will be assumed that every eigenvector, \mathbf{w} , of the two-sided eigenvalue problem is normalised such that

$$\mathbf{w}'\mathbf{B}\mathbf{w} = 1.$$

The eigenvectors of the two-sided eigenvalue problem in (1.6.22) that satisfy this constraint are uniquely defined up to multiplication by minus one.

It will now be shown that when $\lambda_i \neq \lambda_j$, the solutions $\mathbf{w}_{(i)}$ and $\mathbf{w}_{(j)}$ of $\mathbf{A}\mathbf{w} = \lambda_i\mathbf{B}\mathbf{w}$ and $\mathbf{A}\mathbf{w} = \lambda_j\mathbf{B}\mathbf{w}$ respectively are orthogonal in the metric, \mathbf{B} , that is $\mathbf{w}'_{(j)}\mathbf{B}\mathbf{w}_{(i)} = 0$. Let $\mathbf{w}_{(i)}$ and $\mathbf{w}_{(j)}$ be solutions of $\mathbf{A}\mathbf{w} = \lambda_i\mathbf{B}\mathbf{w}$ and $\mathbf{A}\mathbf{w} = \lambda_j\mathbf{B}\mathbf{w}$ respectively with $\lambda_i \neq \lambda_j$. This implies that the following two equations hold:

$$\begin{aligned} \mathbf{w}'_{(i)}\mathbf{A}\mathbf{w}_{(j)} &= \lambda_j\mathbf{w}'_{(i)}\mathbf{B}\mathbf{w}_{(j)} \\ \mathbf{w}'_{(j)}\mathbf{A}\mathbf{w}_{(i)} &= \lambda_i\mathbf{w}'_{(j)}\mathbf{B}\mathbf{w}_{(i)} \end{aligned}$$

Since $\mathbf{w}'_{(i)}\mathbf{A}\mathbf{w}_{(j)} = \mathbf{w}'_{(j)}\mathbf{A}\mathbf{w}_{(i)}$, it follows that

$$\lambda_j \mathbf{w}'_{(i)}\mathbf{B}\mathbf{w}_{(j)} = \lambda_i \mathbf{w}'_{(j)}\mathbf{B}\mathbf{w}_{(i)}$$

and since $\lambda_i \neq \lambda_j$, it follows that

$$\mathbf{w}'_{(i)}\mathbf{B}\mathbf{w}_{(j)} = \mathbf{w}'_{(j)}\mathbf{B}\mathbf{w}_{(i)} = 0.$$

■

The set of n solutions of the system of equations in (1.6.22) can be simultaneously represented by

$$\mathbf{A}\mathbf{W} = \mathbf{B}\mathbf{W}\mathbf{\Lambda} \tag{1.6.23}$$

where $\mathbf{\Lambda}$ is an $n \times n$ diagonal matrix, the diagonal elements of which are the eigenvalues of the two-sided eigenvalue problem and \mathbf{W} is an $n \times n$ matrix with j th column vector, $\mathbf{w}_{(j)}$, equal to the eigenvector of the two-sided eigenvalue problem which is associated with the eigenvalue given by the j th diagonal element of $\mathbf{\Lambda}$. Note that the diagonal elements of $\mathbf{\Lambda}$ can be ordered in any way, as long as the column vectors of \mathbf{W} are ordered accordingly. In the remainder of this thesis it will be assumed that the eigenvalues of a two-sided eigenvalue problem in (1.6.22) are ordered in descending order as the diagonal elements of $\mathbf{\Lambda}$. For convenience the j th column vector of \mathbf{W} , that is the eigenvector of the two-sided eigenvalue problem associated with the j th largest eigenvalue, λ_j , will be referred to as the j th eigenvector of the two-sided eigenvalue problem. Note that since

$$\mathbf{w}'_{(i)}\mathbf{B}\mathbf{w}_{(j)} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

it follows that

$$\mathbf{W}'\mathbf{B}\mathbf{W} = \mathbf{I}.$$

Note that if equation (1.6.23) is pre-multiplied by \mathbf{W} , then the following expression

is obtained:

$$\mathbf{W}'\mathbf{A}\mathbf{W} = \mathbf{\Lambda}$$

is obtained. It is evident that given any symmetric matrix \mathbf{A} and positive definite matrix \mathbf{B} , there exists a non-singular transformation matrix, \mathbf{W} , which is such that $\mathbf{W}'\mathbf{B}\mathbf{W} = \mathbf{I}$ and $\mathbf{W}'\mathbf{A}\mathbf{W} = \mathbf{\Lambda}$ where $\mathbf{\Lambda}$ is a diagonal matrix with its diagonal elements written in non-increasing order. The matrix \mathbf{W} is formed from the solutions to the two-sided eigenvalue problem - the j th column vector of \mathbf{W} , $\mathbf{w}_{(j)}$, is the solution of

$$\mathbf{A}\mathbf{w} = \lambda_j\mathbf{B}\mathbf{w}$$

where λ_j is the j th largest eigenvalue of the two-sided eigenvalue problem in (1.6.22). The vector $\mathbf{w}_{(j)}$ is called the eigenvector of the two-sided eigenvalue problem in (1.6.22) corresponding to the j th largest eigenvalue of that two-sided eigenvalue problem.

The only task left is to find the solutions of the two-sided eigenvalue problem. Note that since \mathbf{B} is positive definite, it has a unique square root matrix. Let the svd of \mathbf{B} be given by

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' .$$

It follows that the square root matrix of \mathbf{B} is given by

$$\mathbf{B}^{1/2} = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{V}$$

where

$$[\mathbf{\Lambda}^{1/2}]_{ij} = \begin{cases} [\mathbf{\Lambda}]_{ii}^{1/2} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

(see Section 1.6.2). By manipulating the expression $\mathbf{A}\mathbf{W} = \mathbf{B}\mathbf{W}\mathbf{\Lambda}$ algebraically using $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}$ and $\mathbf{B}^{1/2}\mathbf{B}^{-1/2} = \mathbf{I}$, it can be shown that the i th column vector of $\mathbf{B}^{-1/2}\mathbf{U}$, where \mathbf{U} is the matrix the i th column vector of which is the eigenvector of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ corresponding to its i th largest eigenvalue, is the i th eigenvector of

the two-sided eigenvalue problem:

$$\mathbf{A}\mathbf{W} = \mathbf{B}\mathbf{W}\mathbf{\Lambda} \quad (1.6.24)$$

$$\longrightarrow \mathbf{A}\mathbf{W} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}\mathbf{W}\mathbf{\Lambda}$$

$$\longrightarrow \mathbf{B}^{-1/2}\mathbf{A}\mathbf{W} = \mathbf{B}^{1/2}\mathbf{W}\mathbf{\Lambda}$$

$$\longrightarrow \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{B}^{1/2}\mathbf{W} = \mathbf{B}^{1/2}\mathbf{W}\mathbf{\Lambda}$$

$$\longrightarrow (\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})(\mathbf{B}^{1/2}\mathbf{W}) = (\mathbf{B}^{1/2}\mathbf{W})\mathbf{\Lambda}. \quad (1.6.25)$$

It is evident that the column vectors of $\mathbf{B}^{1/2}\mathbf{W}$ are the eigenvectors of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ with the corresponding eigenvalues given by the diagonal elements of $\mathbf{\Lambda}$. Let $\mathbf{U} = \mathbf{B}^{1/2}\mathbf{W}$ so that:

$$\mathbf{W} = \mathbf{B}^{-1/2}\mathbf{U}.$$

This shows that n solutions of $\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}$ can be found by firstly finding n eigenvectors of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ and then pre-multiplying the matrix with j th column vector given by the eigenvector of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ that corresponds to the j th largest eigenvalue of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, by $\mathbf{B}^{-1/2}$. It is evident from equation (1.6.25) that the eigenvalues of the two-sided eigenvalue problem in (1.6.22) are identical to the eigenvalues of the matrix $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$. This means that the number of non-zero eigenvalues of the two-sided eigenvalue problem in equation (1.6.24) is equal to the rank of the matrix $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, which is equal to the rank of the matrix \mathbf{A} since $\mathbf{B}^{-1/2}$ is a non-singular matrix. It is shown below that the eigenvectors of the two-sided eigenvalue problem in (1.6.22) will only be orthogonal in the metric \mathbf{B} and adhere to the constraints, $\{\mathbf{w}'_{(i)}\mathbf{B}\mathbf{w}_{(i)} = 1\}$, if they are calculated from the set of orthonormal eigenvectors of the matrix $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, that is if the column vectors of $\mathbf{U} = \mathbf{B}^{1/2}\mathbf{W}$ are orthonormal:

$$\mathbf{W}'\mathbf{B}\mathbf{W} = \mathbf{I}$$

$$\longleftrightarrow \mathbf{U}'\mathbf{B}^{-1/2}\mathbf{B}\mathbf{B}^{-1/2}\mathbf{U} = \mathbf{I}$$

$$\longleftrightarrow \mathbf{U}'\mathbf{U} = \mathbf{I}.$$

1.6.13 The generalised svd (The svd in a metric other than \mathbf{I})

Consider the two-sided eigenvalue problem in (1.6.24) where $\mathbf{W}'\mathbf{B}\mathbf{W} = \mathbf{I}_p$ and

$$\mathbf{A} = \mathbf{E}'\mathbf{E}$$

that is the matrix \mathbf{A} is at least positive semi-definite. Let the number of non-zero eigenvalues of the two-sided eigenvalue problem in (1.6.24) be denoted by q , where $q \leq p$. Hence, the matrix $\mathbf{\Lambda}$ has the following structure:

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

It will be shown below that the matrix $\mathbf{U}_q = \mathbf{E}\mathbf{M}_q\mathbf{\Lambda}_q^{-1/2}$ is an orthonormal matrix:

$$\begin{aligned} \mathbf{A}\mathbf{W}_q &= \mathbf{B}\mathbf{W}_q\mathbf{\Lambda}_q \\ \longrightarrow \mathbf{E}'\mathbf{E}\mathbf{W}_q &= \mathbf{B}\mathbf{W}_q\mathbf{\Lambda}_q \\ \mathbf{W}_q'\mathbf{E}'\mathbf{E}\mathbf{W}_q &= \mathbf{\Lambda}_q \\ \mathbf{\Lambda}_q^{-1/2}\mathbf{W}_q'\mathbf{E}'\mathbf{E}\mathbf{W}_q\mathbf{\Lambda}_q^{-1/2} &= \mathbf{I}_q \\ \longrightarrow \mathbf{U}_q'\mathbf{U}_q &= \mathbf{I}_q \end{aligned}$$

where $\mathbf{U}_q = \mathbf{E}\mathbf{W}_q\mathbf{\Lambda}_q^{-1/2}$.

Let \mathbf{U} be an orthogonal matrix the first q column vectors of which are identical to the column vectors of \mathbf{U}_q . Due to the structure of the matrix $\mathbf{\Lambda}$, the matrix $\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{W}^{-1}$ can also be expressed in the following way:

$$\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{W}^{-1} = \mathbf{U}_q\mathbf{\Lambda}_q^{1/2}\mathbf{W}_q.$$

Since $\mathbf{W}_q'\mathbf{W}_q = \mathbf{I}_q$, it follows that

$$\mathbf{U}_q = (\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{W}^{-1})\mathbf{W}_q\mathbf{\Lambda}_q^{-1/2}.$$

Due to the fact that $\mathbf{U}_q = \mathbf{E}\mathbf{W}_q\mathbf{\Lambda}_q^{-1/2}$, it follows that

$$\mathbf{E}\mathbf{W}_q\mathbf{\Lambda}_q^{-1/2} = (\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{W}^{-1})\mathbf{W}_q\mathbf{\Lambda}_q^{-1/2}. \quad (1.6.26)$$

It is evident that

$$\mathbf{E} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{W}^{-1} \quad (1.6.27)$$

is a solution of equation (1.6.26). Since $\mathbf{W}'\mathbf{B}\mathbf{W} = \mathbf{I}$, the matrix \mathbf{E} can also be expressed as:

$$\mathbf{E} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{W}'\mathbf{B}. \quad (1.6.28)$$

Both equation (1.6.27) and equation (1.6.28) are referred to as the generalised svd of the matrix \mathbf{E} in the metric \mathbf{B} .

In order to obtain the generalised svd of the matrix of group means, $\overline{\mathbf{X}}$, in the metric \mathbf{W} , the within-group matrix of sums of squares and cross products, the starting point is the two-sided eigenvalue problem,

$$\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda}$$

where $\mathbf{M}'\mathbf{W}\mathbf{M} = \mathbf{I}_p$. Following the argument above, it is evident that the generalised svd of the matrix $\overline{\mathbf{X}}$ in the metric \mathbf{W} is given by

$$\overline{\mathbf{X}} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{M}^{-1} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{M}'\mathbf{W}. \quad (1.6.29)$$

1.6.14 The generalised Eckart-Young theorem (The Eckart-Young theorem in a metric other than \mathbf{I})

The generalised Eckart-Young theorem follows as a direct consequence of the Eckart-Young theorem and the generalised svd. Given a $J \times p$ matrix \mathbf{E} of rank q , where $q \leq p \leq J$, with generalised svd in the metric \mathbf{B} ,

$$\mathbf{E} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{W}^{-1}$$

where \mathbf{W} is such that $\mathbf{W}\mathbf{W}' = \mathbf{B}^{-1}$, the generalised Eckart-Young theorem in the metric \mathbf{B} states that for any r , where $0 \leq r \leq q$,

$$\min_{\mathbf{F}: \text{rank}(\mathbf{F}) \leq r} \left\{ \text{tr} \left\{ (\mathbf{E} - \mathbf{F}) \mathbf{B} (\mathbf{E} - \mathbf{F})' \right\} \right\} = \text{tr} \left\{ (\mathbf{E} - \widehat{\mathbf{E}}_r) \mathbf{B} (\mathbf{E} - \widehat{\mathbf{E}}_r)' \right\}$$

where

$$\widehat{\mathbf{E}}_r = \mathbf{U}_r \mathbf{\Lambda}_r^{1/2} \mathbf{W}^r .$$

This is shown below:

$$\begin{aligned} \text{tr} \left\{ (\mathbf{E} - \mathbf{F}) \mathbf{B}^{-1} (\mathbf{E} - \mathbf{F})' \right\} &= \text{tr} \left\{ (\mathbf{E} - \mathbf{F}) \mathbf{W}\mathbf{W}' (\mathbf{E} - \mathbf{F})' \right\} \\ &= \text{tr} \left\{ (\mathbf{E}\mathbf{W} - \mathbf{F}\mathbf{W}) (\mathbf{E}\mathbf{W} - \mathbf{F}\mathbf{W})' \right\} \\ &= \text{tr} \left\{ (\mathbf{U}\mathbf{\Lambda}^{1/2} - \mathbf{F}\mathbf{W}) (\mathbf{U}\mathbf{\Lambda}^{1/2} - \mathbf{F}\mathbf{W})' \right\} . \end{aligned}$$

According to the Eckart-Young theorem, the trace,

$$\text{tr} \left\{ (\mathbf{U}\mathbf{\Lambda}^{1/2} - \mathbf{F}\mathbf{W}) (\mathbf{U}\mathbf{\Lambda}^{1/2} - \mathbf{F}\mathbf{W})' \right\}$$

will be minimised across all $J \times p$ matrices of rank less than or equal to r , $\mathbf{F}\mathbf{W}$, if

$$\begin{aligned} \mathbf{F}\mathbf{W} &= \mathbf{U}_r \mathbf{\Lambda}_r^{1/2} [\mathbf{I}_r \quad \mathbf{0}] \\ \longrightarrow \mathbf{F}\mathbf{W} &= [\mathbf{U}_r \mathbf{\Lambda}_r^{1/2} \quad \mathbf{0}] \\ \longrightarrow \mathbf{F} &= [\mathbf{U}_r \mathbf{\Lambda}_r^{1/2} \quad \mathbf{0}] \mathbf{W}^{-1} \\ \longrightarrow \mathbf{F} &= \mathbf{U}_r \mathbf{\Lambda}_r^{1/2} \mathbf{W}^r \\ \longrightarrow \mathbf{F} &= \mathbf{E}\mathbf{W}_r \mathbf{W}^r . \end{aligned}$$

It follows that the trace, $\text{tr} \left\{ (\mathbf{E} - \mathbf{F}) \mathbf{B}^{-1} (\mathbf{E} - \mathbf{F})' \right\}$, will be minimised across all matrices of rank less than or equal to r , \mathbf{F} , if

$$\mathbf{F} = \mathbf{U}_r \mathbf{\Lambda}_r^{1/2} \mathbf{W}^r = \mathbf{E}\mathbf{W}_r \mathbf{W}^r = \widehat{\mathbf{E}}_r .$$

Given the generalised svd of the matrix of group means, $\bar{\mathbf{X}}$, in the metric $\mathbf{W} = (\mathbf{M}\mathbf{M}')^{-1}$ provided in (1.6.29), the generalised Eckart-Young theorem in the metric \mathbf{W} states that the trace

$$\text{tr} \left\{ (\bar{\mathbf{X}} - \mathbf{F}) \mathbf{W}^{-1} (\bar{\mathbf{X}} - \mathbf{F})' \right\}$$

will be minimised across all matrices of rank less than or equal to r , \mathbf{F} , if

$$\mathbf{F} = \mathbf{U}_r \mathbf{\Lambda}_r^{1/2} \mathbf{M}_r = \bar{\mathbf{X}} \mathbf{M}_r (\mathbf{M}^{-1})' = \widehat{\bar{\mathbf{X}}}.$$

Chapter 2 - PCA and the PCA biplot

2.1 Introduction

Principal component analysis (PCA) is a multivariate linear dimension reduction technique and probably the most popular of the techniques that fall into that category. PCA dates back to publications by Pearson (1901) and Hotelling (1933), who independently of each other arrived at PCA following two very different routes. While Pearson searched to find the straight line or hyperplane which is best fitting to a higher dimensional configuration of points, Hotelling aimed to summarise the total sample variance associated with the set of measured variables by means of a few uncorrelated linear combinations of the measured variables.

Three reasons why PCA is such a popular dimension reduction technique are that (1) PCA provides a nested solution i.e. if $k > r$, then the r -dimensional PCA solution is contained within the k -dimensional PCA solution, (2) it is easy to understand and (3) much research has been done on the topic.

In order to fully understand the construction of the PCA biplot as well as the interpretation thereof, a sound understanding of PCA is required. All three of the above mentioned approaches to PCA will therefore be discussed in this chapter. However, since the aim of this thesis is to investigate the quality of the PCA biplot in terms of its ability to reproduce the values in the data matrix at hand, PCA will be viewed from Pearson's perspective when discussing the construction and interpretation of the PCA biplot.

2.2 Deriving PCA

The approaches to PCA discussed in the next three sections will be discussed in view of a data set comprising n individuals or objects, which will henceforth be referred to as samples, each of which is measured on the same p variables, where $p \leq n$. It is evident that the measurement vector associated with a sample is contained in the p -dimensional space \mathbb{R}^p , which is defined by p orthogonal Cartesian axes lying in the directions of the p -dimensional unit vectors $\{\mathbf{e}_i\}$. This p -dimensional space will henceforth be referred to as the measurement space. Each of the p Cartesian axes in the measurement space represents one of the p measured variables. The Cartesian axis representing the k th measured variable will henceforth be referred to as the k th Cartesian axis.

The two above mentioned approaches to PCA will now be discussed in the order in which they originated, namely that of Pearson (1901) first and then that of

Hotelling (1933).

2.2.1 Pearson's approach to PCA

Pearson (1901) searched to find the best fitting (affine) hyperplane, that is the best fitting $(p - 1)$ -dimensional affine subspace, to a configuration of points in a p -dimensional vector space. Given a particular hyperplane, it seems natural that, in order to best represent the configuration of points in that hyperplane, each point should be represented by the point in the hyperplane it is nearest to. Pearson used the Pythagorean distance metric to measure distance. Note that in much of the available literature, Pythagorean distance is referred to as Euclidean distance. However, since the latter term can be used to refer to any Euclidean embeddable distance metric (Gower and Hand, 1996), the term Pythagorean distance will be used in this thesis to avoid ambiguity. Given that distance is measured by the Pythagorean distance metric, it follows that given an arbitrary hyperplane, the point on that hyperplane that is nearest to a particular point in the p -dimensional configuration, is the orthogonal projection of that point onto the hyperplane. Hence, Pearson defined the best fitting hyperplane to a p -dimensional configuration of points to be that hyperplane which corresponds to the smallest possible sum of squared Pythagorean distances between the points in the p -dimensional configuration and their orthogonal projections onto the hyperplane, that is the hyperplane that corresponds to the smallest possible sum of squared residuals. The configuration of points in the best fitting hyperplane which is obtained by orthogonally projecting the configuration of points in the measurement space onto the hyperplane, is the best $(p - 1)$ -dimensional representation of the configuration in the measurement space in terms of the least squares criterion. In his article written in 1901, Pearson derived expressions for the best fitting straight line as well as the best fitting hyperplane to a configuration of points in a higher, say p -dimensional, space.

Consider a configuration of n points in p -dimensional space. Let the coordinates of the i th point be denoted by \mathbf{x}_i and \mathbf{X} be the $n \times p$ matrix (where $p \leq n$) the i th row vector of which is given by \mathbf{x}_i' . Let the coordinates of the centroid of the n points be denoted by $\bar{\mathbf{x}}$, that is:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i .$$

Consider an arbitrary hyperplane in the p -dimensional measurement space which lies at a perpendicular distance f from the origin. Every point, \mathbf{x} which lies on this hyperplane satisfies the equation

$$\boldsymbol{\beta}'\mathbf{x} = f .$$

The vector, $\boldsymbol{\beta}$ is called the normal vector of the hyperplane - it is the vector which

lies orthogonal to the hyperplane. Let $\boldsymbol{\beta}$ be normalised such that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. Note that since $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$, the elements of $\boldsymbol{\beta}$ are equal to the direction cosines of the hyperplane, $\boldsymbol{\beta}'\mathbf{x} = f$ (Stewart, 2003).

The perpendicular Pythagorean distance between a point, \mathbf{x}_0 , and the hyperplane, $\boldsymbol{\beta}'\mathbf{x} = f$ is given by

$$\frac{|\boldsymbol{\beta}'\mathbf{x}_0 - f|}{\|\boldsymbol{\beta}\|} = |\boldsymbol{\beta}'\mathbf{x}_0 - f|.$$

The sum of the squared perpendicular Pythagorean distances between the n points in the measurement space and the hyperplane, $\boldsymbol{\beta}'\mathbf{x} = f$, follows as

$$\begin{aligned} U &= \sum_{i=1}^n (\boldsymbol{\beta}'\mathbf{x}_i - f)^2 \\ &= (\mathbf{X}\boldsymbol{\beta} - \mathbf{1}f)'(\mathbf{X}\boldsymbol{\beta} - \mathbf{1}f) \\ &= \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{1}f - f\mathbf{1}'\mathbf{X}\boldsymbol{\beta} + f^2\mathbf{1}'\mathbf{1} \\ &= \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - n f \boldsymbol{\beta}'\bar{\mathbf{x}} - n f \bar{\mathbf{x}}'\boldsymbol{\beta} + n f^2 \\ \longrightarrow U &= \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2n f \boldsymbol{\beta}'\bar{\mathbf{x}} + n f^2. \end{aligned}$$

The first thing that Pearson (1901) illustrated in his article is that the best fitting hyperplane to a configuration of points in a vector space passes through the centroid of the configuration (that is Huygens' principle (Greenacre, 1984)). He did so by differentiating the sum of the squared Pythagorean distances between the points in the measurement space and their orthogonal projections onto an arbitrary hyperplane with respect to the distance of that hyperplane from the origin:

$$\begin{aligned} \frac{dU}{df} &= -2n\boldsymbol{\beta}'\bar{\mathbf{x}} + 2nf \\ \frac{dU}{df} = 0 &\longrightarrow -2n\boldsymbol{\beta}'\bar{\mathbf{x}} + 2nf = 0 \\ \longrightarrow f &= \boldsymbol{\beta}'\bar{\mathbf{x}}. \end{aligned}$$

It is evident from the derivation above that the best fitting hyperplane passes through the centroid of the data points and is therefore of the following form:

$$\boldsymbol{\beta}'\mathbf{x} = \boldsymbol{\beta}'\bar{\mathbf{x}}.$$

Next Pearson showed that if $\boldsymbol{\beta}'\mathbf{x} = f$ is the best fitting hyperplane to the configura-

tion of points in the measurement space, then

$$U = n\boldsymbol{\beta}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}$$

and $\boldsymbol{\beta}$ is a solution of the system of equations,

$$(\widehat{\boldsymbol{\Sigma}} - \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{0} \quad (2.2.1)$$

where λ denotes the mean square residual, that is

$$\lambda = \frac{U}{n}$$

and $\widehat{\boldsymbol{\Sigma}}$ denotes the sample covariance matrix associated with the stochastic vector variable, \mathbf{x} , based on the n data points, $\{\mathbf{x}_i\}$, that is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}.$$

Since $\|\boldsymbol{\beta}\| = 1$, the elements of $\boldsymbol{\beta}$ cannot all be zero and hence the determinant, $|\boldsymbol{\Sigma} - \lambda\mathbf{I}|$, must equal zero. Since the solutions to the system of equations in equation (2.2.1) are to be non-trivial, the mean square residual, λ , is an eigenvalue of the sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}$. Let λ_i denote the i th largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$, $i \in [1 : p]$. Since U will be a minimum if and only if λ is a minimum, it follows that the mean square residual of the best fitting hyperplane to the configuration of points in the p -dimensional measurement space is equal to the smallest eigenvalue, λ_p , of the sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}$. It follows that in order for the hyperplane, $\boldsymbol{\beta}'\mathbf{x} = f$, to be the best fitting hyperplane to the configuration of points in the measurement space, $\boldsymbol{\beta}$ must be a solution of the system of equations,

$$(\widehat{\boldsymbol{\Sigma}} - \lambda_p\mathbf{I})\boldsymbol{\beta} = \mathbf{0}$$

that is, $\boldsymbol{\beta}$ must be the eigenvector of $\widehat{\boldsymbol{\Sigma}}$ which corresponds to the eigenvalue λ_p . It follows that the best fitting hyperplane to the configuration of points in the measurement space passes through the centroid of the points, $\bar{\mathbf{x}}$, and is orthogonal to the eigenvector of $\widehat{\boldsymbol{\Sigma}}$ which corresponds to the smallest eigenvalue, λ_p , of $\widehat{\boldsymbol{\Sigma}}$, or equivalently, orthogonal to the eigenvector of $\widehat{\boldsymbol{\Sigma}}^{-1}$ which corresponds to the largest

eigenvalue, $\frac{1}{\lambda_p}$, of $\widehat{\Sigma}^{-1}$. It follows that if $\bar{\mathbf{x}} = \mathbf{0}$, then the best fitting hyperplane is spanned by the $p - 1$ orthonormal eigenvectors of $\widehat{\Sigma}$ which corresponds to the $p - 1$ largest eigenvalues of $\widehat{\Sigma}$. It is known from the principal axis theorem (Anton and Rorres, 2000) that the greatest axis of the ellipsoid,

$$\mathbf{x}'\widehat{\Sigma}\mathbf{x} = c^2 \quad (2.2.2)$$

lies in the direction of the eigenvector of $\widehat{\Sigma}$ which corresponds to the smallest eigenvalue, λ_p , of $\widehat{\Sigma}$ (see Section 1.6.8) or equivalently, in the direction of the eigenvector of $\widehat{\Sigma}^{-1}$ which corresponds to the largest eigenvalue, $\frac{1}{\lambda_p}$, of $\widehat{\Sigma}^{-1}$. This means that the best fitting hyperplane to the configuration of points in the measurement space passes through the centroid of the points, $\bar{\mathbf{x}}$, and lies orthogonal to the greatest axis of the ellipsoid in 2.2.2. Pearson also derived that the best fitting straight line to the p -dimensional configuration of points, $\{\mathbf{x}_i\}$, passes through the centroid of those points, $\bar{\mathbf{x}}$, and lies in the direction of the eigenvector of $\widehat{\Sigma}$ corresponding to the largest eigenvalue, λ_1 , of $\widehat{\Sigma}$, that is in the direction of the minor axis of the ellipse in equation (2.2.2). Let the spectral decomposition of $\widehat{\Sigma}$ be given by

$$\widehat{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$$

where the diagonal elements of $\mathbf{\Lambda}$ are arranged in descending order. The approximations of \mathbf{X} in the best fitting hyperplane and best fitting straight line are given by

$$\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}\mathbf{V}_{p-1}\mathbf{V}'_{p-1} + \mathbf{1}\bar{\mathbf{x}}' \quad \text{and} \quad \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}\mathbf{V}_1\mathbf{V}'_1 + \mathbf{1}\bar{\mathbf{x}}'$$

respectively. Pearson furthermore showed that the best fitting hyperplane to a p -dimensional configuration of points passes through the best fitting straight line to that configuration of points.

Note that when the matrix \mathbf{X} is centred such that $\mathbf{1}'\mathbf{X} = \mathbf{0}'$, the Pythagorean intersample distances in the best fitting straight line and hyperplane defined by Pearson will yield the optimal representation of the true Pythagorean intersample distances in one and $(p - 1)$ -dimensional space respectively (Section 1.6.11).

Consider the University data set provided in Table 12.9 of Johnson and Wichern (2002). This data set contains information on 25 universities in the United States. For each university, the data set provides the average SAT score of the entering freshmen (*SAT*), the percentage of the entering freshmen who were in the top 10 percent of their high school class (*Top10*), the percentage of applicants which were accepted to attend the particular university (*Accept*), the student-faculty ratio (*SFRatio*), the

estimated annual expense of attending the particular university (*Expense*) and the graduation rate in percentage form (*Grad*). The University data set will be used to illustrate different aspects of PCA and the PCA biplot throughout this chapter as well as Chapter 3. In order to illustrate the concept of the best fitting straight line, consider the measurements of the variables *Top10* and *Grad* of the *University* data set. The best fitting straight line to the two-dimensional configuration of data points described by these two variables is illustrated in Figure 2.1. It is evident that there is a strong positive correlation between the variables *Top10* and *Grad*.

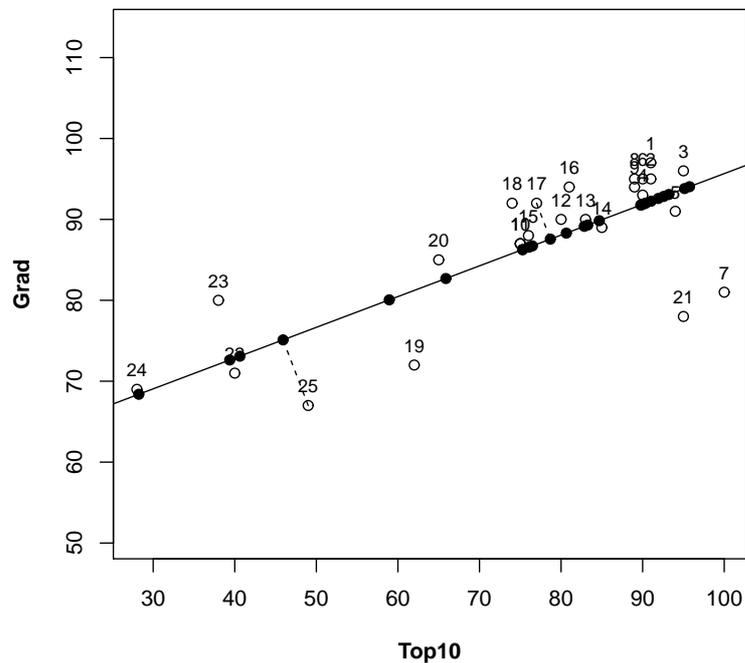


Figure 2.1: The top ten percentages (*Top10*) and graduation rates (*Grad*) (empty circles) of the 25 universities of the *University* data set along with the best fitting straight line (solid line) and the approximated data points (solid circles). The two dashed lines illustrates the orthogonal projection of the data points corresponding to the 17th and 25th universities of the *University* data set onto the best fitting straight line to the two-dimensional configuration of points.

Consider the squared Pythagorean distance between the i th and m th row vectors of \mathbf{X} :

$$\|\mathbf{x}_i - \mathbf{x}_m\|^2 = \sum_{j=1}^p (x_{ij} - x_{mj})^2 .$$

It is evident that for each of the p variables, an absolute deviation of one makes a contribution of one to the squared intersample Pythagorean distance. This means that the Pythagorean distance metric implicitly assumes that an absolute deviation of one is equally significant for all the measured variables, irrespective of the magnitudes of their standard deviations. If however the measured variables do not have identical standard deviations, this is not true and when the measured variables have widely differing standard deviations, the ‘significance’ of an absolute deviation of one differs greatly amongst the variables. Consequently, when the measured variables have greatly differing standard deviations, the Pythagorean intersample distances will not provide trustworthy information regarding the true intersample relationships. It follows that the Pythagorean intersample distances in the best fitting straight line and hyperplane to the p -dimensional configuration of points associated with the unstandardised measurement vectors do not provide the most accurate representation of the true intersample relationships in one and $(p - 1)$ -dimensional space respectively. Hence, when the measured variables have widely differing standard deviations and a lower dimensional representation of the data in which the intersample relationships are represented as accurately as possible, is desired, the variables must first be standardised to have identical standard deviations - standardising to unit standard deviation is often convenient. The configuration of points obtained by orthogonally projecting the p -dimensional configuration of points associated with the standardised measurement vectors onto the best fitting hyperplane (or straight line) to that configuration of points will yield the most accurate representation of the true intersample relationships in a $(p - 1)$ -dimensional (or one-dimensional) affine subspace of the measurements space.

This is also evident upon consideration of the minimisation criteria which defines the best fitting straight line (or hyperplane). Consider the following expression of the sum of the squared Pythagorean distances between the points in the measurement space associated with the set of unstandardised measurement vectors, $\{\mathbf{x}_i\}$, and their orthogonal projections onto a straight line (or hyperplane) in the measurement space (that is the sum of squared residuals or SSR):

$$\sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \hat{x}_{ij})^2 = \sum_{i=1}^n \sum_{j=1}^p \hat{\sigma}_j^2 \left(\frac{x_{ij}}{\hat{\sigma}_j} - \frac{\hat{x}_{ij}}{\hat{\sigma}_j} \right)^2$$

where $\hat{\mathbf{x}}_i$ denotes the orthogonal projection of \mathbf{x}_i onto the straight line and $\hat{\sigma}_j$ denotes the sample standard deviation of the j th measured variable. Since the deviations corresponding to variables with greater standard deviations carry greater weight in the minimisation criteria, the approximated measurements of variables with larger standard deviations will typically be more accurate than those of variables with smaller standard deviations. When a small number of variables amongst the p measured variables have standard deviations which are of much greater magnitude than those of the other variables, the approximation process will be ‘dominated’ by those highly variable variables. Consequently, the measurements of the few variables with large standard deviations will be approximated much more accurately than the

measurements of the rest of the variables. When however the minimization criteria is taken to be the SSR corresponding to the standardised (and centred) data matrix, the approximation process is not dominated by certain variables merely as a result of their standard deviations being of greater magnitude than those of the other variables and hence the resulting lower dimensional representation of the data will typically be a more accurate representation of the truth than the lower dimensional representation resulting from the minimisation of the SSR associated with the unstandardised measurements. It follows that when the measured variables have widely differing standard deviations, the straight line (or hyperplane) in which the data is to be represented should be the best fitting straight line (or hyperplane) to the p -dimensional configuration of points associated with the standardised measurements.

By the same argument as followed before, the best fitting hyperplane to the p -dimensional configuration of standardised measurements lies orthogonal to that eigenvector of the sample correlation matrix, \mathbf{P} , which corresponds to the smallest eigenvalue of \mathbf{P} , or equivalently, orthogonal to the greatest axis of the ellipsoid,

$$\mathbf{z}'\mathbf{P}\mathbf{z} = c^2 \quad (2.2.3)$$

where c is an arbitrary non-zero constant. Similarly, the best fitting straight line to the p -dimensional configuration of standardised measurements lies in the direction of the eigenvector of \mathbf{P} which corresponds to the largest eigenvalue of \mathbf{P} , or equivalently, in the direction of the minor axis of the ellipsoid in (2.2.3). The approximation to the original data matrix can be obtained by multiplying the ij th element of the approximation to the standardised data matrix with the standard deviation of the j th variable and then adding the sample mean of the j th variable, $i \in [1:n]$, $j \in [1:p]$.

2.2.2 Hotelling's approach to PCA

Hotelling arrived at PCA as a result of investigating whether there exists a more fundamental set of independent variables, perhaps less in number than the original measured variables, that would determine the observed values of the measured variables. The motivation behind this investigation was that such a smaller set of variables would reduce the dimensionality of the data, which might ease interpretation and possibly allow graphical representation of the data.

Hotelling referred to the smaller, more fundamental set of independent variables as "components of the complex depicted by the" measured variables (Hotelling, 1933). He started out by considering only normally distributed components having zero means, unit variances and zero correlations. Instead of working with the original measured variables, Hotelling worked with standardised variables which he obtained by subtracting from each variable its mean and then dividing the difference by the variable's standard deviation, since working with these standardised variables produces simpler formulas. Denoting the i th measured variable by x_i^* and its mean

and variance by μ_i^* and σ_{ii}^* respectively, the i th standardised measured variable is given by

$$z_i = \frac{x_i^* - \mu_i^*}{\sqrt{\sigma_{ii}^*}} \quad \forall i \in [1 : p] .$$

Hotelling further confined his search by limiting the standardised variables to be linear combinations of the components. Denoting the i th component by y_i , the relationship between the components and the standardised variables follows as

$$z_i = \sum_{j=1}^p a_{ij} y_j \quad \forall i \in [1 : p] \quad (2.2.4)$$

where a_{ij} is a constant, $i, j \in [1 : p]$. Note that there is no constant term in the summation in equation (2.2.4) since the expected values of z_i and y_j are both zero, $i, j \in [1 : p]$. Let \mathbf{a}_i denote the vector, the j th element of which is equal to a_{ij} . The expression for z_i given in equation (2.2.4) can now be rewritten as follows:

$$z_i = \mathbf{a}_i' \mathbf{y} .$$

Let \mathbf{A} denote the matrix the i th row vector of which is \mathbf{a}_i' . The expressions of the p standardised measured variables can now be represented simultaneously as

$$\mathbf{z} = \mathbf{A} \mathbf{y} . \quad (2.2.5)$$

Since it is desirable to express the standardised variables in terms of the smallest number of components as possible, a procedure is required to select the components in order of decreasing importance and reject those components which are of little importance. Hotelling drew an analogy between the process of selecting components with the process of fitting empirical curves as well as the process of selecting variables to predict a response via a regression equation. When for example fitting a series of the form $y = a + bx + cx^2 + \dots$, the number of terms fitted are limited by the decreasing contributions of the higher order terms to the variance of the response. Similarly, when selecting variables to predict a response via a regression equation, the number of variables included in the regression equation is limited by the diminishing contributions of the predictor variables to the residual variance of the response. These analogies suggest that the first component should be that variable whose contribution to the total variance of the p standardised measured variables

is as great as possible while the second component should be that variable whose contribution to the total residual variance of the standardised measured variables is as great as possible given that it is independent of the first component. The i th component should therefore be the variable whose contribution to the total variance of the standardised measured variables which is unaccounted for by the first $i - 1$ components, is as great as possible given that it is independent of the first $i - 1$ components, $i \in [1 : p]$. Hotelling called the independent components defined by this procedure, the principal components and the procedure used to obtain the principal components he called the “method of principal components” (Hotelling, 1933). The reason why Hotelling constrained the components to be independent of each other is to prevent two or more components from containing the same, or some of the same, information - this is a necessary constraint to ensure that the data is explained as well as possible by as few components as possible. It is important to note that the principal components derived here are population principal components since they are derived from population quantities, while the principal components that were derived by Pearson are sample principal components since they are derived from sample quantities.

Before defining the contribution of a component to the total variance, $\sum_{i=1}^p \text{var}(z_i)$, associated with the stochastic vector variable, \mathbf{z} , consider the following expression of the variance of z_i resulting from the relationship between the standardised variables and the principal components:

$$\begin{aligned} \text{var}(z_i) &= \text{var}(\mathbf{a}'_i \mathbf{y}) \\ &= \mathbf{a}'_i \text{cov}(\mathbf{y}, \mathbf{y}') \mathbf{a}_i \\ &= \mathbf{a}'_i \mathbf{I} \mathbf{a}_i \\ &= \sum_{j=1}^p a_{ij}^2. \end{aligned}$$

It is evident that the contribution of the j th component, y_j , to the variance of z_i is given by a_{ij}^2 , $i, j \in [1 : p]$. It follows that the contribution of y_j to the total variance, $\sum_{i=1}^p \text{var}(z_i)$, is given by

$$\sum_{i=1}^p a_{ij}^2.$$

Denote the vector, the i th element of which is a_{ij} , by $\mathbf{a}_{(j)}$. The contribution of y_j to $\sum_{i=1}^p \text{var}(z_i)$ can then be rewritten in the following way:

$$\sum_{i=1}^p a_{ij}^2 = \mathbf{a}'_{(j)} \mathbf{a}_{(j)} = \|\mathbf{a}_{(j)}\|^2.$$

Denote the contribution of the first principal component, y_1 , to $\sum_{i=1}^p \text{var}(z_i)$, by S , that is:

$$S = \mathbf{a}'_{(1)} \mathbf{a}_{(1)} = \|\mathbf{a}_{(1)}\|^2 .$$

Note that due to the relationships between the principal components and the standardised variables, the elements of $\mathbf{a}_{(1)}$ are bound to satisfy certain constraints which has to be taken into account when maximising S . These constraints are evident from the following expression of the covariance (correlation) between z_i and z_k , ρ_{ik} , $i, k \in [1 : p]$:

$$\begin{aligned} \rho_{ik} &= \text{COV}(z_i, z_k) \\ &= \mathbf{a}'_i \text{COV}(\underline{\mathbf{y}}, \underline{\mathbf{y}}) \mathbf{a}_k \\ &= \mathbf{a}'_i \mathbf{I} \mathbf{a}_k \\ &= \mathbf{a}'_i \mathbf{a}_k . \end{aligned}$$

Remember that since $\text{var}(z_i) = 1$, it follows that

$$\mathbf{a}'_i \mathbf{a}_i = \rho_{ii} = 1 .$$

Hotelling maximised S using differentiation techniques together with Lagrangian multipliers to take the constraints on the values of a_{ij} into account, $i, j \in [1 : p]$. Hotelling's derivation illustrated that in order for S to be maximised under the constraints, $\mathbf{a}'_i \mathbf{a}_k = \rho_{ik}$, for $i, k \in [1 : p]$, the elements of $\mathbf{a}_{(1)}$ must satisfy

$$\mathbf{a}'_{(1)} \mathbf{a}_{(j)} = \delta_{1j} \lambda \text{ where } \delta_{1j} = \begin{cases} 1 & \text{if } j = 1 \\ 0 & \text{if } j \neq 1 \end{cases}$$

where λ is some constant, $j \in [1 : p]$. Note that

$$S = \mathbf{a}'_{(1)} \mathbf{a}_{(1)} = \lambda .$$

Hotelling furthermore derived that λ is a root of the characteristic equation,

$$|\mathbf{P} - \lambda \mathbf{I}| = 0$$

and that $\mathbf{a}_{(1)}$ is the solution of the system,

$$(\mathbf{P} - \lambda_1 \mathbf{I}) \mathbf{a}_{(1)} = \mathbf{0} \quad (2.2.6)$$

where \mathbf{P} denotes the population covariance (correlation) matrix associated with the vector variable, \mathbf{z} , that is

$$[\mathbf{P}]_{ij} = \rho_{ij}.$$

It follows that S is an eigenvalue of the population correlation matrix, \mathbf{P} . Let λ_i denote the i th largest eigenvalue of \mathbf{P} , $i \in [1 : p]$. It follows that the maximum value of S is equal to λ_1 , the largest eigenvalue of the matrix, \mathbf{P} , and that $\mathbf{a}_{(1)}$ is the eigenvector of \mathbf{P} that corresponds to λ_1 which is normalised such that $\mathbf{a}'_{(1)} \mathbf{a}_{(1)} = \lambda_1$. The vector, $\mathbf{a}_{(1)}$, can therefore be found by taking any solution of equation (2.2.6), dividing each element by the square root of the sum of the squared elements of the solution and then multiplying each resulting element by $\sqrt{\lambda_1}$. For example, if \mathbf{b} is an eigenvector of \mathbf{P} corresponding to the eigenvalue, λ_1 , normalised such that $\mathbf{b}'\mathbf{b} = 1$. The vector $\mathbf{a}_{(1)}$ can then be obtained as follows:

$$\mathbf{a}_{(1)} = \frac{\lambda_1^{1/2}}{\sqrt{\sum_{j=1}^p b_j^2}} \mathbf{b}.$$

After the first principal component is found, the component which makes the largest contribution to the residual variance, $\sum_{i=2}^p \text{var}(z_i)$, must be found. Following the same argument as before, Hotelling deduced that the contribution of the j th principal component to the total variance, i.e. $\mathbf{a}'_{(j)} \mathbf{a}_{(j)}$, is equal to the j th largest eigenvalue, λ_j , of \mathbf{P} and that $\mathbf{a}_{(j)}$ is the eigenvector of \mathbf{P} corresponding to λ_j , which is normalised such that $\mathbf{a}'_{(j)} \mathbf{a}_{(j)} = \lambda_j$ and that $\mathbf{a}'_{(j)} \mathbf{a}_{(i)} = 0 \forall i \neq j, i, j \in [2 : p]$. Next, Hotelling derived the expression of the i th principal component in terms of the standardised variables:

$$y_i = \frac{1}{\lambda_i} \mathbf{a}'_{(i)} \mathbf{z} \quad (2.2.7)$$

where $i \in [1 : p]$. The value of the i th principal component for a given sample is referred to as the i th principal component score of that sample, $i \in [1 : p]$. When y_i is defined as in equation (2.2.7), the variance of y_i is equal to one and the i th principal component is uncorrelated with (or equivalently, since the principal components

are normally distributed, independent from) the other $p - 1$ principal components, $i \in [1 : p]$:

$$\begin{aligned}
 \text{var}(\underline{y}_i) &= \frac{1}{\lambda_i} \mathbf{a}'_{(i)} \mathbf{P} \mathbf{a}_{(i)} \frac{1}{\lambda_i} \\
 &= \frac{1}{\lambda_i} \mathbf{a}'_{(i)} \frac{1}{\lambda_i} \mathbf{a}_{(i)} \frac{1}{\lambda_i} \\
 &= \frac{1}{\lambda_i} \mathbf{a}'_{(i)} \mathbf{a}_{(i)} \\
 &= \frac{1}{\lambda_i} \lambda_i \\
 \longrightarrow \text{var}(\underline{y}_i) &= 1 \\
 \text{cov}(\underline{y}_i, \underline{y}_j) &= \frac{1}{\lambda_i} \mathbf{a}'_{(i)} \mathbf{P} \mathbf{a}_{(j)} \frac{1}{\lambda_j} \\
 &= \frac{1}{\lambda_i} \mathbf{a}'_{(i)} \lambda_j \mathbf{a}_{(j)} \frac{1}{\lambda_j} \\
 &= \frac{1}{\lambda_i} \mathbf{a}'_{(i)} \mathbf{a}_{(j)} \\
 \longrightarrow \text{cov}(\underline{y}_i, \underline{y}_j) &= 0 \text{ if } j \neq i .
 \end{aligned}$$

Since the standardised variables, $\{z_i\}$, are defined as linear combinations of normally distributed components, the standardised variables are normally distributed and \mathbf{z} is multivariate normally distributed. Specifically, \mathbf{z} is normally distributed with mean vector, $\mathbf{0}$ and covariance matrix, \mathbf{P} :

$$\mathbf{z} \sim \text{normal}(\mathbf{0}, \mathbf{P}) .$$

It is known that the density of \mathbf{z} is constant on the ellipsoids,

$$\mathbf{z}' \mathbf{P}^{-1} \mathbf{z} = c^2 ,$$

where c is an arbitrary constant. Different values of c produce concentric ellipsoids with identical centroids and principal axes lying in the directions of the eigenvectors of \mathbf{P} (or \mathbf{P}^{-1}). Hotelling showed that the method of principal components, as he defined it, is equivalent to choosing a set of coordinate axes coinciding with the principal axes of the ellipsoids of constant density - the coefficient vectors of the principal components are in the same directions as the principal axes of the ellipsoids of constant density. The fact that the method of principal components, as he defined it, is equivalent to choosing a set of coordinate axes coinciding with the

principal axes of the ellipsoids of constant density also provides an explanation of why it is better to work with the standardised variables than the original measured variables when the measured variables have greatly differing standard deviations. The ellipsoids of constant density can be stretched and squeezed in any way by performing specific linear transformations on \mathbf{x}^* . It follows that the principal components change as a result of performing a linear transformation on \mathbf{x}^* i.e. principal components are scale dependent. Therefore, in order for the principal components to have significant meaning, there needs to be a measurement unit of unique importance and the method of principal components needs to be applied to the variables which are measured in this unit of unique importance. According to Hotelling's definition of the method of principal components, it is the unweighted sum of the variances of the standardised measured variables, $\sum_{i=1}^p \text{var}(z_i)$, which is the quantity to be analyzed. When the method of principal components is applied to the original variables, $\{x_i^*\}$, the quantity to be analyzed is $\sum_{i=1}^p \text{var}(x_i^*) = \sum_{i=1}^p \sigma_{ii}$, which can be viewed as a weighted sum in which variables with greater variance are given a greater weight. Since the variables with greater variance have greater weight in the quantity to be analyzed, the variables with greater variance have greater importance in the analysis. Since the measured variables that have larger variances are given more weight in the quantity to be analyzed, the first few principal components produced will tend to explain the variation in the measured variables with larger variances better than that in the variables with smaller variances i.e. the first few principal components will tend to produce better approximations of the observed measurements of the measured variables with larger variances than for the measured variables with smaller variances. Later in this thesis, it will become clear that this is equivalent to saying that the predictivities of the biplot axes representing measured variables with larger variances will tend to be greater than that of biplot axes representing measured variables with smaller variances. When working with the standardised variables, the variables receive equal weight in the quantity to be analyzed and hence the first few principal components will not be "biased" towards certain variables - they will not produce more accurate approximations for certain variables than for others just because of the relatively large magnitudes of those variables' standard deviations.

The definition of the principal components which has been considered up to now, is not exactly the same as the definition found in most of the available literature. According to most of the literature, Hotelling defined the principal components to be scaled slightly differently. Consider again the relationship between the standardised variables and the components given in equation (2.2.5):

$$\underline{\mathbf{z}} = \mathbf{A}\mathbf{y}.$$

By using the fact that $\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}^{1/2} = \mathbf{I}$, the components of $\underline{\mathbf{z}}$ can be expressed as linear

combinations of the p variables contained as elements of the vector, $\underline{\mathbf{y}}^* = \mathbf{\Lambda}^{1/2}\underline{\mathbf{y}}$:

$$\begin{aligned}\underline{\mathbf{z}} &= \mathbf{A}\underline{\mathbf{y}} \\ \longrightarrow \underline{\mathbf{z}} &= \mathbf{A}\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}^{1/2}\underline{\mathbf{y}} \\ \longrightarrow \underline{\mathbf{z}} &= \mathbf{A}^*\underline{\mathbf{y}}^* \text{ where } \mathbf{A}^* = \mathbf{A}\mathbf{\Lambda}^{-1/2} \text{ and } \underline{\mathbf{y}}^* = \mathbf{\Lambda}^{1/2}\underline{\mathbf{y}}.\end{aligned}$$

If the same argument that was followed in order to derive the expression for y_i given in equation (2.2.7), is followed in order to derive the expression for y_i^* , then the following expression for y_i^* is obtained:

$$\underline{y}_i^* = \mathbf{a}_{(i)}^*\underline{\mathbf{z}} \quad i \in [1 : p].$$

Notice that the i th column vector of \mathbf{A}^* , $\mathbf{a}_{(i)}^*$, is a scalar multiple of the i th column vector of \mathbf{A} , $\mathbf{a}_{(i)}$:

$$\mathbf{a}_{(i)}^* = \frac{1}{\sqrt{\lambda_i}}\mathbf{a}_{(i)}.$$

Since $\mathbf{a}_{(i)}$ is the eigenvector of \mathbf{P} corresponding to the eigenvalue, λ_i , which is normalised such that

$$\mathbf{a}_{(i)}'\mathbf{a}_{(i)} = \lambda_i,$$

it follows that $\mathbf{a}_{(i)}^*$ is an eigenvector of \mathbf{P} corresponding to the eigenvalue, λ_i , normalised such that

$$\mathbf{a}_{(i)}^{*'}\mathbf{a}_{(i)}^* = 1.$$

The next derivation shows that y_i^* and y_j^* are uncorrelated given that $j \neq i$ and that the variance of y_i^* is equal to λ_i :

$$\begin{aligned}\text{cov}(y_i^*, y_j^*) &= \mathbf{a}_{(i)}^{*'}\mathbf{P}\mathbf{a}_{(j)}^* \\ &= \frac{1}{\sqrt{\lambda_i}}\mathbf{a}_{(i)}'\mathbf{P}\mathbf{a}_{(j)}\frac{1}{\sqrt{\lambda_j}}\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{\lambda_i}} \mathbf{a}'_{(i)} \lambda_j \mathbf{a}_{(j)} \frac{1}{\sqrt{\lambda_j}} \\
\longrightarrow \text{cov}(\underline{y}_i^*, \underline{y}_j^*) &= \begin{cases} \lambda_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} .
\end{aligned}$$

It follows that the first principal component can also be defined as the linear combination of the original variables which has the greatest variance and that the i th principal component can be defined as the linear combination of the original variables which has the greatest variance subject on being uncorrelated with the previous $i - 1$ principal components, $i \in [2 : p]$. This is the definition of principal components which is ascribed to Hotelling in most of the available literature and is also the definition which will be used in the remainder of this thesis.

Let $\mathbf{\Lambda}_p$ denote the diagonal matrix, the i th diagonal element of which is equal to the i th largest eigenvalue of \mathbf{P} , namely, λ_i , and $\mathbf{\Phi}$ the matrix the i th column vector of which is the eigenvector of \mathbf{P} corresponding to λ_i which is normalised such that it has a squared norm of one. The spectral decomposition (1.6.1) of \mathbf{P} follows as

$$\mathbf{P} = \mathbf{\Phi} \mathbf{\Lambda}_p \mathbf{\Phi}' .$$

The derivation below shows that the total population variance associated with the vector of standardised measured variables, \mathbf{z} , that is, $\sum_{i=1}^p \text{var}(z_i)$, is equal to the total population variance associated with the vector of principal components, $\mathbf{y} = \mathbf{\Phi}' \mathbf{x}$, that is $\sum_{i=1}^p \text{var}(y_i)$:

$$\begin{aligned}
\sum_{i=1}^p \text{var}(z_i) &= \text{tr}(\mathbf{P}) \\
&= \text{tr}(\mathbf{\Phi} \mathbf{\Lambda}_p \mathbf{\Phi}') \\
&= \text{tr}(\mathbf{\Lambda}_p \mathbf{\Phi}' \mathbf{\Phi}) \\
&= \text{tr}(\mathbf{\Lambda}_p) \\
&= \sum_{i=1}^p \text{var}(\underline{y}_i) \\
\longrightarrow \sum_{i=1}^p \text{var}(z_i) &= \sum_{i=1}^p \text{var}(\underline{y}_i) .
\end{aligned}$$

It is evident that the proportion of the total population variance associated with the vector of standardised measured variables, accounted for by the i th principal

component is given by

$$\frac{\text{var}(y_i)}{\sum_{i=1}^p \text{var}(z_i)} = \frac{\lambda_i}{\sum_{i=1}^p \lambda_i}$$

for $i \in [1 : p]$. It follows that the first r principal components are the r orthonormal linear combinations that account for the greatest proportion of the total variance associated with \mathbf{z} , namely $\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p \lambda_i}$.

Note that since \mathbf{P} is a positive definite matrix, its eigenvalues are positive and can therefore be expressed as the squares of appropriate constants. Since \mathbf{P} is a $p \times p$ symmetric matrix, it has p orthonormal (and therefore linear independent) eigenvectors (Anton and Rorres, 2000), hence there are p possible principal components. Since the multiplication of a vector by -1 does not change its length, the orthonormal eigenvectors of a matrix are uniquely defined up to multiplication by -1 . This is the reason why, when it comes to the interpretation of the principal components, it is not the individual signs of the coefficients, but the pattern of the signs which is important. Since the set of p coefficient vectors defining the p principal components is orthonormal, the p principal components define a new coordinate system which is obtained by rotating the old coordinate system, the rotation being performed by the matrix with column vectors given by the eigenvectors of \mathbf{P} . The axes defining the new coordinate system therefore represents the directions of maximum variability. Note that since the coefficient vectors of the principal components are orthonormal, the coordinates of the standardised (and centred) measurements, $\{\mathbf{z}_i\}$, with respect to the new coordinates axes given by the coefficient vectors of the p principal components are given by the principal component scores associated with the standardised measurements:

$$(\mathbf{Z}\Phi_r\Phi_r')\Phi_r = \mathbf{Z}\Phi_r.$$

In particular, the coordinate of the i th standardised measurement vector, \mathbf{z}_i , with respect to the coordinate axis which lies in the direction of the coefficient vector of the j th principal component is given by the j th principal component score associated with the i th standardised measurement vector, $i, j \in [1 : p]$.

Remember that when the p eigenvalues of \mathbf{P} are distinct, the p orthonormal eigenvectors and therefore the p principal components are uniquely defined up to multiplication by -1 , while if two or more eigenvalues are the same, then the principal component with population variance equal to the value of the common eigenvalue, is not uniquely defined since there is more than one eigenvector (coefficient vector) corresponding to that eigenvalue. For convenience, but without loss of generality, it will henceforth be assumed that all non-zero eigenvalues of \mathbf{P} are unique. See section 2.3 for a brief discussion about principal components with identical and/or zero variances.

Usually, the information which is available is only for a number of observations from the population of interest. When this is the case, PCA is used to summarise the variation in the sample. Hotelling's aim, in the case where only data from a sample is analysed, is to summarise the sample variance by means of as few uncorrelated linear combinations of the measured variables as possible. The principal components derived from a sample are referred to as sample principal components and indicate the directions of maximum sample variability.

Let $\widehat{\Sigma}$ and $\widehat{\mathbf{P}}$ respectively denote the sample covariance matrix and sample correlation matrix associated with a set of independent drawings from some p -dimensional population. When PCA is performed on the standardised measurements, the i th sample principal component is defined to be the linear combination of the standardised variables, $\mathbf{a}'_{(i)}\mathbf{z}$ where

$$\mathbf{a}_{(i)} = \underset{\mathbf{a}}{\operatorname{argmax}} \{ \widehat{\operatorname{var}}(\mathbf{a}'\mathbf{z}) \}$$

while also satisfying

$$\mathbf{a}'_{(i)}\mathbf{a}_{(i)} = 1$$

and

$$\widehat{\operatorname{cov}}\left(\mathbf{a}'_{(k)}\mathbf{z}, \left(\mathbf{a}'_{(i)}\mathbf{z}\right)'\right) = 0 \quad k < i, i, k \in [1 : p].$$

Following the same arguments as for the (population) principal components, it can be shown that the coefficient vector of the i th sample principal component is given by the unit length eigenvector of the sample correlation matrix, $\widehat{\mathbf{P}}$, that corresponds to the i th largest eigenvalue of $\widehat{\mathbf{P}}$, that is $\hat{\lambda}_i$, and that the sample variance of this principal component is given by $\hat{\lambda}_i$, $i \in [1 : p]$. Using the spectral decomposition of $\widehat{\mathbf{P}}$ it can be shown that the total sample variance associated with the standardised measured variables is equal to the total sample variance associated with the p sample principal components. The proportion of the total sample variance associated with the standardised measured variables accounted for by the i th sample principal component is therefore given by

$$\frac{\hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i}.$$

It follows that the first r sample principal components corresponding to the standardised measurements explain the maximum proportion of the total sample variance associated with the p standardised variables that can be explained by r uncorrelated linear combinations of the standardised variables. Hence, the r -dimensional configuration which contains the largest possible proportion of the variance associated with the (standardised) data set is therefore the configuration in which each sample is represented by the point with coordinates given by the first r principal component scores corresponding to the standardised sample.

The i th principal component associated with the unstandardised measurements is defined to be the linear combination of the measured variables with coefficient vector $\mathbf{a}_{(i)}$, where

$$\mathbf{a}_{(i)} = \underset{\mathbf{a}}{\operatorname{argmax}} \{ \operatorname{var}(\mathbf{a}'\mathbf{x}) \}$$

while also satisfying

$$\mathbf{a}'_{(i)}\mathbf{a}_{(i)} = 1$$

and

$$\operatorname{cov}\left(\mathbf{a}'_{(k)}\mathbf{x}, \left(\mathbf{a}'_{(i)}\mathbf{x}\right)'\right) = 0 \quad k < i, i, k \in [1 : p].$$

It can be shown that the coefficient vector of the i th principal component will then be given by the eigenvector of the covariance matrix, $\mathbf{\Sigma}$, associated with the vector of measured variables, which corresponds to the i th largest eigenvalue of $\mathbf{\Sigma}$. Subsequently, the variance of the i th principal component is given by the i th largest eigenvalue of $\mathbf{\Sigma}$ and the proportion of the total variance associated with the vector of measured variables which is accounted for by the i th principal component is given by the ratio of the i th largest eigenvalue of $\mathbf{\Sigma}$ to the summation of all p eigenvalues of $\mathbf{\Sigma}$. Substituting the sample covariance matrix for the population covariance matrix in the above yields the definition and properties of the sample principal components derived from the unstandardised measurements.

Note that since the first principal component is that linear combination of the measured variables which explains the greatest possible proportion of the total variability associated with the set of measured variables, the first principal component associated with the unstandardised measurements will tend to be dominated by the variables with the greatest standard deviations. The proportion of the total variance explained by the first principal component associated with the unstandardised measurements will make it appear to be a better one-dimensional 'summary measure' of the observed data than it truly is. The first principal component associated with

the standardised measurements will not be dominated by certain variables simply as a result of their relatively large standard deviations and will therefore be a better one-dimensional ‘summary’ measure of the observed data than the first principal component associated with the unstandardised. The proportion of the total variance explained by the first principal component associated with the standardised measurements, that is $\frac{1}{p}$, will usually be smaller than the proportion of the total variance explained by the first principal component associated with the unstandardised measurements, thus making the former principal component seem to be a poorer summary measure of the observed data than the latter principal component. The same argument holds for the first few principal components. It follows that when the measured variables have widely differing standard deviations, the variables should be standardised such that they have equal sized variances prior to performing PCA. Standardising to unit variance is usually a convenient option. In the rest of this thesis it will be assumed that ‘standardising the variables’ means standardising the variables to unit variance, unless stated otherwise.

Returning attention to the principal components derived from the standardised measurements, note that the first sample principal component lies in the same direction as the best fitting straight line to the configuration of standardised and centred measurements, $\{\mathbf{z}_i\}$, defined by Pearson (1901). Also, the p th sample principal component lies orthogonal to the best fitting hyperplane defined by Pearson (1901). This implies that the best fitting hyperplane to a p -dimensional configuration of points (as defined by Pearson (1901)) is spanned by the coefficient vectors of the first $p - 1$ sample principal components (derived by Hotelling (1933)). It is known from the Eckart-Young theorem that the r eigenvectors of $\widehat{\mathbf{P}}$ that correspond to the r largest eigenvalues of $\widehat{\mathbf{P}}$, span the best fitting r -dimensional subspace to the p -dimensional configuration of standardised and centred measurements, $\{\mathbf{z}_i\}$ (see Section 1.6.11). It follows that the coefficient vectors of the first r sample principal components derived from the standardised measurements, span the best fitting r -dimensional subspace to the p -dimensional configuration of standardised and centred measurements, $r \in [1 : p]$. Hence, saying that PCA chooses as r -dimensional display space that r -dimensional subspace which accounts for the greatest proportion of the total sample variance associated with the vector of standardised variables is equivalent to saying that PCA chooses as r -dimensional display space that r -dimensional subspace of the p -dimensional measurement space which yields the smallest possible sum of squared residuals between the points in the measurement space and their orthogonal projections onto the subspace. The r -dimensional representation of \mathbf{Z} associated with (sample) PCA, that is the representation of \mathbf{Z} in the r -dimensional PCA display space, is given by

$$\widehat{\mathbf{Z}} = \mathbf{Z}\mathbf{V}_r\mathbf{V}_r'$$

where \mathbf{V}_r is the matrix the j th column vector of which equal to the eigenvector of $\widehat{\mathbf{P}}$ which corresponds to the j th largest eigenvalue of $\widehat{\mathbf{P}}$, $j \in [1 : r]$. It is evident that the r -dimensional approximation to the standardised measurement vector variable,

\mathbf{z} , associated with (sample) PCA is given by

$$\hat{\mathbf{z}}' = \mathbf{z}'\mathbf{V}_r\mathbf{V}_r'.$$

The fact that the coefficient vectors of the sample principal components (that is, the column vectors of \mathbf{V}_r) are orthonormal implies that the coordinates of the point that represents \mathbf{z}_i prime in the r -dimensional PCA display space with respect to the column vectors of \mathbf{V}_r , are given by the first r sample principal component scores of \mathbf{z}_i , that is $\mathbf{z}_i'\mathbf{V}_r\mathbf{V}_r'$. The fact that the coefficient vectors of the first r sample principal components derived from the standardised measurements span the best fitting r -dimensional subspace for all $r \in [1 : p]$ also implies that the best fitting r -dimensional subspace is contained within the best fitting $(r + 1)$ -dimensional subspace - that is, the PCA solution is a nested solution. Since \mathbf{Z} is centred such that $\mathbf{1}'\mathbf{Z} = \mathbf{0}'$, the Pythagorean distances between the p -dimensional configuration of points, $\{\mathbf{z}_i\}$, are optimally approximated by the corresponding Pythagorean distances in the r -dimensional PCA display space. Similarly, the r -dimensional PCA display based on the unstandardised measurement vectors, $\{\mathbf{x}_i\}$, is that r -dimensional subspace of the measurement space which yields the smallest possible sum of squared residuals between the points $\{\mathbf{x}_i\}$ and their orthogonal projections onto the subspace, or equivalently, that r -dimensional subspace in which the true Pythagorean distances between the points, $\{\mathbf{x}_i\}$, are optimally approximated. This definition highlights the fact that PCA is an MDS technique and that the distance metric associated with PCA is the Pythagorean distance metric.

In the rest of this chapter as well as in Chapter 3, PCA will be discussed as it relates to the PCA biplot constructed from a number of samples drawn from a population. It is therefore the sample principal components which will be investigated from this point onwards. For convenience the sample principal components will henceforth be referred to simply as the principal components. Also, since it is the quality of the PCA biplot as approximation to the data set at hand which is of interest in this thesis, PCA will henceforth be viewed from the perspective of Pearson (1901).

2.3 Principal components with zero and/or equal variances

Two complications which may arise when performing PCA are principal components with zero variances and principal components with identical variances. Both of these situations occur very rarely in practice.

Since the variance of the k th principal component is equal to the k th eigenvalue of $\mathbf{X}'\mathbf{X}$ (or equivalently, the squared k th singular value of \mathbf{X}) the variances of q principal components will be identical if and only if the corresponding q eigenvalues of $\mathbf{X}'\mathbf{X}$ are identical. When q eigenvalues are equal, the corresponding q eigenvectors span a unique q -dimensional eigenspace. In this eigenspace, the q eigenvectors are,

2.4. INTERPRETATION OF THE COEFFICIENTS OF THE PRINCIPAL COMPONENTS

61

apart from being orthogonal to each other, arbitrary which means that the principal components associated with the q identical eigenvalues are not uniquely defined. However, due to sample variability, the occurrence of principal components with identical sample variances is very rare in practice.

The variance of a principal component will be zero if and only if the corresponding singular value of \mathbf{X} is zero. A principal component with zero variance therefore indicates an exact linear relationship between the measured variables so that the values of one of the variables in the relationship can be determined exactly from the values of the other variables in the relationship. In order to see this, let \mathbf{u} and \mathbf{v} be the left and right singular vector of \mathbf{X} corresponding to the singular value, d , that is

$$\mathbf{X}\mathbf{v} = d\mathbf{u}.$$

When $d = 0$, the following is true:

$$\begin{aligned} \mathbf{X}\mathbf{v} &= \mathbf{0} \\ \longrightarrow \sum_{j=1}^p v_j x_{ij} &= 0 \quad \forall i \in [1 : n]. \end{aligned}$$

A variable which can be determined exactly from other measured variables is redundant and can be removed from the data set without any loss of information regarding the intersample relationships. When q principal components have zero variance, q singular values of \mathbf{X} are zero (that is, the rank of \mathbf{X} is equal to $p - q$) and hence q exact linear relationships between the measured variables exist. This means that q variables, one from each exact linear relationship, are redundant and can be removed from the data set without any loss of information. Ideally, exact linear relationships between the measured variables should be identified and the redundant variables removed prior to performing PCA. In the remainder of this chapter and the next, the general case where the matrix \mathbf{X} is of rank q , where $q \leq p \leq n$, will be considered. However, only the case where the q non-zero singular values of \mathbf{X} are distinct will be studied.

2.4 Interpretation of the coefficients of the principal components

The magnitude of a variable's coefficient in a particular principal component measures the contribution of the variable to that principal component in the presence of the other measured variables, that is, it measures the variable's multivariate contribution to the principal component. Let \mathbf{X} denote the centred data matrix with

svd given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' .$$

Since the j th principal component is given by $y_j = \mathbf{x}'\mathbf{v}_{(j)}$, this means that the magnitude of v_{kj} measures the contribution of the k th measured variable, x_k , to the j th principal component, $k \in [1 : p]$, $j \in [1 : p]$. The correlation between a measured variable and a principal component on the other hand measures the measured variable's univariate contribution to that principal component. Consider the expression of the estimated correlation between the k th measured variable and the j th principal component:

$$\begin{aligned} \hat{\text{corr}}(x_k, y_j) &= \frac{\hat{\text{cov}}(\mathbf{e}'_k \mathbf{x}, \mathbf{v}'_{(j)} \mathbf{x})}{\sqrt{\hat{\text{var}}(\mathbf{e}'_k \mathbf{x})} \sqrt{\hat{\text{var}}(\mathbf{v}'_{(j)} \mathbf{x})}} \\ &= \frac{\mathbf{e}'_k \hat{\text{cov}}(\mathbf{x}, \mathbf{x}') \mathbf{v}_{(j)}}{\sqrt{\mathbf{e}'_k \hat{\text{cov}}(\mathbf{x}, \mathbf{x}') \mathbf{e}_k} \sqrt{\mathbf{v}'_{(j)} \hat{\text{cov}}(\mathbf{x}, \mathbf{x}') \mathbf{v}_{(j)}}} \\ &= \frac{\mathbf{e}'_k \hat{\Sigma} \mathbf{v}_{(j)}}{\sqrt{\mathbf{e}'_k \hat{\Sigma} \mathbf{e}_k} \sqrt{\mathbf{v}'_{(j)} \hat{\Sigma} \mathbf{v}_{(j)}}} \\ &= \frac{\mathbf{e}'_k \mathbf{X}' \mathbf{X} \mathbf{v}_{(j)}}{\sqrt{\mathbf{e}'_k \mathbf{X}' \mathbf{X} \mathbf{e}_k} \sqrt{\mathbf{v}'_{(j)} \mathbf{X}' \mathbf{X} \mathbf{v}_{(j)}}} \\ &= \frac{\mathbf{e}'_k (d_j^2 \mathbf{v}_{(j)})}{\sqrt{\mathbf{x}'_{(k)} \mathbf{x}_{(k)}} \sqrt{\mathbf{v}'_{(j)} (d_j^2 \mathbf{v}_{(j)})}} \\ &= \frac{d_j^2 v_{kj}}{\sqrt{\mathbf{x}'_{(k)} \mathbf{x}_{(k)}} \sqrt{d_j^2}} \\ \rightarrow \hat{\text{corr}}(x_k, y_j) &= \frac{\sqrt{d_j^2} v_{kj}}{\sqrt{\mathbf{x}'_{(k)} \mathbf{x}_{(k)}}} . \end{aligned}$$

If the matrix, \mathbf{X} upon which PCA is performed is standardised, then the expression of the correlation between the k th measured variable and the j th principal component becomes

$$\hat{\text{corr}}(x_k, y_j) = \sqrt{d_j^2} v_{kj} .$$

It is evident that when PCA is performed on the standardised measurements, the variable with the greatest absolute coefficient for a particular principal component, is also the variable most strongly correlated with that principal component.

2.5 The number of principal components to retain

When it is desired to graphically represent a (centred) data matrix \mathbf{X} in a PCA display, the first step is deciding on the number of principal components that will be used in the approximation of \mathbf{X} . If a target is specified for the proportion of the total sample variance associated with the vector variable $\underline{\mathbf{x}}$ as measured by the one-dimensional measure, $\sum_{i=1}^p \hat{\text{var}}(x_i)$, then any number of principal components that account for a proportion of the total variance equal to or greater than the target proportion can be used to approximate \mathbf{X} .

The required number of principal components can be determined by considering for every $r \in [1:p]$ the proportion of the variance accounted for by the first r principal components. Since the total sample variance associated with the vector variable $\underline{\mathbf{x}}$ is equal to the total sample variance associated with the vector of principal components, the proportion of the total sample variance associated with $\underline{\mathbf{x}}$ that is accounted for by the first r principal components is equal to

$$\frac{\sum_{i=1}^r d_i^2}{\sum_{i=1}^p d_i^2}. \quad (2.5.1)$$

The plot of the ratio in (2.5.1) against r , with r increasing from left to right on the x -axis, is a cumulative version of the scree plot associated with \mathbf{X} , which is the plot of d_i^2 against i for $i \in [1:p]$ (Cattell, 1966). Since only one, two and three-dimensional displays can be visualised, a three-dimensional PCA display should be used to represent \mathbf{X} in the event that the target proportion of the total variance to be accounted for is greater than the proportion accounted for by the first three principal components.

If a target proportion is not specified then it must be determined how many principal components are required to approximate \mathbf{X} sufficiently accurate. Usually the first few principal components account for most of the total variance associated with $\underline{\mathbf{x}}$ while the rest of the principal components account for a negligibly small proportion of the variance. Principal components that account for a negligible proportion of the total variance associated with $\underline{\mathbf{x}}$ can be discarded in the approximation of \mathbf{X} without much, if any loss of information. It must therefore be determined which principal components account for a substantial proportion of the variance associated with $\underline{\mathbf{x}}$, that is which principal components contribute substantially to the approximation to \mathbf{X} , and which account for negligibly small proportions of the total variance. This information is provided by the relative sizes of the sample variances of the principal components, or equivalently the relative sizes of the eigenvalues of the matrix $\mathbf{X}'\mathbf{X}$, and hence can be obtained from the scree plot associated with \mathbf{X} . At the point to the left of which the gradients of the straight lines connecting the points in the scree

plot are high and to the right of which the gradients are low, the scree plot reflects an elbow-shaped form. The number of principal components before the ‘elbow’ usually summarise the total sample variance associated with $\underline{\mathbf{x}}$ to a sufficient extent (given that $\mathbf{X}'\mathbf{X}$ does not have a very large number of small eigenvalues) and is therefore usually a sufficient number of principal components to use in the approximation of \mathbf{X} . Choosing to approximate \mathbf{X} using three or fewer principal components has the advantage that it allows visualisation of the approximation of \mathbf{X} .

2.6 The traditional (classical) biplot

Since “there are many patterns and relationships that are easier to discern in graphical displays than by any other data analysis method” (Everitt, 1994), it is always desirable to graphically represent a data set to be investigated and to do so as accurately as possible. Given that humans can only visualise objects which are at most three-dimensional, it is the graphical representation of a data matrix in one, two or three-dimensional space which is usually of interest.

The lower dimensional space in which the data matrix is graphically represented is referred to as the display space. For generality, let r denote the dimension of the display space, $1 \leq r \leq p$. The ordinary r -dimensional MDS configuration associated with a PCA of a data set is the r -dimensional configuration obtained by orthogonally projecting the configuration of points representing the samples comprising a data set in the measurement space onto the best fitting r -dimensional subspace to that configuration. Although this r -dimensional configuration is optimal in its representation of the samples of the data set, it is lacking in that it does not provide any information on the measured variables of the data set. Gabriel (1971) proposed that a joint representation of the samples and variables of a data set, which he called a biplot, be used to represent the data set. In a biplot each row (i.e. sample) and each column (i.e. variable) of the data matrix is represented by a vector emanating from the origin. These vectors are such that the inner product of a vector representing a row of the data matrix and a vector representing a column of the data matrix is equal to an approximation of the corresponding element of the data matrix. The ‘bi’ in biplot refers to the fact that the biplot is a joint map of two modes, namely observations and variables and does not refer to the dimension of the display. The space in which a biplot is constructed is referred to as the biplot space and will henceforth be denoted by \mathcal{L} .

Consider an $n \times p$ data matrix \mathbf{X} which is centred such that each column has a zero mean, that is such that $\mathbf{1}'\mathbf{X} = \mathbf{0}$. Let the rank of \mathbf{X} be denoted by q , $q \leq p \leq n$. The construction of the biplot is based on the fact that any $n \times p$ matrix of rank q can be expressed as the inner product of an $n \times q$ matrix of rank q and a $q \times p$ matrix of rank q (Section 1.6.5). It follows that \mathbf{X} can be expressed as

$$\mathbf{X} = \mathbf{A}\mathbf{B}'$$

where \mathbf{A} is an $n \times q$ matrix of rank q and \mathbf{B} is a $p \times q$ matrix of rank q . It follows

from the expression, $\mathbf{X} = \mathbf{AB}'$, that every element of \mathbf{X} can be expressed in the form of the inner product between a row vector of \mathbf{A} and a column vector of \mathbf{B}' :

$$x_{ij} = \mathbf{a}'_i \mathbf{b}_j \quad \forall (i, j), \quad i \in [1 : n] \quad \text{and} \quad j \in [1 : p] .$$

Since the row vectors of \mathbf{A} and the column vectors of \mathbf{B}' are q -dimensional, it follows that \mathbf{X} can be perfectly represented by $n + p$ vectors in q -dimensional space. Seeing as

$$\begin{aligned} x_{ij} &= \mathbf{a}'_i \mathbf{b}_j \\ \mathbf{a}_i &= c\mathbf{a}_m \longrightarrow \mathbf{x}_i = c\mathbf{x}_m \\ \text{and } \mathbf{b}_j &= d\mathbf{b}_k \longrightarrow \mathbf{x}_{(j)} = d\mathbf{x}_{(k)} \end{aligned}$$

where c and d are arbitrary constants, the row vectors of \mathbf{A} and \mathbf{B} can be viewed as representing the rows (samples) and columns (variables) of the matrix \mathbf{X} respectively. The i th row of \mathbf{X} and the j th column of \mathbf{X} can therefore be represented by the i th row vector of \mathbf{A} emanating from the origin and the j th row vector of \mathbf{B} emanating from the origin, respectively, $i \in [1 : n]$, $j \in [1 : p]$. Henceforth the vectors stretching from the origin to the points with coordinate vectors given by the i th row vector of \mathbf{A} and the j th row vector of \mathbf{B} will be referred to as the i th row marker and j th column marker respectively, $i \in [1 : n]$, $j \in [1 : p]$. Gabriel suggested that the rows of \mathbf{X} be represented by the endpoints of the row markers only so that the representations of the rows and columns of \mathbf{X} can be easily differentiated in the biplot. The biplot proposed by Gabriel is known as the traditional or classical biplot.

When $q = r$, \mathbf{X} can be expressed as the inner product of an $n \times r$ matrix of rank r and an $r \times p$ matrix of rank r , hence \mathbf{X} can be perfectly represented by $n + p$ vectors in r -dimensional space. When $q > r$, it is however not possible to perfectly represent \mathbf{X} in r -dimensional space. What can be done in such a situation is to perfectly represent a rank r approximation, $\widehat{\mathbf{X}}$, of \mathbf{X} in the r -dimensional biplot space. Since the rank of $\widehat{\mathbf{X}}$ is equal to r , $\widehat{\mathbf{X}}$ can be expressed as

$$\widehat{\mathbf{X}} = \mathbf{AB}'$$

where \mathbf{A} is an $n \times r$ matrix of rank r and \mathbf{B}' is a $r \times p$ matrix of rank r i.e. $\widehat{\mathbf{X}}$ can be perfectly represented by $n + p$ vectors in r -dimensional space. Representing the data by means of a traditional biplot therefore in effect models the data as the sum of an inner-product, $\mathbf{a}'\mathbf{b}$, which is 'explained' by the biplot (i.e. that is contained in the biplot space, \mathcal{L}) and a residual 'error' term, ϵ , which is not explained by the

biplot (i.e. that is not contained in the biplot space):

$$\begin{aligned} x_{ij} &= \mathbf{a}'_i \mathbf{b}_j + \epsilon \\ \longrightarrow x_{ij} &= \hat{x}_{ij} + \epsilon \end{aligned} \quad (2.6.1)$$

Consider the following expression of the inner product between two vectors, \mathbf{a} and \mathbf{b} :

$$\mathbf{a}'\mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\theta_{\mathbf{a},\mathbf{b}}) \quad (2.6.2)$$

Since an inner product is defined in terms of norms and angles, it is important that the aspect ratio of the biplot plotting region be equal to one. If the aspect ratio of the plotting region is not equal to one, the inner products between the row and column markers will still give the correct approximations of the elements of \mathbf{X} but the lengths of the row and column markers and the angles between them will appear different to their true values. For this reason, conclusions drawn from the visual inspection of a biplot with an aspect ratio other than one, are likely to be incorrect. The expression of the inner product between two vectors in (2.6.2) highlights the main weakness of the traditional biplot, namely that inner products, through which the traditional biplot approximates the elements of \mathbf{X} , are difficult to visualise. However, there is information about the approximations that can easily be discerned from the mere inspection of the traditional biplot:

1. If two row markers lie in the same direction, i.e. $\mathbf{a}_i = c\mathbf{a}_m$, it implies that the corresponding two row vectors of $\widehat{\mathbf{X}}$ are proportional, i.e. $\hat{\mathbf{x}}_i = c\hat{\mathbf{x}}_m$.
2. If two column markers lie in the same direction, i.e. $\mathbf{b}_j = d\mathbf{b}_k$, it implies that the corresponding two column vectors of $\widehat{\mathbf{X}}$ are proportional, i.e. $\hat{\mathbf{x}}_j = d\hat{\mathbf{x}}_k$.
3. If a row marker and a column marker are orthogonal to each other, i.e. $\mathbf{a}'_i \mathbf{b}_j = 0$, the corresponding element of $\widehat{\mathbf{X}}$, \hat{x}_{ij} , is equal to zero.

Note that the factorization of $\widehat{\mathbf{X}}$ into two matrices of rank r is not unique - this is evident from the following expression:

$$\widehat{\mathbf{X}} = \mathbf{A}\mathbf{T}'(\mathbf{T}^{-1})'\mathbf{B}'$$

where \mathbf{T}' is a $q \times q$ non-singular matrix. Since \mathbf{T} is a non-singular matrix, the rank of $\mathbf{A}\mathbf{T}'$ is the same as the rank of \mathbf{A} and also the rank of $(\mathbf{T}^{-1})'\mathbf{B}'$ is the same as the rank of \mathbf{B}' . It follows that $\widehat{\mathbf{X}}$ can be perfectly represented in r -dimensional space by n row markers given by the n row vectors of $\mathbf{A}\mathbf{T}'$ and p column markers given by the row vectors of $\mathbf{B}\mathbf{T}^{-1}$. In order to see the effect of the non-singular transformations

performed on \mathbf{A} and \mathbf{B} on the produced biplot, consider the svd's of the matrices \mathbf{T}' and \mathbf{T}^{-1} :

$$\begin{aligned}\mathbf{T}' &= \mathbf{G}\mathbf{\Theta}\mathbf{H}' \\ \mathbf{T}^{-1} &= \mathbf{G}\mathbf{\Theta}^{-1}\mathbf{H}' .\end{aligned}$$

Substituting the svd of \mathbf{T}' for \mathbf{T}' and the svd of \mathbf{T}^{-1} for \mathbf{T}^{-1} in the expression of $\widehat{\mathbf{X}}$ gives

$$\widehat{\mathbf{X}} = \mathbf{A}\mathbf{G}\mathbf{\Theta}\mathbf{H}'\mathbf{H}\mathbf{\Theta}^{-1}\mathbf{G}'\mathbf{B}' .$$

Since $\mathbf{A}\mathbf{T}' = \mathbf{A}\mathbf{G}\mathbf{\Theta}\mathbf{H}'$, the transformation from the row markers given by the n row vectors of \mathbf{A} to the row markers given by the n row vectors of $\mathbf{A}\mathbf{T}'$ consist of a rotation and/or reflection due to \mathbf{G} , a dilation (or contraction) and possible reflection due to $\mathbf{\Theta}$ and another rotation and/or reflection of due to \mathbf{H}' . Similarly, the transformation of the column markers given by the p row vectors of \mathbf{B} to the column markers given by the p row vectors of $\mathbf{B}\mathbf{T}^{-1}$ consist of a rotation and/or reflection due to \mathbf{G} , a dilation (or contraction) and possible reflection due to $\mathbf{\Theta}^{-1}$ and another rotation and/or reflection due to \mathbf{H}' . The only difference between the transformation, $\mathbf{A} \rightarrow \mathbf{A}\mathbf{T}'$, and the transformation, $\mathbf{B} \rightarrow \mathbf{B}\mathbf{T}^{-1}$, is the dilation or contraction part of the transformation. In the transformation, $\mathbf{A} \rightarrow \mathbf{A}\mathbf{T}'$, the dilation or contraction (as well as possible reflection) is performed by the diagonal elements of $\mathbf{\Theta}$ while the diagonal elements of $\mathbf{\Theta}^{-1}$, which are the reciprocals of the diagonal elements of $\mathbf{\Theta}$, performs the dilation or contraction (as well as possible reflection) in the transformation, $\mathbf{B} \rightarrow \mathbf{B}\mathbf{T}^{-1}$. Due to the fact that the elements of a row vector of \mathbf{A} are dilated (or contracted) by different values - the j th element of a row vector being dilated (or contracted) by the j th diagonal element of $\mathbf{\Theta}$ - the angles between the row vectors of $\mathbf{A}\mathbf{T}'$ as well as the distances between the points with coordinate vectors given by the row vectors of $\mathbf{A}\mathbf{T}'$ will differ from the angles and distances corresponding to the row vectors of \mathbf{A} . Collinear row vectors of \mathbf{A} will be transformed to collinear row vectors of $\mathbf{A}\mathbf{T}'$ - only the distances between the points with coordinate vectors given by the row vectors of $\mathbf{A}\mathbf{T}'$ will differ from the corresponding distances between the points with coordinate vectors given by the row vectors of \mathbf{A} . The same can be said about the row vectors of $\mathbf{B}\mathbf{T}^{-1}$ and the row vectors of \mathbf{B} . It follows that the relationships between the row markers as well as the relationships between the column markers depend entirely on the chosen non-singular transformation. Since the lengths of the row markers, the angles between the row markers and the distances between the endpoints of the row markers are all functions of the inner products between the row markers, the mentioned lengths, angles and distances will be unaffected by a transformation that does not affect the inner products between the row markers. The inner products between the row vectors of \mathbf{A} are unaffected by the linear transformation brought

about by multiplication by the matrix \mathbf{T}' (called the transformation matrix) when

$$\mathbf{AT}'\mathbf{TA}' = \mathbf{AA}'.$$

It is evident that \mathbf{T}' will satisfy the above equation when $\mathbf{T}'\mathbf{T} = \mathbf{I}$ i.e. when the matrix \mathbf{T}' , is an orthogonal matrix. When the non-singular matrix, \mathbf{T} , is an orthogonal matrix, $\mathbf{T}'\mathbf{T} = \mathbf{I}$ and $\mathbf{TT}' = \mathbf{I}$ and hence

$$\begin{aligned} \mathbf{AT}'\mathbf{TA}' &= \mathbf{AIA}' \\ \rightarrow \mathbf{AT}'\mathbf{TA}' &= \mathbf{AA}' . \end{aligned}$$

Similarly, in order for the lengths of the column markers given by the row vectors of \mathbf{B} to be unaffected by the transformation, $\mathbf{B} \rightarrow \mathbf{BT}^{-1}$, the matrix \mathbf{T} must satisfy $\mathbf{BT}^{-1}\mathbf{T}^{-1}\mathbf{B}' = \mathbf{BB}'$. It is evident that the matrix \mathbf{T} will satisfy the above equation if and only if $(\mathbf{T}'\mathbf{T})^{-1} = \mathbf{I}$ i.e. if and only if $\mathbf{T}'\mathbf{T} = \mathbf{I}$ i.e. if and only if \mathbf{T} and hence also \mathbf{T}' is an orthogonal matrix. It follows that if the matrix of transformation \mathbf{T}' is an orthogonal matrix, then the inner products between the row vectors of \mathbf{A} as well as the inner products between the row vectors of \mathbf{B} , will be unaffected by the transformations, $\mathbf{A} \rightarrow \mathbf{AT}'$ and $\mathbf{B} \rightarrow \mathbf{BT}^{-1}$ respectively, and hence the properties of the row and column markers (namely the lengths of the markers, the angles between the markers and the distances between the endpoints of the markers) will be unaffected by the transformations. Note that when \mathbf{T} is an orthogonal matrix, $\mathbf{T}^{-1} = \mathbf{T}'$ so that the transformation performed on the column markers is given by $\mathbf{B} \rightarrow \mathbf{BT}'$. Hence, when \mathbf{T} is an orthogonal matrix, the row markers and column markers are transformed in exactly the same way. A transformation which does not affect the properties of the row and column markers is desired when the row and column markers are such that certain aspects of interest are approximated in the produced biplot. If the row and column markers given by the row vectors of \mathbf{A} and \mathbf{B} respectively produce a biplot with desirable properties in that certain characteristics of interest are approximated in the biplot, then only transformations for which the relationships between the row markers and the relationships between the column markers are unaffected, should be performed, that is only orthogonal transformations should be performed. It is known that an orthogonal transformation (i.e. multiplication by an orthogonal matrix) results in a rotation and/or a reflection being performed. It follows that performing the same rotation or reflection on the row and column markers of an existing PCA biplot will not change the approximation of the data matrix produced by the biplot or the properties of the row and column markers. In order to see this, let \mathbf{Q} denote an orthogonal matrix and $\mathbf{A}^* = \mathbf{AQ}'$ and $\mathbf{B}^* = \mathbf{BQ}'$, then:

$$\widehat{\mathbf{X}} = \mathbf{AB}'$$

$$\begin{aligned}
&\longrightarrow \widehat{\mathbf{X}} = \mathbf{A}\mathbf{Q}'\mathbf{Q}\mathbf{B}' \\
&\longrightarrow \widehat{\mathbf{X}} = \mathbf{A}^*\mathbf{B}^{*'} \\
\mathbf{A}^*\mathbf{A}^{*'} &= \mathbf{A}\mathbf{Q}'\mathbf{Q}\mathbf{A}' \\
&= \mathbf{A}\mathbf{A}' \\
\mathbf{B}^*\mathbf{B}^{*'} &= \mathbf{B}\mathbf{Q}'\mathbf{Q}\mathbf{B}' \\
&= \mathbf{B}\mathbf{B}' .
\end{aligned}$$

Since \mathbf{Q} is an orthogonal matrix, multiplication by \mathbf{Q} performs a reflection and/or a rotation. If $|\mathbf{Q}| = 1$, multiplication by \mathbf{Q} performs a rotation while if $|\mathbf{Q}| = -1$, multiplication by \mathbf{Q} performs either a reflection only or a reflection and a rotation. Note that since \mathbf{Q} is a non-singular matrix, the rank of $\mathbf{A}^* = \mathbf{A}\mathbf{Q}'$ is the same as that of \mathbf{A} and the rank of $\mathbf{B}^{*'} = \mathbf{Q}\mathbf{B}'$ is the same as that of \mathbf{B} .

In order for the biplot to be used to represent the true relationships between the samples in the measurement space or the true relationships between the variables in the measurement space, specific constraints need to be imposed on the row and column markers. If for example the relationships between the samples in the full measurement space are to be represented by the same relationships between the corresponding row markers in the biplot, it must be true that

$$\widehat{\mathbf{X}}\widehat{\mathbf{X}}' = \mathbf{A}\mathbf{A}' . \quad (2.6.3)$$

Generally however, $\widehat{\mathbf{X}}\widehat{\mathbf{X}}' = \mathbf{A}\mathbf{B}'\mathbf{B}\mathbf{A}'$. In order for equation (2.6.3) to hold, the matrix of column markers, \mathbf{B} , must therefore satisfy

$$\mathbf{B}'\mathbf{B} = \mathbf{I}$$

that is, \mathbf{B} must be an orthogonal matrix. Similarly, when the relationships between the variables in the measurement space are to be represented by the same relationships between the corresponding columns markers, the following must be true:

$$\widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \mathbf{B}\mathbf{B}' . \quad (2.6.4)$$

Given that generally, $\widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \mathbf{B}\mathbf{A}'\mathbf{A}\mathbf{B}'$, it is evident that the appropriate constraint to be imposed in order for equation (2.6.4) to hold, is given by

$$\mathbf{A}'\mathbf{A} = \mathbf{I}$$

which implies that \mathbf{A} must be an orthonormal matrix. Note that when $\mathbf{B}'\mathbf{B} = \mathbf{I}$, the inner products between the column vectors of \mathbf{X} are represented by the inner products between the column markers in the metric, $\mathbf{A}'\mathbf{A}$:

$$\widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \mathbf{B}\mathbf{A}'\mathbf{A}\mathbf{B}' .$$

Similarly, when $\mathbf{A}'\mathbf{A} = \mathbf{I}$, the inner products between the row vectors of \mathbf{X} are represented by the inner products between the row markers in the metric, $\mathbf{B}'\mathbf{B}$:

$$\widehat{\mathbf{X}}\widehat{\mathbf{X}}' = \mathbf{A}\mathbf{B}'\mathbf{B}\mathbf{A}' .$$

Since the aim is to represent \mathbf{X} as well as possible in the r -dimensional display space, the rank r approximation, $\widehat{\mathbf{X}}$, to \mathbf{X} chosen to represent \mathbf{X} in the r -dimensional display space must be such that some function of the deviations between the elements of \mathbf{X} and the elements of $\widehat{\mathbf{X}}$ are minimised. An example of such a function, which is not only mathematically tractable but also a logical choice, is the function used by Gabriel (1971) in the construction of the traditional biplot, namely the sum of squared residuals,

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \hat{x}_{ij})^2 &= \text{tr} \{ (\mathbf{X} - \widehat{\mathbf{X}}) (\mathbf{X} - \widehat{\mathbf{X}})' \} \\ &= \|\mathbf{X} - \widehat{\mathbf{X}}\|^2 . \end{aligned}$$

If the sum of the squared residuals is taken as the measure of lack of fit, then the matrix which will most accurately represent \mathbf{X} in r -dimensional space is the best least squares rank r approximation of \mathbf{X} . It follows that the r -dimensional traditional biplot of a centred data matrix \mathbf{X} , is an r -dimensional joint representation of the rows (samples) and columns (variables) of \mathbf{X} which perfectly reproduces the best least squares rank r approximation, $\widehat{\mathbf{X}} = \mathbf{A}\mathbf{B}'$, of \mathbf{X} via inner-products between the row and column markers given by the row vectors of the matrices \mathbf{A} and \mathbf{B} respectively. Consider the svd of \mathbf{X} :

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' .$$

According to the Eckart-Young theorem, the best least squares rank r approximation

of \mathbf{X} is given by

$$\widehat{\mathbf{X}} = \mathbf{U}_r \mathbf{D}_r \mathbf{V}_r'.$$

Recall from Chapter 1 that the best rank r approximation to \mathbf{X} can be expressed as the orthogonal projection of \mathbf{X} onto $\mathcal{V}(\mathbf{V}_r)$:

$$\widehat{\mathbf{X}} = \mathbf{U}_r \mathbf{D}_r \mathbf{V}_r' = \mathbf{X} \mathbf{V}_r \mathbf{V}_r'.$$

Since the row vectors of $\widehat{\mathbf{X}}$ are the orthogonal projections of the corresponding row vectors of \mathbf{X} onto the column space of \mathbf{V}_r , the row vectors of the matrix of residuals, $\mathbf{X} - \widehat{\mathbf{X}}$, are contained in the orthogonal complement of the column space of \mathbf{V}_r . This means that the r -dimensional traditional biplot space is identical to the column space of the matrix \mathbf{V}_r and the ϵ -term in equation (2.6.1) is contained in the orthogonal complement of the biplot space.

The matrix $\widehat{\mathbf{X}}$, which is of rank r , can be expressed as the inner product of two rank r matrices, \mathbf{A} and \mathbf{B}' , in the following way:

$$\begin{aligned} \widehat{\mathbf{X}} &= \mathbf{U}_r \mathbf{D}_r \mathbf{V}_r' \\ \longrightarrow \widehat{\mathbf{X}} &= (\mathbf{U}_r \mathbf{D}_r^\alpha) (\mathbf{D}_r^{1-\alpha} \mathbf{V}_r') \\ \longrightarrow \widehat{\mathbf{X}} &= \mathbf{A} \mathbf{B}' \end{aligned}$$

where $\mathbf{A} = \mathbf{U}_r \mathbf{D}_r^\alpha$, $\mathbf{B}' = \mathbf{D}_r^{1-\alpha} \mathbf{V}_r'$ and alpha is a scale parameter, $0 \leq \alpha \leq 1$ (see Section 1.6.5). It will be explained that as the value of α moves from zero to one, the focus in the biplot is shifted from the representation of the columns of \mathbf{X} to the rows of \mathbf{X} . It is evident that the approximation to \mathbf{X} that is produced by the biplot with row markers given by the row vectors of \mathbf{A} and column markers given by the row vectors of \mathbf{B} , is exactly the same for every possible value of α , $0 \leq \alpha \leq 1$. The following approximations can be obtained from the r -dimensional traditional biplot given any value of α :

1. the individual elements of \mathbf{X} : $x_{ij} \sim \mathbf{a}_i \mathbf{b}_j$;
2. the difference in a specific sample's measurements on two different variables:
 $x_{ij} - x_{ik} \sim \mathbf{a}_i' (\mathbf{b}_j - \mathbf{b}_k)$;
3. the difference between two different samples' measurements on the same variable: $x_{ij} - x_{mj} \sim (\mathbf{a}_i - \mathbf{a}_m)' \mathbf{b}_j$, and
4. the interaction of two samples with two variables: $x_{ij} - x_{mj} - x_{ik} + x_{mk} \sim (\mathbf{a}_i - \mathbf{a}_m)' (\mathbf{b}_j - \mathbf{b}_k)$.

The biplots corresponding to $\alpha = 0$ or $\alpha = 1$ are very useful when it comes to interpretation since in each of these biplots certain aspects of \mathbf{X} are approximated as well as possible. When $\alpha = 1$, the row markers are given by the row vectors of $\mathbf{A} = \mathbf{U}_r \mathbf{D}_r$ and the column markers are given by the column vectors of $\mathbf{B}' = \mathbf{V}_r'$. Hence, \mathbf{A} and \mathbf{B} satisfies the following equations:

$$\begin{aligned} \mathbf{A}\mathbf{A}' &= \mathbf{U}_r \mathbf{D}_r^2 \mathbf{U}_r' \\ \text{and } \mathbf{B}'\mathbf{B} &= \mathbf{V}_r' \mathbf{V}_r = \mathbf{I}_r. \end{aligned}$$

It is evident that $\alpha = 1$ ensures that the relationships between the samples in the full measurement space are represented by the same relationships between the corresponding row markers. This implies that the Pythagorean distances between the endpoints of the row markers of the r -dimensional biplot will represent the corresponding Pythagorean distances between the points with coordinate vectors given by the row vectors of \mathbf{X} .

Note that the elements of the i th row of $\mathbf{U}_r \mathbf{D}_r$ are equal to the first r principal component scores of the i th row vector of \mathbf{X} and that the elements of the j th column of \mathbf{V}_r' are equal to the j th coefficients of the first r principal components:

$$\mathbf{X}\mathbf{V}_r = \mathbf{U}\mathbf{D}\mathbf{V}\mathbf{V}_r = \mathbf{U}_r \mathbf{D}_r.$$

Note that the i th row of $\mathbf{X}\mathbf{V}_r = \mathbf{U}_r \mathbf{D}_r$ gives the coordinates of the point $\mathbf{x}_i' \mathbf{V}_r \mathbf{V}_r$, that is the orthogonal projection of \mathbf{x}_i' onto $\mathcal{V}(\mathbf{V}_r)$, with respect to the column vectors of \mathbf{V}_r . It follows that the r -dimensional PCA biplot space is spanned by the coefficient vectors of the first r principal components. For the reasons mentioned above, the biplot produced by $\alpha = 1$ is called the Principal Component Analysis biplot, or PCA biplot for short. Note that the plot containing only the row markers is the r -dimensional MDS configuration corresponding to a PCA performed on \mathbf{X} .

Recall that the column space of \mathbf{V}_r is the best fitting r -dimensional subspace to the configuration of points, $\{\mathbf{x}_i\}$. It follows that the PCA biplot represents the samples as accurately as possible, in that it yields the smallest possible sum of squared residuals between the samples in the measurement space, \mathbb{R}^p , and the points representing the samples in the biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$, and represents the relationships between the samples as accurately as possible in that it minimises the sum of the differences between the squared Pythagorean intersample distances in the measurement space and the corresponding squared fitted Pythagorean intersample distances in the biplot space. It follows that in addition to the four aspects of the data set which are approximated in any traditional biplot, the traditional PCA biplot can also be used to approximate the Pythagorean distances between the samples in the measurement space.

Since the points representing the samples in the traditional biplot have coordinates given by the principal component scores of these samples and the traditional

biplot space is spanned by the coefficient vectors of the principal components, it is clear that the traditional biplot with $\alpha = 1$ is scale dependent. Recall that PCA should be performed on the standardised measurements of a data set when the measured variables have widely differing standard deviations. By the same argument, the traditional PCA biplot should be constructed from the standardised data matrix when the standard deviations of the measured variables differ substantially.

The traditional PCA biplot will now be illustrated at the hand of the *University* data set. Consider the standard deviations of the *University* data set provided in Table 2.1.

Table 2.1: *The standard deviations of the measured variables of the University data set.*

SAT	Top10	Accept	SFRatio	Expenses	Grad
1.084	19.434	19.727	4.067	14.425	9.058

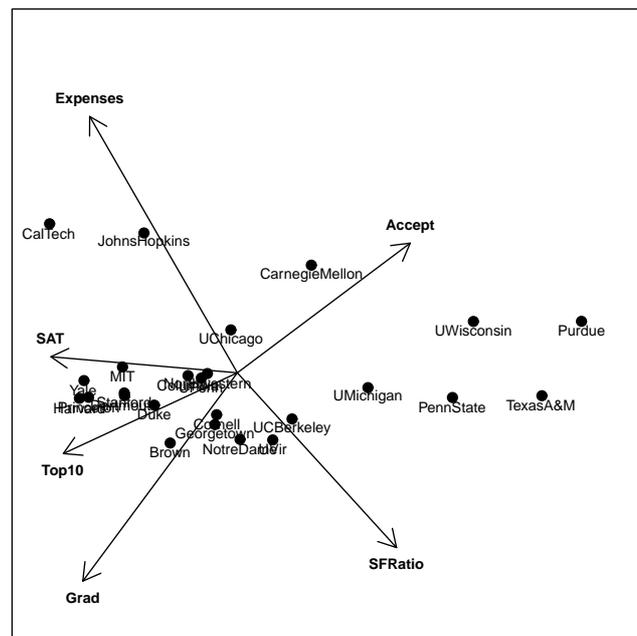


Figure 2.2: *The two-dimensional traditional PCA biplot (i.e. $\alpha = 1$) constructed from the standardised measurements of the University data set.*

Due to the fact that the standard deviations of the measured variables of the *University* data set differ substantially, the two-dimensional traditional PCA biplot of the *University* data set, which is provided in Figure 2.2, was constructed from the standardised measurements of the data set. Note that in Figure 2.2 the vectors

representing the measured variables have been extended to be more clearly visible.

When $\alpha = 0$, the row markers of the traditional biplot are given by the row vectors of the matrix $\mathbf{A} = \mathbf{U}_r$ and the column markers are given by the column vectors of the matrix $\mathbf{B}' = \mathbf{D}_r \mathbf{V}'_r$. The matrices \mathbf{A} and \mathbf{B} therefore satisfy the equations,

$$\begin{aligned} \mathbf{A}'\mathbf{A} &= \mathbf{U}'_r \mathbf{U}_r = \mathbf{I}_r \\ \text{and } \mathbf{B}\mathbf{B}' &= \mathbf{V}_r \mathbf{D}_r^2 \mathbf{V}'_r. \end{aligned}$$

Since $\alpha = 0$ corresponds to the constraint $\mathbf{A}'\mathbf{A} = \mathbf{I}$, $\alpha = 0$ ensures that the relationships between the variables in the full measurement space are represented by the same relationships between the corresponding column markers in the biplot space, \mathcal{L} . This implies that the Pythagorean distances between the endpoints of the column markers in the biplot represent the corresponding Pythagorean distances between the points with coordinate vectors given by the column vectors of \mathbf{X} . Note that the j th column marker (i.e. the j th row of $\mathbf{V}_r \mathbf{D}_r$) gives the coordinates of the orthogonal projection of the j th column of \mathbf{X} , $\mathbf{x}_{(j)}$, onto the column space of \mathbf{U}_r , in terms of the column vectors of \mathbf{U}_r . Note that the column space of \mathbf{U}_r is the best fitting r -dimensional linear subspace to the centred matrix, \mathbf{X}' . This implies that the Pythagorean distances between the points with coordinate vectors given by the column vectors of \mathbf{X} are optimally approximated by the Pythagorean distances between the endpoints of the column markers (see Section 1.6.11). Since the column vectors of \mathbf{U}_r are the first r right singular vectors of \mathbf{X}' , they give the directions of the first r principal axes of the configuration of points with coordinate vectors given by the row vectors of \mathbf{X}' and hence define the first r principal components corresponding to a PCA performed on the matrix, \mathbf{X}' . The elements of the j th column marker are therefore given by the first r principal component scores corresponding to the j th row of \mathbf{X}' , or equivalently the j th column of \mathbf{X} .

It can be shown that the Pythagorean distance between the i th and j th row markers in the p -dimensional traditional biplot with $\alpha = 0$ is proportional to the (sample) Mahalanobis distance between the points, \mathbf{x}_i and \mathbf{x}_j , the constant of proportionality equaling \sqrt{n} . Consider the squared Mahalanobis distance between \mathbf{x}_i and \mathbf{x}_j :

$$\begin{aligned} (\mathbf{x}_i - \mathbf{x}_j)' \widehat{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j) &= (\mathbf{x}_i - \mathbf{x}_j)' \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1} (\mathbf{x}_i - \mathbf{x}_j) \\ &= n (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{V} \mathbf{D}^{-2} \mathbf{V}' (\mathbf{x}_i - \mathbf{x}_j) \\ &= n (\mathbf{x}'_i \mathbf{V} \mathbf{D}^{-1} - \mathbf{x}'_j \mathbf{V} \mathbf{D}^{-1}) (\mathbf{x}'_i \mathbf{V} \mathbf{D}^{-1} - \mathbf{x}'_j \mathbf{V} \mathbf{D}^{-1})' \\ &= n (\mathbf{u}'_i \mathbf{D} \mathbf{V}' \mathbf{V} \mathbf{D}^{-1} - \mathbf{u}'_j \mathbf{D} \mathbf{V}' \mathbf{V} \mathbf{D}^{-1}) (\\ &\quad \mathbf{u}'_i \mathbf{D} \mathbf{V}' \mathbf{V} \mathbf{D}^{-1} - \mathbf{u}'_j \mathbf{D} \mathbf{V}' \mathbf{V} \mathbf{D}^{-1})' \\ &= n (\mathbf{u}_i - \mathbf{u}_j)' (\mathbf{u}_i - \mathbf{u}_j) \end{aligned}$$

$$= n \sum_{k=1}^n (u_{ik} - u_{jk})^2 .$$

Since the squared Pythagorean distance between the i th and j th row markers of the r -dimensional biplot with $\alpha = 0$ is given by

$$\sum_{k=1}^r (u_{ik} - u_{jk})^2 ,$$

it is proportional to an approximation of the squared Mahalanobis distance between \mathbf{x}_i and \mathbf{x}_j . The relative magnitudes of the intersample Pythagorean distances between the row markers of the r -dimensional biplot therefore approximates the relative magnitudes of the Mahalanobis distances between the samples in the measurement space.

When the biplot is constructed from the matrix of unstandardised (and centred) measurements, $\mathbf{X} = \mathbf{UDV}'$, the inner product matrix, $\mathbf{X}'\mathbf{X} = \mathbf{VD}^2\mathbf{V}'$, is proportional to the sample covariance matrix corresponding to the vector of measured variables, $\widehat{\Sigma}$. Hence the inner product between the i th and j th column marker of the p -dimensional biplot corresponding to $\alpha = 0$ is proportional to the sample covariance between the i th and j th measured variables and the squared length of the i th column marker is proportional to the sample variance of the i th measured variable:

$$\begin{aligned} \mathbf{BB}' &= \mathbf{VD}^2\mathbf{V} \\ \rightarrow \mathbf{BB}' &= c\widehat{\Sigma} \\ \rightarrow \mathbf{b}'_i\mathbf{b}_j &= \begin{cases} c\hat{\sigma}_{ij} & \text{if } i \neq j \\ c\hat{\sigma}_{ii} & \text{if } i = j \end{cases} . \end{aligned}$$

It follows that the relative magnitudes of the lengths of the column markers equal the relative magnitudes of the standard deviations of the measured variables. The p -dimensional biplot corresponding to $\alpha = 0$ also perfectly represents the sample correlations between the measured variables - the cosine of the angle between two column markers equals the sample correlation between the corresponding two measured variables:

$$\begin{aligned} \cos(\theta_{\mathbf{b}_{i,j}}) &= \frac{\mathbf{b}'_i\mathbf{b}_j}{\sqrt{\mathbf{b}'_i\mathbf{b}_i}\sqrt{\mathbf{b}'_j\mathbf{b}_j}} \\ &= \frac{[c\widehat{\Sigma}]_{ij}}{\sqrt{[c\widehat{\Sigma}]_{ii}}\sqrt{[c\widehat{\Sigma}]_{jj}}} \end{aligned}$$

$$\begin{aligned}
 &= \frac{[\widehat{\Sigma}]_{ij}}{\sqrt{[\widehat{\Sigma}]_{ii}}\sqrt{[\widehat{\Sigma}]_{jj}}} \\
 \longrightarrow \cos(\theta_{\mathbf{b}_{i,j}}) &= r_{ij}.
 \end{aligned}$$

The size of the angle between two column markers is therefore a decreasing function of the sample correlation coefficient between the corresponding two measured variables. In the r -dimensional biplot corresponding to $\alpha = 0$, the i th column marker is given by the i th row of $\mathbf{V}_r \mathbf{D}_r^2$ and hence the inner product between the i th and j th column markers is given by the (ij) th element of the matrix $\widehat{\mathbf{X}}' \widehat{\mathbf{X}} = \mathbf{V}_r \mathbf{D}_r^2 \mathbf{V}_r'$, which is the best rank r approximation of $\mathbf{V} \mathbf{D}_p^2 \mathbf{V}' = c \widehat{\Sigma}$. It follows that the inner product between the i th and j th column markers of the r -dimensional biplot corresponding to $\alpha = 0$ is proportional to an approximation of the sample covariance between the i th and j th measured variables. Similarly, the squared length of the i th column marker is proportional to an approximation of the sample variance of the i th measured variable. The relative magnitudes of the lengths of the column markers therefore represent the relative magnitudes of the approximations to the sample standard deviations of the measured variables produced by the biplot. Since this biplot approximates the elements of $\widehat{\Sigma} = c \mathbf{S}$ as well as possible in terms of least squares, approximations to the elements of the sample correlation matrix, \mathbf{R} , can be obtained from this biplot in the following way:

$$\begin{aligned}
 \hat{r}_{ij} &= \frac{\mathbf{b}'_i \mathbf{b}_j}{\sqrt{\mathbf{b}'_i \mathbf{b}_i} \sqrt{\mathbf{b}'_j \mathbf{b}_j}} \\
 \longrightarrow \hat{r}_{ij} &= \cos(\theta_{\mathbf{b}_{i,j}}).
 \end{aligned}$$

It is evident that the cosine of the angle between the i th and j th column markers approximates the correlation coefficient between the i th and j th measured variable. Note however that the approximations to the elements of \mathbf{R} which are obtained in this way are not equal to the elements of the best rank- r approximation to \mathbf{R} in terms of least squares.

When the biplot is constructed from the matrix of standardised measurements, $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}'$, the inner product matrix, $\mathbf{X}' \mathbf{X} = \mathbf{V} \mathbf{D}_p^2 \mathbf{V}'$, is proportional to the sample correlation matrix, \mathbf{R} , associated with the vector of measured variables:

$$\begin{aligned}
 \mathbf{R} &= c \mathbf{X}' \mathbf{X} \\
 &= c \mathbf{V} \mathbf{D}_p^2 \mathbf{V}'.
 \end{aligned}$$

It follows that the inner product between the i th and j th column markers of the p -dimensional biplot corresponding to $\alpha = 0$ is proportional to the sample correlation

between the i th and j th measured variables:

$$\begin{aligned} \mathbf{BB}' &= \mathbf{VD}^2\mathbf{V} \\ \longrightarrow \mathbf{BB}' &= c\mathbf{R} \\ \longrightarrow \mathbf{b}'_i\mathbf{b}_j &= \begin{cases} cr_{ij} & \text{if } i \neq j \\ c & \text{if } i = j \end{cases} . \end{aligned}$$

The cosine of the angle between the i th and j th column markers of the p -dimensional biplot corresponding to $\alpha = 0$ and constructed from the standardised measurements is equal to the sample correlation between the i th and j th measured variables:

$$\begin{aligned} \cos(\theta_{\mathbf{b}_{i,j}}) &= \frac{\mathbf{b}'_i\mathbf{b}_j}{\sqrt{\mathbf{b}'_i\mathbf{b}_i}\sqrt{\mathbf{b}'_j\mathbf{b}_j}} \\ &= \frac{[\mathbf{cR}]_{ij}}{\sqrt{[\mathbf{cR}]_{ii}}\sqrt{[\mathbf{cR}]_{jj}}} \\ \longrightarrow \cos(\theta_{\mathbf{b}_{i,j}}) &= r_{ij} . \end{aligned}$$

The inner product between the i th and j th column markers in the r -dimensional biplot corresponding to $\alpha = 0$ is equal to the (ij) th element of the matrix, $\widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \mathbf{V}_r\mathbf{D}_r^2\mathbf{V}'_r$, which is the best rank r approximation of the matrix $\mathbf{X}'\mathbf{X} = \mathbf{VD}_p^2\mathbf{V}' = c\mathbf{R}$. It follows that the inner product between the i th and j th column markers in the r -dimensional biplot is proportional to an approximation of the sample correlation between the i th and j th measured variables while the cosine of the angle between the two column markers approximates the sample correlation.

Note that when $\alpha = 1$ and the biplot is constructed from the unstandardised (or standardised) measurements, the inner product of the i th and j th column markers of the p -dimensional biplot is not proportional to the sample covariance (or correlation) between the i th and j th measured variables and hence the cosine of the angle between the two column markers will not equal the sample correlation between the i th and j th measured variables. Consequently, the cosine of the angle between two column markers of the r -dimensional biplot will not approximate the sample correlation coefficient between the corresponding two measured variables.

The two-dimensional traditional biplot constructed from standardised measurements of the *University* data set and with $\alpha = 0$, is provided in Figure 2.3. Visual inspection of the biplot in Figure 2.3 suggests a weak positive linear relationship between the variables *Expenses* and *Grad* as well as a very weak negative linear relationship between the variables *Grad* and *SFRatio*. The sample correlation matrix of the *University* data set provided in Table 2.2 confirms the weak positive linear relationship between the variables *Expenses* and *Grad* (correlation of 0.394) but indicates a relatively strong negative linear relationship between the variables

Grad and *SFRatio* (correlation of -0.561). The reason for the poor representation of the linear relationship between the variables *Grad* and *SFRatio* is the loss of information due to the reduction in the dimensionality of the data from six (the true dimensionality of the data) to two (the dimensionality of the biplot space). The traditional biplot also suggests very strong positive linear relationship between the variable *Top10* and each of the variables *SAT* and *Grad* and a very strong negative linear relationship between the variable *Accept* and each the variables *Top10* and *Grad* as well as a strong negative linear relationship between the variables *Expenses* and *SFRatio*. The sample correlation matrix in Table 2.2 confirms the signs and strengths of these linear relationships.

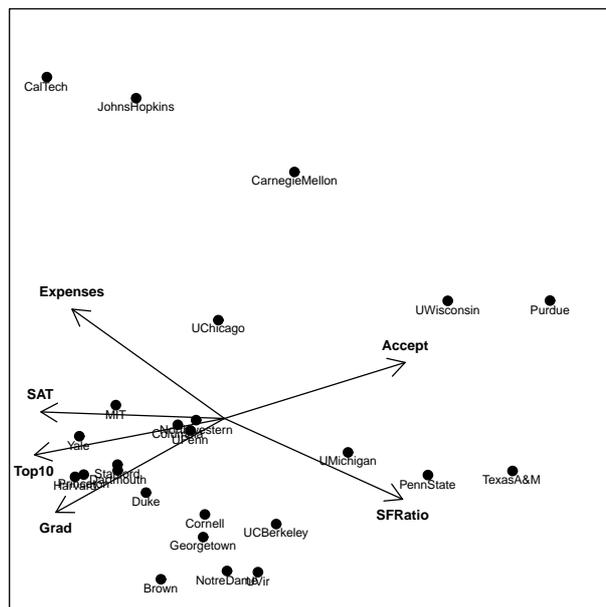


Figure 2.3: The two-dimensional traditional biplot constructed from the standardised measurements of the University data set with $\alpha = 0$.

Table 2.2: The sample correlation matrix associated with the University data set.

	SAT	Top10	Accept	SFRatio	Expenses	Grad
SAT	1.0000	0.923	-0.886	-0.813	0.779	0.748
Top10	0.923	1.000	-0.859	-0.643	0.611	0.746
Accept	-0.886	-0.859	1.000	0.632	-0.558	-0.820
SFRatio	-0.813	-0.643	0.632	1.000	-0.782	-0.561
Expenses	0.779	0.611	-0.558	-0.782	1.000	0.394
Grad	0.748	0.746	-0.820	-0.561	0.394	1.000

Upon inspection of the traditional PCA biplot in Figure 2.2, it is evident that the angles between the column markers convey very misleading information regarding the correlations between at least some of the variables if the angles are interpreted as being indicative of the strengths and signs of the linear relationships between the corresponding variables. For example, the angle between the column markers representing the variables *Accept* and *Expenses* seems to indicate a weak positive linear relationship between the two variables, whereas in truth the linear relationship is a relatively strong negative one. Also, the angle between the column markers corresponding to *Expenses* and *Grad* seem to indicate a relatively weak negative linear relationship between the two variables whereas in truth the linear relationship between the variables is a relatively weak positive one. The angle between the column markers corresponding to *Grad* and *SFRatio* seem to indicate a relatively weak positive linear relationship between the two variables whereas in reality it is a relatively strong negative linear relationship. The angle between the column markers representing the variables *Accept* and *SFRatio* seems to indicate that these variables are almost uncorrelated while in reality the two variables have a relatively strong positive linear relationship. This is evidence of the fact that the cosines of the angles between the column markers of the traditional PCA biplot (i.e. $\alpha = 1$) do not approximate the correlation coefficients between the corresponding variables.

In conclusion, when $\alpha = 1$, the rows (samples) of \mathbf{X} and the relationships between the rows of \mathbf{X} are represented as well as possible while when $\alpha = 0$, the columns (variables) of \mathbf{X} and the relationships between the columns of \mathbf{X} are represented as well as possible. Naturally, when $\alpha > 0.5$, the rows of \mathbf{X} will be better represented in the biplot than the columns of \mathbf{X} while $\alpha < 0.5$ will result in the columns of \mathbf{X} being better represented in the biplot than the rows of \mathbf{X} . It is evident that as the value of α moves from zero to one, the focus is shifted from the accurate representation of the columns of \mathbf{X} to the accurate representation of the rows of \mathbf{X} . The value of α therefore determines the properties of the PCA biplot with respect to which aspect of the data set are most accurately represented in the biplot. Given the value of α , the biplot is however still not unique - performing the same orthogonal transformation on the matrix of row markers and the matrix of column markers, produces a biplot with exactly the same properties as the original.

2.7 The biplot proposed by Gower and Hand (1996)

The main weakness of the traditional biplot is the difficulty of visualising inner products and hence the difficulty of visualising the approximations to the elements of the data matrix represented in the biplot. To address this difficulty Gower and Hand (1996) proposed that in the biplot the variables be represented by calibrated axes that are such that the approximated measurements of a sample can be read off from the axes as in an ordinary scatter plot. These axes are formed by extending the vectors representing the variables in the traditional biplot in both directions to the edges of the biplot. Biplots constructed in this way can therefore be viewed as multivariate analogues of ordinary scatter plots (Gower and Hand, 1996). The cali-

brated axes just described are referred to as predictive biplot axes and accordingly the PCA biplot in which the variables are represented by the predictive biplot axes is called a predictive PCA biplot. Another set of axes, referred to as the interpolative biplot axes, is used to position the points representing the samples in the biplot space. The PCA biplot in which the variables are represented by the interpolative biplot axes is called the interpolative PCA biplot. The interpolation and prediction processes as well as the construction of the interpolative and predictive PCA biplot axes will be discussed in more detail in Section 2.7.1.

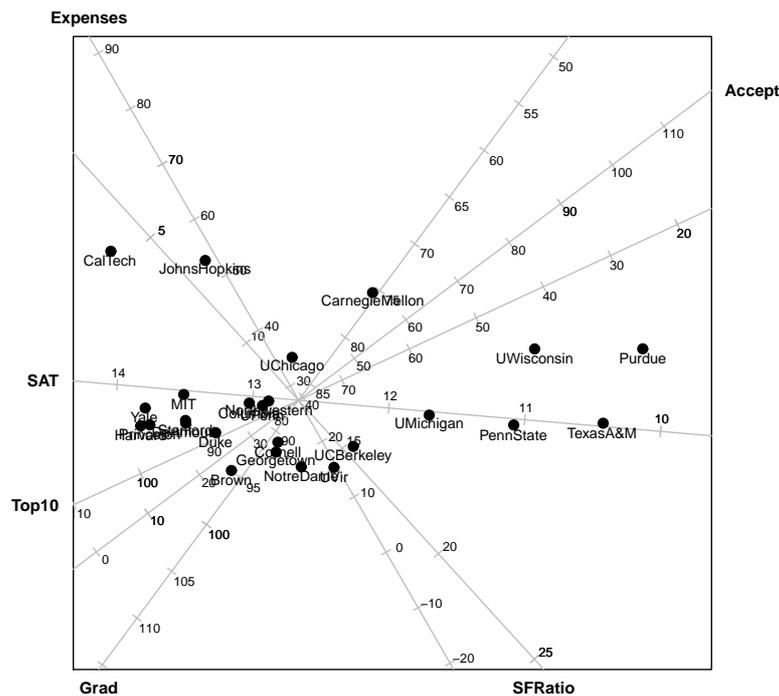


Figure 2.4: *The two-dimensional predictive PCA biplot constructed from the standardised measurements of the University data set.*

The biplots that are obtained by making the adjustments proposed by Gower and Hand (1996) to the traditional biplots constructed with $\alpha = 1$ and $\alpha = 0$ are respectively called the PCA biplot and the correlation biplot. Since the first r right singular vectors of \mathbf{X} form an orthogonal basis for the r -dimensional PCA biplot space, they define a set of orthogonal coordinate axes that can be used as scaffolding for the construction of the PCA biplot. Similarly the first r left singular vectors of \mathbf{X} form an orthogonal basis for the r -dimensional correlation biplot and hence can be used as scaffolding for the construction of the correlation biplot. However, since it is the approximation of the elements of \mathbf{X} which is of interest, the scaffolding is not displayed in either of the biplots. The differences between the traditional PCA biplot proposed by Gabriel (1971) and the PCA biplot proposed by Gower and Hand (1996) are evident upon comparison of Figure 2.4 in which the two-dimensional predictive

PCA biplot of the *University* data set is illustrated and Figure 2.2.

The predictive correlation biplot of the *University* data set is provided in Figure 2.5. Since the biplot axes of the correlation biplot and the column markers in the corresponding traditional biplot constructed with $\alpha = 0$ are collinear, the interpretation of the angles between the biplot axes in the correlation biplot is identical to the interpretation of the angles between the vectors representing the measured variables in the traditional biplot.

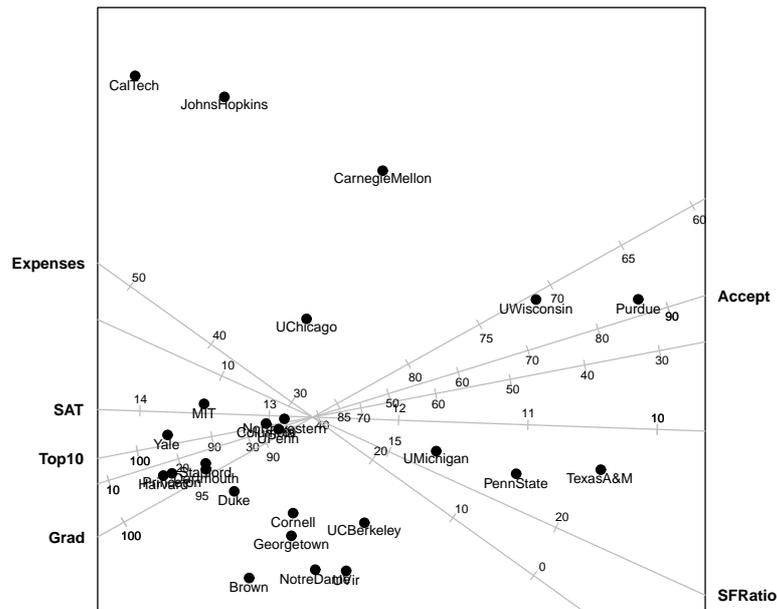


Figure 2.5: *The two-dimensional predictive correlation biplot constructed from the standardised measurements of the University data set.*

In the rest of this thesis biplots will be constructed as proposed by Gower and Hand (1996).

2.7.1 The construction of the PCA biplot

In the next two sections it will be explained how to position the observed samples in the PCA biplot space using a set of linear axes called the interpolative biplot axes, how to obtain the predicted measurements of the samples using a set of linear axes called the predictive biplot axes as well as how to construct these two sets of biplot axes.

2.7.1.1 Interpolation and the interpolative biplot

Interpolation is the process of finding the position of a sample in the biplot space, \mathcal{L} , given the sample's measurements on the original variables. Interpolation is per-

formed by relating the given values to a set of biplot axes referred to as interpolative biplot axes. Accordingly the joint plot in which the measured variables are represented by the interpolative biplot axes and the samples are represented by the points the positions of which are found using the interpolative biplot axes, is called the interpolative PCA biplot.

For convenience the interpolative biplot axis corresponding to the k th measured variable will be referred to as the k th interpolative biplot axis. Let \mathbf{X}^* denote the original observed data matrix and \mathbf{X} denote the matrix upon which the construction of the PCA biplot is based. Let \bar{x}_j^* and $\hat{\sigma}_{jj}^*$, respectively denote the sample mean and sample variance of the j th measured variable calculated from the matrix \mathbf{X}^* , $j \in [1:p]$. Recall that when the measured variables have widely differing standard deviations, the PCA biplot should be based on the centred and standardised data matrix, i.e. $\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}^*\mathbf{A}^{-1}$, where \mathbf{A} denotes the $p \times p$ diagonal matrix with j th diagonal element equal to $\sqrt{\hat{\sigma}_{jj}^*}$, $j \in [1:p]$. On the other hand, when the measured variables have very similar standard deviations, the PCA biplot can be constructed from the centred but unstandardised measurements i.e. $\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}^*$.

Consider the svd of \mathbf{X} :

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' .$$

Recall that the point that represents the i th sample in the r -dimensional PCA biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$, is the orthogonal projection of \mathbf{x}_i onto the biplot space. This point in \mathcal{L} is referred to as the interpolant of \mathbf{x}_i . The coordinate vector of the interpolant of \mathbf{x}_i in terms of the basis of the p -dimensional measurement space given by the p p -dimensional unit vectors $\{\mathbf{e}_k\}_{k=1}^p$, is given by

$$\mathbf{x}_i'\mathbf{V}_r\mathbf{V}_r' .$$

The coordinate vector of this point in terms of the basis of \mathcal{L} given by the column vectors of \mathbf{V}_r is given by

$$\mathbf{x}_i'\mathbf{V}_r .$$

A method called the ‘vector sum method’ can be used to find the position of the interpolant of a sample \mathbf{x} in the biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$. This method relies on the following expression of the interpolant $\mathbf{x}'\mathbf{V}_r$:

$$\mathbf{x}'\mathbf{V}_r = p \frac{1}{p} \sum_{k=1}^p x_k (\mathbf{e}_k'\mathbf{V}_r) .$$

The ‘vector sum method’ entails locating the centroid of the p points $x_1\mathbf{e}'_1\mathbf{V}_r$, $x_2\mathbf{e}'_2\mathbf{V}_r$, ... and $x_p\mathbf{e}'_p\mathbf{V}_r$, and then extending the vector stretching from the origin to this centroid p times - the endpoint of this extended vector gives the position of the interpolant $\mathbf{x}'\mathbf{V}_r$.

Note that $\mathbf{e}'_k\mathbf{V}_r$ is the interpolant of the unit point \mathbf{e}_k on the k th Cartesian axis. The interpolant $\mathbf{e}'_k\mathbf{V}_r$ therefore represents one unit of the k th variable, x_k . Remember that this one unit is in terms of the scale of the elements of the k th column of the matrix \mathbf{X} . Similarly, the interpolant $\mu\mathbf{e}'_k\mathbf{V}_r$, where μ is an arbitrary constant, represents μ units of the k th variable, x_k . The k th interpolative axis is defined by points of the form $\mu\mathbf{e}'_k\mathbf{V}_r$. This means that the k th interpolative axis is linear and lies collinear to the k th row vector of \mathbf{V}_r which stretches from the origin. It follows that if $\mathbf{e}_k \in \mathcal{V}(\mathbf{V}_r)$, the k th interpolative biplot axis of the r -dimensional PCA biplot and the Cartesian axis that represents the k th variable in the p -dimensional measurement space, will be collinear. If the interpolative biplot axes are calibrated in the same scales as the elements of \mathbf{X} then point $\mu\mathbf{e}'_k\mathbf{V}_r$ is calibrated with the value μ such that the position of the interpolant of \mathbf{x} can be found by locating the point on the i th interpolative biplot axis that is calibrated with the value x_i for all $i \in [1:p]$, finding the centroid of these p points and then extending the vector emanating from the origin to this centroid p times - the endpoint of this extended vector gives the position of the point representing \mathbf{x} in the PCA biplot. It is evident that the calibrations on the k th interpolative biplot axis of the r -dimensional PCA biplot increases in the direction of the k th row vector of \mathbf{V}_r and decreases in the opposite direction. If on the other hand the interpolative biplot axes are calibrated in the same scales as the elements of \mathbf{X}^* , the point $\mu\mathbf{e}'_i\mathbf{V}_r$ will be calibrated with the value μ^* , where $\mu^*_{*i} = \mu + \bar{x}_i^*$ if the PCA biplot was constructed from the unstandardised measurements and $\mu^*_i = \mu\sqrt{\hat{\sigma}_{ii}} + \bar{x}_i^*$ if the PCA biplot was constructed from the standardised measurements. It follows that if the interpolative biplot axes are calibrated in terms of the scales of the elements of \mathbf{X}^* , then the position of \mathbf{x} can be found by locating the point on the i th biplot axis that is calibrated with the value x_i^* for all $i \in [1:p]$ and then extending the vector emanating from the origin to the centroid of these p points p times - the endpoint of this extended vector gives the position of the interpolant of \mathbf{x} .

The two-dimensional interpolative PCA biplot constructed from the standardised measurements of the *University* data set is provided in Figure 2.6. In this biplot it is illustrated how the position of the point representing Purdue University (*Purdue*) in the biplot is found using the vector sum approach. The points on the six biplot axes that are calibrated with the appropriate x_i^* values are indicated with solid triangles while the centroid of these six points is indicated with a solid square.

Samples other than those upon which the construction of the PCA biplot was based can be interpolated onto the existing PCA biplot in order to visualise their positions relative to the other samples. In order to interpolate the new sample \mathbf{x}^* onto the PCA biplot constructed from the matrix \mathbf{X} , each element of \mathbf{x}^* first needs to be transformed such that it is measured in the same scale as the elements of the corresponding column of the matrix \mathbf{X} . If the existing PCA biplot was constructed from the unstandardised measurements i.e. $\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}^*$, the new sample \mathbf{x}^* must be transformed to $\mathbf{x} = \mathbf{x}^* - \bar{\mathbf{x}}^*$, where $\bar{\mathbf{x}}^* = \frac{1}{n}\mathbf{1}'\mathbf{X}^*$. If the existing PCA biplot

was constructed from the standardised measurements i.e. $\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}\mathbf{A}^{-1}$, \mathbf{x}^* must be transformed to $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{x}^* - \bar{\mathbf{x}}^*)$. The point representing the new sample in the r -dimensional PCA biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$, is the orthogonal projection of \mathbf{x} onto \mathcal{L} . The position of this point can be obtained using the interpolative biplot axes in exactly the same way as explained above.

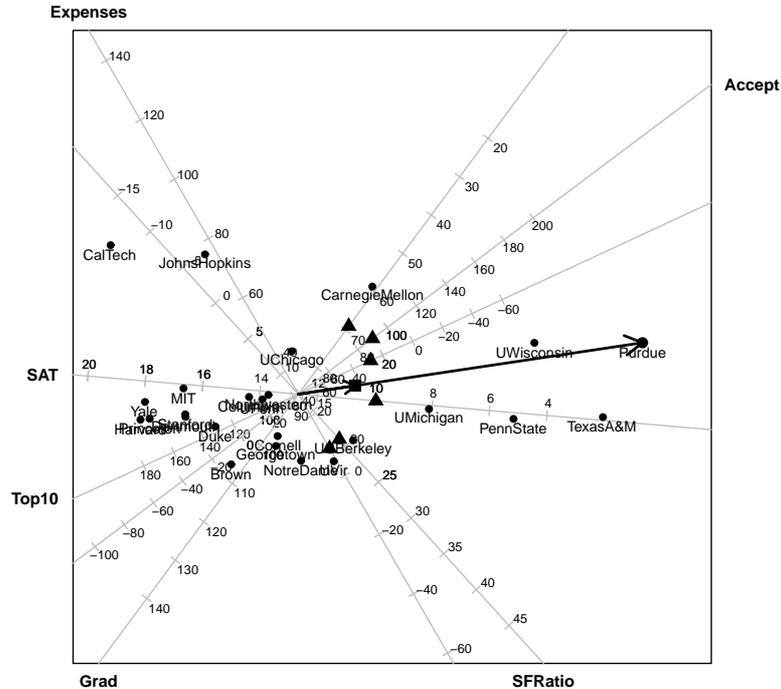


Figure 2.6: *The two-dimensional interpolative biplot constructed from the standardised measurements of the University data set, illustrating the vector-sum approach for Purdue University.*

2.7.1.2 Prediction and the predictive PCA biplot

Prediction is the process of inferring a sample's measurements on the measured variables, given the sample's position in the biplot space, \mathcal{L} . Hence, prediction is a process inversely related to interpolation. Prediction, like interpolation, is performed by relating the set of given values to a set of biplot axes. The biplot axes used for prediction purposes are called predictive biplot axes. For convenience the predictive biplot axis representing the k th measured variable will be referred to as the k th predictive biplot axis. The k th predictive biplot axis is calibrated such that the approximated measurement of a given sample on the k th measured variable can be read off from the k th biplot axis at the orthogonal projection of the point representing that sample in \mathcal{L} onto the k th predictive biplot axis, $k \in [1 : p]$. It follows that prediction is performed on a given sample by orthogonally projecting the point representing that sample in \mathcal{L} onto each of the p predictive biplot axes. The PCA

biplot in which the measured variables are represented by predictive biplot axes is called a predictive PCA biplot.

Firstly it is important to note that since the biplot space, \mathcal{L} , is contained in the measurement space, \mathbb{R}^p , every point in the biplot space is also contained in \mathbb{R}^p and hence can be expressed both in terms of the basis of \mathbb{R}^p and in terms of the basis of \mathcal{L} . Given a point \mathbf{z} in the r -dimensional PCA biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$, where the coordinates of \mathbf{z} are with respect to the column vectors of \mathbf{V}_r , the process of prediction consists of finding the coordinates of \mathbf{z} with respect to the basis vectors of the p -dimensional measurement space, \mathbb{R}^p . Let the coordinate vector of \mathbf{z} with respect to the basis of \mathbb{R}^p be denoted by \mathbf{x} . Since \mathbf{x} lies in the biplot space, \mathbf{x} is projected onto itself when it is orthogonally projected onto the biplot space, that is

$$\mathbf{x}'\mathbf{V}_r\mathbf{V}_r' = \mathbf{x}'.$$

Since the coordinate vector of \mathbf{x}' with respect to the basis of \mathcal{L} given by the column vectors of \mathbf{V}_r are given by $\mathbf{z}' = \mathbf{x}'\mathbf{V}_r$, it follows that

$$\mathbf{x}' = \mathbf{z}'\mathbf{V}_r'.$$

Note that the elements of the point $\mathbf{z}'\mathbf{V}_r'$ are in terms of the same scales as the elements of \mathbf{X} . If the prediction is to be given in terms of the scales of the elements of the original observed data matrix, \mathbf{X}^* , then $\mathbf{z}'\mathbf{V}_r'$ needs to be transformed. If the PCA biplot was constructed from the unstandardised measurements i.e. $\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}^*$, then the prediction of \mathbf{z} in terms of the scales corresponding to the elements of \mathbf{X}^* , is given by $\mathbf{z}'\mathbf{V}_r' + \bar{\mathbf{x}}^{*'}.$ If however the PCA biplot was constructed from the standardised measurements i.e. $\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}^*\mathbf{A}^{-1}$, then the prediction of \mathbf{z} in terms of the scales corresponding to the elements of \mathbf{X}^* is given by $\mathbf{z}'\mathbf{V}_r'\mathbf{A} + \bar{\mathbf{x}}^{*}.$

Consider again the coordinate vector of the point $\mathbf{z} \in \mathcal{L}$ with respect to the basis of \mathbb{R}^p , that is

$$\mathbf{x}' = \mathbf{z}'\mathbf{V}_r'.$$

A point which predicts the value μ for the k th variable, x_k , therefore satisfies

$$\mu = \mathbf{z}'\mathbf{V}_r'\mathbf{e}_k. \quad (2.7.1)$$

Hence, all the points in \mathcal{L} that predict the value μ for x_k lie on the $(r - 1)$ -dimensional hyperplane in \mathcal{L} that lies orthogonal to the vector $\mathbf{V}_r'\mathbf{e}_k$. The hyperplanes corre-

sponding to different values of μ lie parallel to each other and are all orthogonal to the vector $\mathbf{V}'_r \mathbf{e}_k$. The value of x_k predicted by a point \mathbf{z} in \mathcal{L} is therefore the value of μ which is such that \mathbf{z} lies on the hyperplane i.e.

$$\mu = \mathbf{z}' \mathbf{V}'_r \mathbf{e}_k .$$

If a line passing through the origin and orthogonal to these hyperplanes is constructed and calibrated with the value μ at the point where it intersects with the hyperplane $\mu = \mathbf{z}' \mathbf{V}'_r \mathbf{e}_k$, then the predicted value for x_k corresponding to a point \mathbf{z}^* in \mathcal{L} , can be obtained by projecting the point \mathbf{z}^* orthogonally onto the line and reading off the calibration at the point onto which \mathbf{z}^* projected. Every point \mathbf{z} on the line passing through the origin and orthogonal to the hyperplanes of the form, $\mu = \mathbf{z}' \mathbf{V}'_r \mathbf{e}_k$, can be expressed in the following way:

$$\mathbf{z}' = \sigma \mathbf{e}'_k \mathbf{V}_r$$

where σ is a constant. If $\sigma \mathbf{e}'_k \mathbf{V}_r$ is substituted for \mathbf{z}' in equation (2.7.1), then the following expression for the predicted value, μ , for x_k is obtained:

$$\mu = \sigma \mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r \mathbf{e}_k .$$

The value of σ corresponding to the value of μ is therefore given by

$$\sigma = \frac{\mu}{\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r \mathbf{e}_k}$$

and hence

$$\mathbf{z}' = \frac{\mu}{\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r \mathbf{e}_k} \mathbf{e}'_k \mathbf{V}_r . \quad (2.7.2)$$

The line on which every point is of the form in (2.7.2) and which is calibrated as explained above is called the k th predictive biplot axis of the PCA biplot. If the biplot axes are to be calibrated in terms of the scales corresponding to the elements of \mathbf{X}^* , then the point in (2.7.2) should be calibrated with the value $\mu + \bar{x}_k^*$ if $\mathbf{X} = (\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}') \mathbf{X}^*$ and the value $\mu \sqrt{\hat{\sigma}_{kk}^*} + \bar{x}_k^*$ if $\mathbf{X} = (\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}') \mathbf{X}^* \mathbf{A}^{-1}$. It is evident that like the k th interpolative biplot axis, the k th predictive biplot axis is

linear and lies collinear to the k th row vector of the matrix \mathbf{V}_r emanating from the origin, $k \in [1:p]$. It follows that if $\mathbf{e}_k \in \mathcal{V}(\mathbf{V}_r)$, the k th predictive biplot axis and the Cartesian axis that represents the variable x_k in the measurement space, are collinear. It is evident that the k th interpolative and k th predictive biplot axes of the r -dimensional PCA biplot are collinear for all $r \in [1:p]$. Note however that these two axes are only differently calibrated for $r < p$ - the k th interpolative and k th predictive biplot axes of the p -dimensional PCA biplot are identically calibrated:

$$\frac{\mu}{\mathbf{e}'_k \mathbf{V} \mathbf{V}' \mathbf{e}_k} \mathbf{e}'_k \mathbf{V} = \mu \mathbf{e}'_k \mathbf{V}.$$

As mentioned before, prediction is inversely related to interpolation. Prediction is actually equivalent to a process called back-projection (Gower and Hand, 1996). The point on the k th interpolative biplot axis that is calibrated with the value μ is given by the orthogonal projection of $\mu \mathbf{e}_k$ onto \mathcal{L} while the point on the k th predictive biplot axis that is calibrated with the value μ is given by the back-projection of $\mu \mathbf{e}_k$ onto \mathcal{L} (Gower and Hand, 1996).

It is evident that if $\mathbf{x}_i \in \mathcal{L} = \mathcal{V}(\mathbf{V}_r)$, then the prediction of \mathbf{x}_i read off from the p predictive biplot axes is

$$\hat{\mathbf{x}}_i = \mathbf{z}'_i \mathbf{V}'_r = \mathbf{x}'_i \mathbf{V}_r \mathbf{V}'_r = \mathbf{x}'_i$$

where $\mathbf{z}'_i = \mathbf{x}'_i \mathbf{V}_r$. That is, if \mathbf{x}_i lies in the biplot space, then it will be perfectly predicted by the predictive biplot axes. Hence, if the dimension of the biplot space, r , is equal to the rank of \mathbf{X} , q , then each of the n samples, $\{\mathbf{x}_i\}$, will be perfectly predicted by the predictive biplot axes. It is however not necessarily the case that a new sample that has been interpolated onto the biplot, will be perfectly predicted when $r = q$. Note however that when $r = p$, every sample, old and new, will be perfectly predicted by the predictive biplot axes since \mathbf{V} is an orthogonal matrix:

$$\hat{\mathbf{x}} = \mathbf{z}'_i \mathbf{V}' = \mathbf{x}'_i \mathbf{V} \mathbf{V}' = \mathbf{x}$$

As an example of a predictive PCA biplot, the two-dimensional predictive PCA biplot constructed from the standardised measurements of the *University* data set is provided in Figure 2.7. The fact that the predictive and interpolative biplot axes corresponding to a particular variable only differ with respect to their calibration, is evident upon comparison of Figure 2.7 and the two-dimensional interpolative biplot in Figure 2.6. In Figure 2.7 the prediction process is illustrated for the University of California, Berkeley (*UCBerkeley*), as well as Purdue University (*Purdue*). The predicted measurements produced by the two-dimensional predictive PCA biplot for the University of California, Berkeley, and Purdue University are compared to the

true measurements of the two universities in Table 2.3.

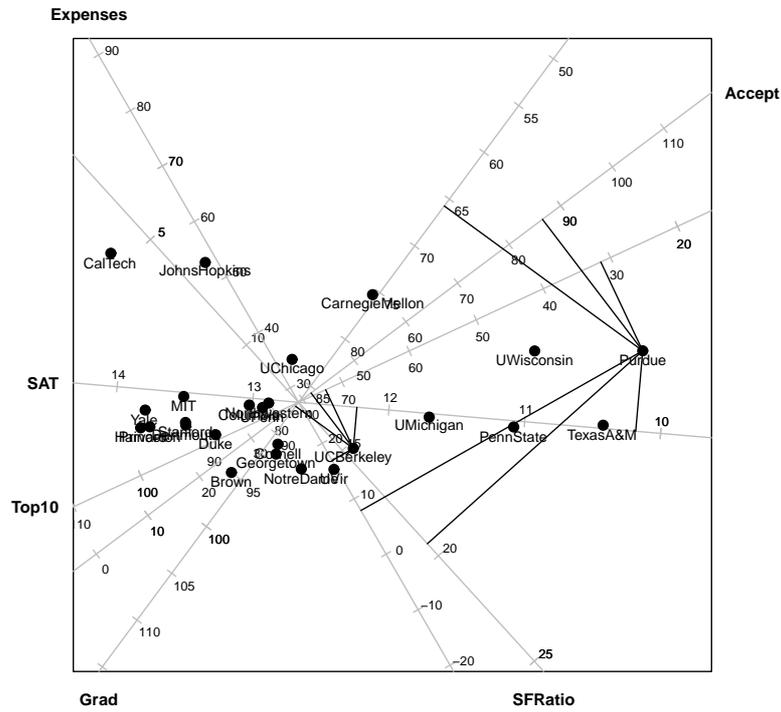


Figure 2.7: The two-dimensional predictive PCA biplot constructed from the standardised measurements of the University data set.

Table 2.3: The predictions of the measurements of the University of California, Berkeley (UCBerkeley), and Purdue University (Purdue) produced by the two-dimensional predictive PCA biplot constructed from the standardised measurements of the University data set.

		SAT	Top10	Accept	SFRatio	Expenses	Grad
UCBerkeley	predictions	12.235	72.450	41.611	15.205	16.906	87.148
UCBerkeley	true	12.400	95	40	17	15.140	78
Purdue	predictions	10.183	31.437	86.463	19.429	7.877	65.733
Purdue	true	10.050	28	90	19	9.066	69

Keeping the standard deviations of the measured variables (provided in Table 2.1) in mind, it is evident that all the measurements of Purdue University are accurately predicted by the two-dimensional PCA biplot - Purdue University is therefore accurately represented in the two-dimensional PCA biplot. The measurements of the University of California, Berkeley, with respect to the variables *SAT*, *Accept*, *SFRatio* and *Expenses*, are accurately predicted by the biplot while its measurements with respect to the variables *Top10* and *Grad*, are very poorly predicted

by the biplot - University of California, Berkeley, is therefore poorly represented in the two-dimensional PCA biplot. This illustrates that the same predictive PCA biplot can represent some samples accurately while representing others poorly. It is important to note that when a sample is poorly represented in a biplot, conclusions about the sample and its relationships to other samples should not be drawn based on its position in the biplot. The quality of the representation of individual samples in the predictive PCA biplot will be discussed in Section 3.4.2.

2.7.1.3 The relationship between prediction and multivariate regression analysis

The process of prediction can also be viewed from a multivariate regression analysis point of view. The definition of prediction can be formulated as predicting \mathbf{X} from $\mathbf{Z} = \mathbf{XV}_r$. Predicting one vector variable from another is the exact task performed by multivariate regression analysis. The only difference between multivariate regression and PCA is that in multivariate regression analysis some variables are defined as predictor variables and some are defined as response variables whereas in PCA the variables do not have different roles. When the prediction process is performed via multivariate regression, the \underline{x} -variables take on the role of the response (dependent) variables while the principal components take on the role of the predictor (independent) variables. The multivariate regression model in matrix notation is given by

$$\mathbf{X} = \mathbf{ZB} + \mathbf{E} \quad (2.7.3)$$

where \mathbf{B} is the matrix of parameters and \mathbf{E} is the matrix of random errors. This linear model approximates the matrix \mathbf{X} by

$$\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\mathbf{B}}$$

where $\widehat{\mathbf{B}}$ is the least squares estimator of \mathbf{B} , namely

$$\widehat{\mathbf{B}} = (\mathbf{Z}'\mathbf{Z})^C \mathbf{Z}'\mathbf{X}.$$

In the immediately preceding expression, $(\mathbf{Z}'\mathbf{Z})^C$ is an arbitrary conditional inverse of $\mathbf{Z}'\mathbf{Z}$. Since $\mathbf{XV}_r = \mathbf{U}_r\mathbf{D}_r$, it follows that

$$\text{rank}(\mathbf{XV}_r) = \text{rank}(\mathbf{U}_r\mathbf{D}_r) = r$$

and hence the matrix $\mathbf{Z} = \mathbf{X}\mathbf{V}_r$ is of full column rank, thereby satisfying an important assumption of regression analysis. This implies that the matrix $\mathbf{Z}'\mathbf{Z}$ is a non-singular matrix. Hence, the estimate of the parameter matrix, \mathbf{B} , is given by

$$\begin{aligned}\widehat{\mathbf{B}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \\ &= (\mathbf{V}_r'\mathbf{X}'\mathbf{X}\mathbf{V}_r)^{-1}\mathbf{V}_r'\mathbf{X}'\mathbf{X} \\ &= (\mathbf{V}_r'\mathbf{V}_p^2\mathbf{D}^2\mathbf{V}_p'\mathbf{V}_r)^{-1}\mathbf{V}_r'\mathbf{V}_p^2\mathbf{D}^2\mathbf{V}_p' \\ &= (\mathbf{D}_r^2)^{-1}\mathbf{D}_r^2\mathbf{V}_r' \\ \longrightarrow \widehat{\mathbf{B}} &= \mathbf{V}_r'\end{aligned}$$

so that

$$\widehat{\mathbf{X}} = \mathbf{X}\mathbf{V}_r\mathbf{V}_r'.$$

Notice that the approximation to \mathbf{X} that is produced by the multivariate regression of \mathbf{X} on \mathbf{Z} is identical to the approximation of \mathbf{X} that is produced by the predictive PCA biplot. The predictive PCA biplot thus represents an approximation to the systematic part of the multivariate regression model in (2.7.3). The columns of the estimated parameter matrix, $\widehat{\mathbf{B}}$, yield the coordinates of the unit points on the corresponding interpolative PCA biplot axes. Since the point on the k th predictive biplot axis that is calibrated with a value, say μ , can be obtained from the point on the k th interpolative biplot axis that is calibrated with the value μ (via the process of back-projection), the multivariate regression analysis method can be used to construct the interpolative and predictive biplot axes also yields the same approximation to \mathbf{X} as does the predictive PCA biplot.

This regression approach can also be used to construct a biplot axis for a new variable (Gower *et al.*, 2011). This does however not form part of the scope of this thesis.

2.8 Data structured into groups

The information that a PCA biplot of a data set consisting of samples structured into a number of predefined groups can provide regarding the group structure underlying that data set is unfortunately quite limited. The reason for this being that (1) the group membership of the samples plays no role in the construction of the PCA biplot and (2) Pythagorean distances does not take the correlations between the measured variables into account. However, different plotting characters and/or different colours can be used to represent samples belonging to different groups to highlight potential differences between the groups. For visualisation of the central locality of the different groups, the group means can be interpolated onto the PCA

biplot (see Section 2.7.1.1). Imposing an α -bag for each group may also add some information about the group structure underlying the data set as it allows for visual appraisal of the amount of overlap and/or separation amongst the groups. When the number of samples in a particular group is too small for the construction of an α -bag, a convex hull can be constructed for that group.

Table 2.4: *The standard deviations of the measured variables of the Ocotea data set.*

VesD	VesL	FibL	RayH	RayW	NumVes
24.53	89.48	214.05	68.39	7.13	5.23

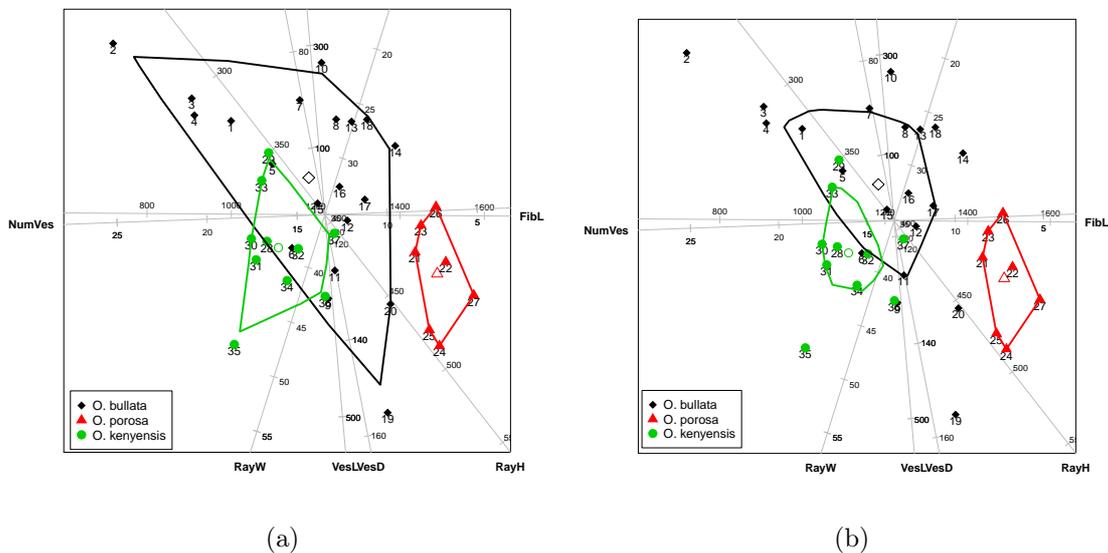


Figure 2.8: (a) *The two-dimensional predictive PCA biplot of the Ocotea data set with 95% bags constructed for $O. bullata$ and $O. porosa$ and a convex hull constructed for $O. kenyensis$; (b) *The two-dimensional predictive PCA biplot of the Ocotea data set with 50% bags constructed for $O. bullata$ and $O. porosa$ and a convex hull constructed for $O. kenyensis$.**

The foregoing concepts will now be illustrated at the hand of the *Ocotea* data set which is available in the R-package ‘*UBbipl*’. The *Ocotea* data set contains information on samples from three species of the Lauraceae family, namely *Ocotea bullata* (*O. bullata*), *Ocotea kenyensis* (*O. kenyensis*) and *Ocotea porosa* (*O. porosa*). The species *O. bullata* and *O. kenyensis* are indigenous to South Africa. The specie *O. porosa* on the other hand is an imported wood used as a substitute for *O. bullata* in the manufacturing of high quality furniture. The *Ocotea* data set contains information on 20 samples belonging to *O. bullata*, seven samples belonging to *O. kenyensis* and ten samples belonging to *O. porosa*. The data set provides measurements of the samples on six variables, namely tangential vessel diameter in μm (*VesD*), vessel element length in μm (*VesL*), fibre length in μm (*FibL*), ray height in μm (*RayH*), Ray width in μm (*RayW*) and number of vessels per square mm (*NumVes*). The stan-

standard deviations of the six measured variables are provided in Table 2.4. Since the measured variables have greatly differing standard deviations, the two-dimensional predictive PCA biplots of the *Ocotea* data set provided in Figure 2.8(a) and Figure 2.8(b), were constructed from the standardised measurements of the *Ocotea* data set.

To allow for visual appraisal of the extent to which the three groups overlap, 95% bags were superimposed onto the biplot in Figure 2.8(a) for the *O. bullata* and *O. porosa* species while a convex hull was superimposed for the *O. kenyensis* species (due to the small number of samples belonging to this species). Figure 2.8(b) contains the same PCA biplot as Figure 2.8(a) with the exception that 50% bags instead of 95% bags are superimposed for the *O. bullata* and *O. porosa* species. Different plotting characters as well as different colours were used to represent the samples (and centroids) of the three different species to ease visualisation of the differences between these species - solid black squares were used for the samples belonging to *O. bullata*, solid red triangles for those belonging to *O. kenyensis* and solid green circles for those belonging to *O. porosa*. For visualisation of the central locations of the three species, the centroid of the samples corresponding to each of the three species was interpolated onto the PCA biplot. The centroid of the samples belonging to *O. bullata* is indicated with a black square (not solid), that of *O. kenyensis* is indicated with a red triangle (not solid) and that of *O. porosa* is indicated with a green circle (not solid).

The 95% bags of *O. bullata* and *O. porosa* overlap substantially, indicating that these two species are likely to be very similar with respect to at least some of the measured variables. The fact that even the 50% bags of these two species overlap provides further evidence of their similarity. In order to know with respect to which variables these two species are similar and with respect to which they differ, their overlap with respect to the individual biplot axes needs to be investigated. It is very important to note that the mere overlap of two groups with respect to a biplot axis alone is not sufficient evidence to conclude that the two groups are similar with respect to the corresponding variable. The overlap of two groups only suggests a possible similarity between the two groups. In addition to the extent of overlap between the groups with respect to the biplot axis, the biplot axis' ability to reproduce the true measurements of the samples on that variable needs to be considered. A measure of the predictive ability of PCA biplot axes will be studied in Section 3.4.1.

Both the 95% bags and the 50% bags of *O. bullata* and *O. porosa* overlap on each of the individual biplot axis. This indicates that the *O. bullata* and *O. porosa* species are probably quite similar with respect to each of the six measured variables. In addition to the overlap between groups with respect to the the individual biplot axes, the overlap between groups with respect to pairs (or sets) of biplot axes should be considered. One possible difference between the *O. bullata* and *O. porosa* species that is suggested by the 50% bags in Figure 2.8(b) is that samples belonging to *O. porosa* that have measurements on *NumVes* and *Fibl* similar to those of samples belonging to *O. bullata*, tend to have greater measurements on *RayW*, *VesL*, *VesD* and *RayH* than the samples belonging to *O. bullata*.

There is no overlap between either of the *O. bullata* and *O. porosa* species'

95% bags and the convex hull of the *O. kenyensis* specie. This indicates that the *O. kenyensis* specie is probably very different from the *O. bullata* and *O. porosa* species with respect to at least some of the measured variables. This does however not imply that the *O. kenyensis* specie differs from the *O. bullata* and *O. porosa* species with respect to all six the measured variables. Projection of the two 95% bags and the convex hull onto the biplot axis representing the variable *RayW* shows a great extent of overlap of the three species with respect to this variable. This is also true for the 50% bags of the *O. bullata* and *O. porosa* species and the convex hull of the *O. kenyensis* specie in Figure 2.8(b). This indicates that the three species are likely to be very similar with respect to the variable *RayW*. On the other hand, not even the 95% bags of *O. bullata* and *O. porosa* overlap with the convex hull of *O. kenyensis* with respect to the biplot axes representing the variables *FibL* and *NumVes*, suggesting that the *O. bullata* and *O. porosa* species probably differ from the *O. kenyensis* specie to a great extent with respect to these two variables. The biplot also seems to suggest that samples belonging to *O. kenyensis* that have measurements on *RayW*, *VesL*, *VesD* and *RayH* similar to those of samples belonging to *O. bullata* and *O. porosa*, tend to have greater measurements on *FibL* and smaller measurements on *NumVes* than the samples belonging to *O. bullata* and *O. porosa*. However, as explained before, no conclusions can be made with certainty prior to considering the predictive abilities of the two biplot axes.

It should once again be emphasised that the PCA biplot is not designed to represent the underlying group structure of a data set. If visualisation of the group structure of a data set is the main interest of the investigator, then the data set should rather be represented by means of a CVA biplot which is designed specifically for this purpose.

2.9 Summary

PCA is a metric MDS technique that uses the Pythagorean distance metric to quantify distances. Two of the most well known routes along which PCA can be derived are those followed by Pearson (1901) and Hotelling (1933). While Pearson searched to find the straight line or hyperplane that fits the observed configuration of points as well as possible in terms of least squares, Hotelling searched for uncorrelated linear combinations of the measured variables that account for as much of the total variability associated with the measured vector variable as possible. A major weakness of PCA is the fact that it is scale dependent. When the standard deviations of the measured variables differ substantially, PCA should be performed on the standardised measurements. If on the other hand the standard deviations of the measured variables are very similar, PCA can be performed on either the standardised measurements or the unstandardised measurements.

The weakness of the ordinary lower-dimensional MDS configuration associated with a PCA performed on a data matrix is the lack of information about the measured variables. In the traditional biplot proposed by Gabriel (1971) each row and column of the data matrix at hand is represented by a vector emanating from the origin. These vectors are such that the inner product between a vector representing a row and a vector representing a column equals an approximation to the correspond-

ing element of the data matrix. Gabriel suggested that the samples be representing by the endpoints of the corresponding vectors alone so that samples and variables are easily differentiated in the biplot. The traditional PCA biplot is a special case of the traditional biplot proposed by Gabriel (1971). In the r -dimensional traditional PCA biplot a sample is represented by a point with coordinate vector given by the first r principal component scores associated with that sample while a variable is represented by a vector stretching from the origin up to the point with coordinate vector given by the coefficients of this variable in the first r principal components.

Given that inner products are difficult to visualise, Gower and Hand (1996) proposed that the measured variables be represented by calibrated axes that are calibrated such that the approximation to a sample's measurement on a variable can be read off from the axis representing that variable in the same way as in the case of an ordinary scatterplot. The axes that are calibrated in this way are called predictive biplot axes. Another set of axes which are used to position the samples of the data set in the biplot space is called the interpolative biplot axes. The predictive and interpolative biplot axes of the PCA biplot are linear and only differ with respect to their calibration.

The PCA biplot is not designed to represent the group structure underlying a data set consisting of samples that are structured into a number of predefined groups. By using different plotting characters and/or colours to represent samples belonging to different groups as well as imposing an α -bag (or convex hull) for each of the groups, certain differences between the groups may be suggested by the PCA biplot. If it is the accurate representation of the group structure underlying such a data set that is the main interest of the investigator, then the data set should rather be represented via a canonical variate analysis (CVA) biplot (on the condition that the data set satisfies the assumptions associated with CVA). This type of biplot is studied in detail in Chapter 4.

Chapter 3 - PCA biplot quality measures

In the previous chapter the construction and interpretation of the PCA biplot was discussed in detail. The conclusions drawn from a PCA biplot are however meaningless if the biplot poorly represents the observed data set. Measures of the quality of the various individual aspects of the PCA biplot are required in order to evaluate to what extent the relationships and predictions suggested by a PCA biplot are representative of reality.

The biplot quality measures discussed in this chapter are all defined as ratios of sums of squared values. In order for these quality measures to be meaningful, certain orthogonality properties must be satisfied. This chapter commences with a discussion of two types of orthogonality properties that underlie the validity of all the quality measures that will be studied in this chapter as well as in Chapter 5.

3.1 Orthogonality properties underlying a PCA biplot

Let \mathbf{X}^* denote the original observed data matrix and \mathbf{X} the transformed data matrix upon which the construction of the PCA biplot is based. Since the construction of the PCA biplot is based on the matrix \mathbf{X} , all the quality measures of the PCA biplot are defined in terms of \mathbf{X} , irrespective of the scales in which the biplot axes are calibrated.

When investigating the quality of the PCA biplot with respect to its ability to reproduce the matrix \mathbf{X} , an appropriate starting point is the identity

$$\mathbf{X} = \widehat{\mathbf{X}} + (\mathbf{X} - \widehat{\mathbf{X}}) \quad (3.1.1)$$

which shows the decomposition of the matrix \mathbf{X} into the approximation to \mathbf{X} produced by the (predictive) PCA biplot, $\widehat{\mathbf{X}}$, and the corresponding residual matrix, $\mathbf{X} - \widehat{\mathbf{X}}$. Note that the vectors stretching from the origin to the points with coordinates \mathbf{x}_i , $\widehat{\mathbf{x}}_i$ and $\mathbf{x}_i - \widehat{\mathbf{x}}_i$, lie in the p -dimensional measurement space, the biplot space, \mathcal{L} , and the orthogonal complement of the biplot space respectively, $i \in [1 : n]$.

Before investigating the matrix identity in (3.1.1), consider the corresponding

identity for a single vector \mathbf{x} :

$$\mathbf{x} = \hat{\mathbf{x}} + (\mathbf{x} - \hat{\mathbf{x}}) . \quad (3.1.2)$$

The identity in (3.1.2) shows how the vector \mathbf{x} is decomposed into a fitted part, $\hat{\mathbf{x}}$, and a residual part, $\mathbf{x} - \hat{\mathbf{x}}$. The quality of the approximation $\hat{\mathbf{x}}$ to \mathbf{x} may be measured by the ratio of the fitted to the total sum of squares,

$$\frac{\hat{\mathbf{x}}' \hat{\mathbf{x}}}{\mathbf{x}' \mathbf{x}} , \quad (3.1.3)$$

given that a certain orthogonality property is satisfied. In order for the ratio in (3.1.3) to be meaningful as a measure of the quality of the approximation, $\hat{\mathbf{x}}$, to \mathbf{x} , the fitted sum of squares, $\hat{\mathbf{x}}' \hat{\mathbf{x}}$, and the residual sum of squares, $(\mathbf{x} - \hat{\mathbf{x}})' (\mathbf{x} - \hat{\mathbf{x}})$, must add up to the total sum of squares, $\mathbf{x}' \mathbf{x}$, that is

$$\mathbf{x}' \mathbf{x} = \hat{\mathbf{x}}' \hat{\mathbf{x}} + (\mathbf{x} - \hat{\mathbf{x}})' (\mathbf{x} - \hat{\mathbf{x}}) . \quad (3.1.4)$$

When $\hat{\mathbf{x}}$ satisfies equation (3.1.4), the decomposition of \mathbf{x} in (3.1.2) is said to be orthogonal.

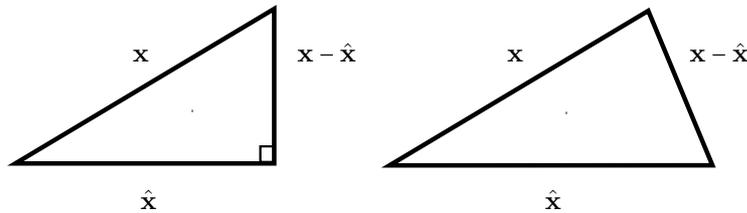


Figure 3.1: *Left: An orthogonal decomposition of a vector; Right: A non-orthogonal decomposition of a vector.*

If the decomposition in (3.1.2) is not orthogonal, the ratio of the fitted to the total sums of squares is not meaningful as a quality measure. This is evident when the elements of \mathbf{x} are measurements on the same variable. In that case, the ratio in (3.1.3) can be interpreted as the proportion of the sample variance of the variable that is accounted for by the approximation since the sample variance of a variable is proportional to its total sum of squares. When the decomposition in (3.1.2) is not orthogonal, it is possible for the fitted sum of squares to be greater than the total sum of squares and hence for the ratio in (3.1.3) to be greater than one. This is illustrated in Figure 3.1 which shows the decomposition of a vector \mathbf{x} into a fitted

3.1. ORTHOGONALITY PROPERTIES UNDERLYING A PCA BIPLLOT 97

part, $\hat{\mathbf{x}}$, and a residual part, $\mathbf{x} - \hat{\mathbf{x}}$. It does however not make sense for the proportion of variance accounted for to be greater than one. An orthogonal decomposition also ensures that the residual sum of squares is minimised and consequently that the ratio of the fitted to total sum of squares is maximised.

Now, consider the decomposition of the matrix \mathbf{X} in (3.1.1). This decomposition can exhibit two types of orthogonality properties - these will be referred to as Type A and Type B orthogonality in correspondence with Gardner-Lubbe *et al.* (2008). Type A orthogonality relates to the rows of \mathbf{X} (i.e. the samples) and is said to be exhibited by the decomposition of \mathbf{X} in (3.1.1) when

$$\mathbf{X}\mathbf{X}' = \widehat{\mathbf{X}}\widehat{\mathbf{X}}' + (\mathbf{X} - \widehat{\mathbf{X}})(\mathbf{X} - \widehat{\mathbf{X}})' . \quad (3.1.5)$$

Type B orthogonality on the other hand relates to the columns of \mathbf{X} (i.e. the variables) and is said to be exhibited by the decomposition of \mathbf{X} in 3.1.1 when

$$\mathbf{X}'\mathbf{X} = \widehat{\mathbf{X}}'\widehat{\mathbf{X}} + (\mathbf{X} - \widehat{\mathbf{X}})'(\mathbf{X} - \widehat{\mathbf{X}}) . \quad (3.1.6)$$

Consider the following expressions of $\mathbf{X}\mathbf{X}'$ and $\mathbf{X}'\mathbf{X}$:

$$\begin{aligned} \mathbf{X}\mathbf{X}' &= (\widehat{\mathbf{X}} + \mathbf{X} - \widehat{\mathbf{X}})(\widehat{\mathbf{X}} + \mathbf{X} - \widehat{\mathbf{X}})' \\ &= \widehat{\mathbf{X}}\widehat{\mathbf{X}}' + (\mathbf{X} - \widehat{\mathbf{X}})(\mathbf{X} - \widehat{\mathbf{X}})' + \widehat{\mathbf{X}}(\mathbf{X} - \widehat{\mathbf{X}})' + (\mathbf{X} - \widehat{\mathbf{X}})\widehat{\mathbf{X}}' \\ \mathbf{X}'\mathbf{X} &= (\widehat{\mathbf{X}} + \mathbf{X} - \widehat{\mathbf{X}})'(\widehat{\mathbf{X}} + \mathbf{X} - \widehat{\mathbf{X}}) \\ &= \widehat{\mathbf{X}}'\widehat{\mathbf{X}} + (\mathbf{X} - \widehat{\mathbf{X}})'(\mathbf{X} - \widehat{\mathbf{X}}) + \widehat{\mathbf{X}}'(\mathbf{X} - \widehat{\mathbf{X}}) + (\mathbf{X} - \widehat{\mathbf{X}})'\widehat{\mathbf{X}} . \end{aligned}$$

It is evident that if $\widehat{\mathbf{X}}$ is such that

$$\widehat{\mathbf{X}}(\mathbf{X} - \widehat{\mathbf{X}})' = \mathbf{0}$$

then the decomposition of \mathbf{X} into $\widehat{\mathbf{X}}$ and $\mathbf{X} - \widehat{\mathbf{X}}$ exhibits Type A orthogonality while if $\widehat{\mathbf{X}}$ satisfies

$$\widehat{\mathbf{X}}'(\mathbf{X} - \widehat{\mathbf{X}}) = \mathbf{0}$$

then the decomposition exhibits Type B orthogonality. It can be shown that if

$$\widehat{\mathbf{X}} = \mathbf{X}\mathbf{Q}_r\mathbf{Q}'_r$$

where \mathbf{Q} is an $p \times p$ orthogonal matrix and $1 \leq r \leq p$, then the decomposition of \mathbf{X} into $\widehat{\mathbf{X}}$ and $\mathbf{X} - \widehat{\mathbf{X}}$ exhibits type A orthogonality (Gardner-Lubbe *et al.*, 2008). A derivation of this result similar to that provided in Gardner-Lubbe *et al.* (2008), is provided below:

$$\begin{aligned} \widehat{\mathbf{X}} &= \mathbf{X}\mathbf{Q}_r\mathbf{Q}'_r \\ \rightarrow \widehat{\mathbf{X}}(\mathbf{X} - \widehat{\mathbf{X}})' &= \widehat{\mathbf{X}}\mathbf{X}' - \widehat{\mathbf{X}}\widehat{\mathbf{X}}' \\ &= \mathbf{X}\mathbf{Q}_r\mathbf{Q}'_r\mathbf{X}' - \mathbf{X}\mathbf{Q}_r\mathbf{Q}'_r\mathbf{Q}_r\mathbf{Q}'_r\mathbf{X} \\ &= \mathbf{X}\mathbf{Q}_r\mathbf{Q}'_r\mathbf{X}' - \mathbf{X}\mathbf{Q}_r\mathbf{Q}'_r\mathbf{X} \\ \rightarrow \widehat{\mathbf{X}}(\mathbf{X} - \widehat{\mathbf{X}})' &= \mathbf{0} \\ \rightarrow \mathbf{X}\mathbf{X}' &= \widehat{\mathbf{X}}\widehat{\mathbf{X}}' + (\mathbf{X} - \widehat{\mathbf{X}})(\mathbf{X} - \widehat{\mathbf{X}})' . \end{aligned}$$

When $\widehat{\mathbf{X}} = \mathbf{X}\mathbf{Q}_r\mathbf{Q}'_r$ and \mathbf{Q} is the orthogonal matrix with column vectors equal to the right singular vectors of \mathbf{X} , the decomposition of \mathbf{X} into $\widehat{\mathbf{X}}$ and $(\mathbf{X} - \widehat{\mathbf{X}})$ exhibits both Type A and Type B orthogonality (Gardner-Lubbe *et al.*, 2008). In order to show this, let the svd of \mathbf{X} be given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V} .$$

Since the matrix of right singular vectors, \mathbf{V} , is an orthogonal matrix, the decomposition of \mathbf{X} into

$$\widehat{\mathbf{X}} = \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r = \mathbf{X}\mathbf{V}_r\mathbf{V}'_r$$

and $\mathbf{X} - \widehat{\mathbf{X}}$ exhibits Type A orthogonality. This decomposition also exhibits Type B orthogonality since the matrix $\widehat{\mathbf{X}}'(\mathbf{X} - \widehat{\mathbf{X}})$ is equal to the null matrix:

$$\begin{aligned} \widehat{\mathbf{X}}'(\mathbf{X} - \widehat{\mathbf{X}}) &= \widehat{\mathbf{X}}'\mathbf{X} - \widehat{\mathbf{X}}'\widehat{\mathbf{X}} \\ &= \mathbf{V}_r\mathbf{D}_r\mathbf{U}'_r\mathbf{U}\mathbf{D}\mathbf{V}' - \mathbf{V}_r\mathbf{D}_r\mathbf{U}'_r\mathbf{U}_r\mathbf{D}_r\mathbf{V}'_r \\ &= \mathbf{V}_r\mathbf{D}_r[\mathbf{I} \quad \mathbf{0}]\mathbf{D}\mathbf{V}' - \mathbf{V}_r\mathbf{D}_r\mathbf{U}'_r\mathbf{U}_r\mathbf{D}_r\mathbf{V}'_r \end{aligned}$$

$$\begin{aligned}
&= \mathbf{V}_r \mathbf{D}_r^2 \mathbf{V}_r' - \widehat{\mathbf{V}}_r \widehat{\mathbf{D}}_r^2 \widehat{\mathbf{V}}_r' \\
\longrightarrow \widehat{\mathbf{X}}' (\mathbf{X} - \widehat{\mathbf{X}}) &= \mathbf{0}.
\end{aligned}$$

Recall that the approximation to \mathbf{X} produced by the r -dimensional (predictive) PCA biplot is given by

$$\widehat{\mathbf{X}} = \mathbf{X} \mathbf{V}_r \mathbf{V}_r'$$

where \mathbf{V} is the matrix with i th column vector equal to the i th right singular vector of \mathbf{X} , $i \in [1 : p]$. Hence, the decomposition of the matrix \mathbf{X} into the approximation produced by the (predictive) PCA biplot, $\widehat{\mathbf{X}}$, and the matrix of residuals, $\mathbf{X} - \widehat{\mathbf{X}}$, exhibits both Type A and Type B orthogonality.

In this chapter, as in Chapter 2, the general case where the $n \times p$ matrix \mathbf{X} is of rank q , where $1 \leq q \leq p$, will be studied while it will be assumed that all q the non-zero singular values of \mathbf{X} are distinct.

3.2 The overall quality of the PCA biplot

It has become standard practice to use the ratio of the fitted sum of squares to the total sum of squares,

$$\frac{\text{tr} \{ \widehat{\mathbf{X}} \widehat{\mathbf{X}}' \}}{\text{tr} \{ \mathbf{X} \mathbf{X}' \}} = \frac{\text{tr} \{ \widehat{\mathbf{X}}' \widehat{\mathbf{X}} \}}{\text{tr} \{ \mathbf{X}' \mathbf{X} \}} \quad (3.2.1)$$

to measure the overall quality of the PCA biplot (see for example Gabriel (1971), Cox and Cox (2001), Gower and Hand (1996), Gardner-Lubbe *et al.* (2008) and Gower *et al.* (2011)). Both Type A and Type B orthogonality underlie the validity of the ratio in equation (3.2.1) as a quality measure:

$$\begin{aligned}
\text{Type A : } \mathbf{X} \mathbf{X}' &= \widehat{\mathbf{X}} \widehat{\mathbf{X}}' + (\mathbf{X} - \widehat{\mathbf{X}}) (\mathbf{X} - \widehat{\mathbf{X}})' \\
\longrightarrow \text{tr} \{ \mathbf{X} \mathbf{X}' \} &= \text{tr} \{ \widehat{\mathbf{X}} \widehat{\mathbf{X}}' \} + \text{tr} \{ (\mathbf{X} - \widehat{\mathbf{X}}) (\mathbf{X} - \widehat{\mathbf{X}})' \} \\
\text{and Type B : } \mathbf{X}' \mathbf{X} &= \widehat{\mathbf{X}}' \widehat{\mathbf{X}} + (\mathbf{X} - \widehat{\mathbf{X}})' (\mathbf{X} - \widehat{\mathbf{X}}) \\
\longrightarrow \text{tr} \{ \mathbf{X}' \mathbf{X} \} &= \text{tr} \{ \widehat{\mathbf{X}}' \widehat{\mathbf{X}} \} + \text{tr} \{ (\mathbf{X} - \widehat{\mathbf{X}})' (\mathbf{X} - \widehat{\mathbf{X}}) \}.
\end{aligned}$$

Note that since the decomposition of \mathbf{X} into $\widehat{\mathbf{X}}$ and $\mathbf{X} - \widehat{\mathbf{X}}$ exhibits Type A orthogonality,

$$\frac{\text{tr}\{\widehat{\mathbf{X}}\widehat{\mathbf{X}}'\}}{\text{tr}\{\mathbf{X}\mathbf{X}'\}} = 1 - \frac{\text{tr}\{(\mathbf{X} - \widehat{\mathbf{X}})(\mathbf{X} - \widehat{\mathbf{X}})'\}}{\text{tr}\{\mathbf{X}\mathbf{X}'\}}. \quad (3.2.2)$$

Hence, using the ratio in (3.2.1) as a measure of the overall quality of the PCA biplot is equivalent to using the ratio

$$\frac{\text{tr}\{(\mathbf{X} - \widehat{\mathbf{X}})(\mathbf{X} - \widehat{\mathbf{X}})'\}}{\text{tr}\{\mathbf{X}\mathbf{X}'\}} = 1 - \frac{\text{tr}\{\widehat{\mathbf{X}}\widehat{\mathbf{X}}'\}}{\text{tr}\{\mathbf{X}\mathbf{X}'\}}$$

as a measure of the overall loss of information resulting from the dimension reduction.

Let the overall quality of the PCA biplot be denoted by Ω . Letting the svd of \mathbf{X} be given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{U}_q\mathbf{D}_q\mathbf{V}_q'$$

the overall quality of the r -dimensional PCA biplot can be expressed as:

$$\Omega = \frac{\text{tr}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})}{\text{tr}(\mathbf{X}'\mathbf{X})} \quad (3.2.3)$$

$$= \frac{\text{tr}\{\mathbf{V}_r\mathbf{D}_r^2\mathbf{V}_r'\}}{\text{tr}\{\mathbf{V}_q\mathbf{D}_q^2\mathbf{V}_q'\}} \quad (3.2.4)$$

$$= \frac{\text{tr}\{\mathbf{D}_r^2\mathbf{V}_r'\mathbf{V}_r\}}{\text{tr}\{\mathbf{D}_q^2\mathbf{V}_q'\mathbf{V}_q\}}$$

$$= \frac{\text{tr}\{\mathbf{D}_r^2\}}{\text{tr}\{\mathbf{D}_q^2\}}$$

$$\longrightarrow \Omega = \frac{\sum_{k=1}^r d_k^2}{\sum_{k=1}^q d_k^2} \quad (3.2.5)$$

where $d_k = [\mathbf{D}]_{kk}$ is the k th largest non-zero singular value of \mathbf{X} or equivalently, the square root of the k th largest non-zero eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$. Since $d_k^2 > 0$ for $k \in [1 : q]$, the overall quality of the PCA biplot can only take on positive values. Since the numerator of Ω is a non-decreasing function of the dimension of the PCA

biplot, r , while the denominator is fixed, Ω is a non-decreasing function of r . The overall quality of the PCA biplot will therefore necessarily equal its maximum value when $r = p$. It is evident from equation (3.2.5) that the maximum value that Ω can attain is one. The condition $r = p$ is sufficient for Ω to attain its maximum value irrespective of the rank of \mathbf{X} but only necessary when \mathbf{X} is of full column rank. It is evident from equation (3.2.5) that the condition which is necessary and sufficient for Ω to equal one is $r = q$. This is also evident from the expression of the overall quality in equation (3.2.2). Since Ω is a decreasing function of the sum of squared residuals,

$$\begin{aligned} \text{tr} \left\{ (\mathbf{X} - \widehat{\mathbf{X}}) (\mathbf{X} - \widehat{\mathbf{X}})' \right\} &= \sum_{i=1}^n (\mathbf{x}_i - \widehat{\mathbf{x}}_i)' (\mathbf{x}_i - \widehat{\mathbf{x}}_i) \\ &= \sum_{i=1}^n \sum_{j=1}^p \left([\mathbf{X}]_{ij}^2 - [\widehat{\mathbf{X}}]_{ij}^2 \right)^2 \end{aligned}$$

it will attain its maximum value of one if and only if the sum of squared residuals attains its minimum value. It is evident that the sum of squared residuals has a minimum value of zero which it will attain if and only if

$$\begin{aligned} [\mathbf{X}]_{ij} &= [\widehat{\mathbf{X}}]_{ij} \quad \forall i \in [1:n], j \in [1:p] \\ \iff \widehat{\mathbf{X}} &= \mathbf{X} \\ \iff \mathbf{x}_i &\in \mathcal{L} \quad \forall i \in [1:n] \\ \iff r &= q. \end{aligned}$$

It follows that Ω will attain its maximum value of one if and only if the dimension of the PCA biplot, r , is equal to the rank of \mathbf{X} , q .

Being a function of the squared singular values of \mathbf{X} , which are scale dependent quantities, the overall quality of the PCA biplot is itself a scale dependent quantity. When the measured variables have widely differing standard deviations, the overall quality of the PCA biplot constructed from the unstandardised measurements will usually be overly optimistic - this will be explained in Section 3.4.1.3. An example illustrating the scale dependence of the overall quality measure of the PCA biplot will be provided in Section 3.4.1.6.

As a result of the fact that the sample variance associated with x_j is proportional to $\mathbf{x}'_{(j)} \mathbf{x}_{(j)}$, the total sample variance associated with $\underline{\mathbf{x}}$ as measured by the one-dimensional measure $\sum_{i=1}^p \text{vâr}(x_i)$, is proportional to $\text{tr}(\mathbf{X}'\mathbf{X})$. Hence, the overall quality of the PCA biplot,

$$\Omega = \frac{\text{tr} \{ \widehat{\mathbf{X}}' \widehat{\mathbf{X}} \}}{\text{tr} \{ \mathbf{X}' \mathbf{X} \}} = \frac{\sum_{j=1}^p \widehat{\mathbf{x}}'_{(j)} \widehat{\mathbf{x}}_{(j)}}{\sum_{j=1}^p \mathbf{x}'_{(j)} \mathbf{x}_{(j)}}$$

can be interpreted as the proportion of the total sample variance associated with $\underline{\mathbf{x}}$ which is accounted for in the PCA biplot.

If a desired proportion of the total sample variance associated with $\underline{\mathbf{x}}$ to be accounted for in the PCA biplot has been specified prior to the investigation of the data, any PCA biplot corresponding to an overall quality that is equal to or greater than the desired proportion can be used to investigate the data. Since only one, two and three-dimensional PCA biplots can be visualised, a three-dimensional PCA biplot should be used to investigate the data in the event that the desired proportion of variance to be accounted for in the biplot is greater than the overall quality associated with the three-dimensional PCA biplot. If however the three-dimensional PCA biplot has a very poor overall quality, the investigator should be very careful about drawing conclusions based on the visual inspection of the biplot alone. Later on in this chapter, it will be explained that it is possible for certain individual samples or variables to be accurately represented in a PCA biplot with poor overall quality. In the event that the PCA biplot upon which the investigation of the data set is to be based has poor overall quality, the investigator should only draw conclusions based on those samples or variables that are accurately represented.

Table 3.1: *The overall quality of the PCA biplot constructed from the standardised measurements of the University data set corresponding to each possible dimensionality of the PCA biplot.*

Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
0.769	0.900	0.948	0.975	0.996	1.000

As an example, consider the overall quality of the r -dimensional PCA biplot constructed from the standardised measurements of the *University* data set for every $r \in [1 : 6]$ as provided in Table 3.1. It is evident from Table 3.1 that if the desired proportion of the total sample variance associated with $\underline{\mathbf{x}}$ (where $\underline{\mathbf{x}}$ is the vector of standardised measured variables) to be accounted for in the PCA biplot is 0.9, the data should be represented in a two or three-dimensional PCA biplot. If on the other hand the desired proportion of total sample variance to be accounted for in the PCA biplot is only 0.75, a one-dimensional PCA biplot will suffice. If the desired proportion of variance to be accounted for in the PCA biplot is greater than 0.948, the data should be represented in a three dimensional PCA biplot, though the desired proportion will not be obtained.

The scree plot associated with \mathbf{X} (Cattell, 1966) can also be used to determine an appropriate dimension for the PCA biplot. The scree plot corresponding to the *University* data set is given in Figure 3.2. It seems as if the tip of the ‘elbow’ in the scree plot is at the point corresponding to the third eigenvalue which means that the first two principal components accounts for a sufficiently large proportion of the sample variance associated with $\underline{\mathbf{x}}$ and hence that the two-dimensional PCA biplot will represent the *University* data set sufficiently accurate.

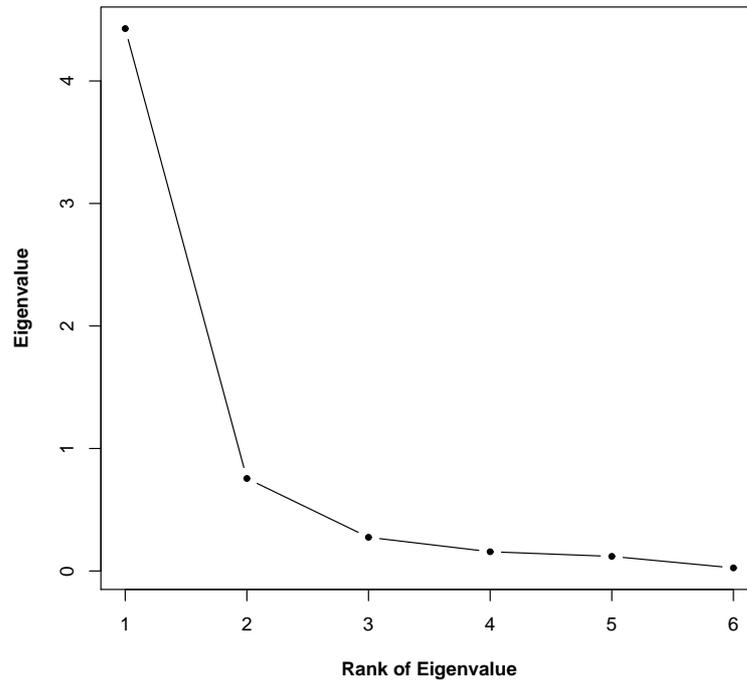


Figure 3.2: *The scree plot corresponding to the (standardised) University data set.*

Note that since the eigenvalues of $\mathbf{X}'\mathbf{X}$ are proportional to the eigenvalues of $\widehat{\Sigma}$ the shape of the scatter plot of

$$\frac{d_k^2}{\sum_{i=1}^q d_i^2} \quad (3.2.6)$$

against $\{k\}_{k=1}^p$ in which the consecutive points are connected by straight lines will look exactly like the shape of the scree plot associated with \mathbf{X} . Note that the ratio in (3.2.6) is equal to the relative contribution of the k th dimension of the PCA biplot to the overall quality of the r -dimensional PCA biplot of \mathbf{X} , or equivalently the relative contribution of the k th principal component to the overall quality, for $k \leq r$ and zero for $k > r$. The scatter plot of $\left\{ \frac{d_k^2}{\sum_{i=1}^q d_i^2} \right\}_{k=1}^p$ against $\{k\}_{k=1}^p$ can therefore be used to determine what the dimension of the PCA biplot should be such that the biplot represents the observed data set sufficiently accurate in the same way that the scree plot is traditionally used to determine the number of principal components to use in an approximation of \mathbf{X} . Remember however that the PCA biplot can be at most three-dimensional. Hence, if the ‘tip of the elbow’ in the plot of $\left\{ \frac{d_k^2}{\sum_{i=1}^q d_i^2} \right\}_{k=1}^p$ against $\{k\}_{k=1}^p$ indicates that more than three principal components should be used to produce a sufficiently accurate approximation of \mathbf{X} , then a three-dimensional

PCA biplot should be used to graphically represent the data set. As mentioned before - if the three-dimensional PCA biplot has a very poor overall quality, the investigator should be very careful when drawing conclusions based on the visual inspection of the biplot alone.

The plot of the overall quality of the PCA biplot against the dimensionality of the biplot space can also be used to determine an appropriate dimension for the PCA biplot to be constructed. Such a scatter plot can be viewed as a cumulative version of the scree plot. The dimension of the PCA biplot which would represent the data sufficiently accurate can therefore be found by identifying the ‘elbow’ that points upwards (instead of the ‘elbow’ that points downwards as in the case of the scree plot). The dimension of the PCA biplot which would represent the data set sufficiently accurate is the dimension corresponding to the point at which the slope of the lines connecting the points in the plot change from steep to not steep. Again, if this dimension is greater than three, a three-dimensional PCA biplot should be used. Since it is a plot of cumulative sums, the ‘elbow’ will usually not be as sharp as in the case of the scree plot and therefore may be better described simply as a decrease in the slope. It is therefore usually easier to determine the appropriate dimension of the PCA biplot from the scree plot or the plot of the relative contributions of the principal components to the overall quality as described earlier.

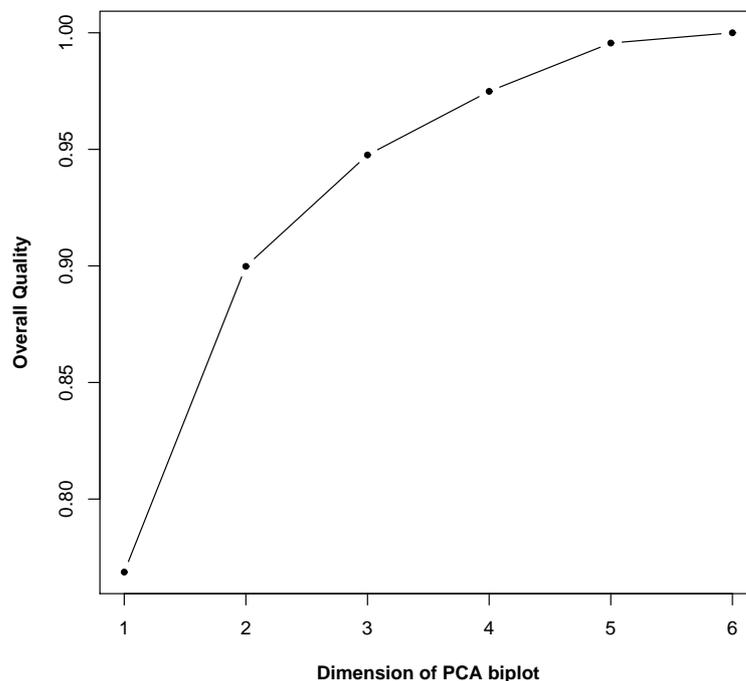


Figure 3.3: *The overall quality of the PCA biplot of the University data set, constructed from the standardised data, corresponding to each possible dimensionality of the PCA biplot.*

Figure 3.3 provides the plot of the overall qualities of the PCA biplots constructed from the standardised measurements of the *University* data set against the corresponding dimensionalities of the biplots. In Figure 3.3 the gradients of the lines connecting the dots only really flattens after the point corresponding to the three-dimensional PCA biplot - the slope of the line joining the points corresponding to the overall qualities of the two-and three-dimensional PCA biplots is still relatively steep. This plot therefore indicates that the three-dimensional PCA biplot should be used to represent the *University* data set.

The problem with the overall quality measure is that it only considers the biplot as a whole and therefore does not necessarily provide accurate information about the quality about the various individual aspects of the biplot. To see this, consider the expression of the overall loss in quality associated with the PCA biplot of a matrix \mathbf{X} :

$$\frac{\text{tr} \{ (\mathbf{X} - \widehat{\mathbf{X}})' (\mathbf{X} - \widehat{\mathbf{X}}) \}}{\text{tr} \{ \mathbf{X}' \mathbf{X} \}} = \frac{\sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \hat{x}_{ij})^2}{\sum_{i=1}^n \sum_{j=1}^p x_{ij}^2}. \quad (3.2.7)$$

Note that it is not necessary for each of the terms of the summation in the numerator of equation (3.2.7) to be very small in order for the overall loss in quality to be very small, or equivalently, in order for the overall quality to be very high. Therefore, a very high overall quality does not necessarily imply that every element of the matrix \mathbf{X} is accurately approximated in the biplot. Similarly, when the overall quality of the PCA biplot is very low, it does not imply that all elements of the matrix \mathbf{X} are poorly approximated in the biplot. It follows that the overall quality of a PCA biplot does not provide information on the quality of the representation of every individual sample or variable. When a particular sample or variable is poorly represented in the PCA biplot, conclusions drawn about that sample or variable based on the visual inspection of the PCA biplot alone, are likely to be erroneous. This emphasises the need for measures of the quality of the representation of the individual samples and variables in the PCA biplot.

Two types of quality measures that focus on individual aspects of the biplot, namely adequacies and predictivities, will be studied in the remainder of this chapter. The term ‘adequacy’ was coined by Gardner (2001) although the measure was suggested earlier by Gower and Hand (1996) as a measure of the quality of the representation of the individual measured variables. Gardner-Lubbe *et al.* (2008) proposed two new quality measures, namely axis predictivities and sample predictivities to measure the quality of the representation of the individual variables and samples respectively. It will be explained in Section 3.3 that the adequacy of the representation of a variable as defined by Gardner (2001) is not a trustworthy measure of the predictive ability of that biplot axis - the axis predictivity of the biplot axis on the other hand, is. It can however be shown that the adequacy of the representation of a variable is a lower bound for the axis predictivity of the corresponding biplot axis. Hence, the adequacy of the representation of a variable can in some circumstances provide useful information about the predictive ability of the corresponding

biplot axis. For this reason as well as to improve the reader's understanding of what exactly is measured by the adequacy measure, adequacies will be studied in section 3.3.

3.3 Adequacies

3.3.1 Definition and properties

The adequacy of the representation of a variable in a PCA biplot, as defined by Gardner (2001), measures how adequately the corresponding biplot axis represents the Cartesian axis representing that variable in the p -dimensional measurement space. The more a biplot axis departs from the corresponding Cartesian axis in the measurement space, the less adequately the Cartesian axis is represented in the PCA biplot space. For example, if the Cartesian axis representing a variable in the measurement space lies in the biplot space, it is represented as adequately as possible in the biplot space since it lies collinear with the corresponding biplot axis. On the other hand, if the Cartesian axis representing a variable in the measurement space lies orthogonal to the PCA biplot space, then that Cartesian axis will be represented by the null vector in the PCA biplot - this is the least adequate possible representation of a Cartesian axis in the biplot space.

Let γ_k denote the adequacy of the representation of the k th variable, x_k , in the biplot space and \mathbf{e}_k denote the p -dimensional unit vector with k th element equal to one and all other elements equal to zero, that is the unit vector along the k th Cartesian axis in the p -dimensional measurement space, $k \in [1:p]$. The adequacy of the representation of the k th measured variable in the r -dimensional PCA biplot is defined as the ratio of the square of the length representing one unit of the k th measured variable in the r -dimensional PCA biplot space to the square of the length representing one unit of the k th measured variable in the p -dimensional measurement space (Gardner (2001), Gardner-Lubbe *et al.* (2008)), that is:

$$\begin{aligned} \gamma_k &= \frac{\|\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r\|^2}{\|\mathbf{e}'_k\|^2} \\ &= \|\mathbf{e}'_k \mathbf{V}_r\|^2 \\ &= [\mathbf{V}_r \mathbf{V}'_r]_{kk} \\ \gamma_k &= \sum_{j=1}^r v_{kj}^2. \end{aligned} \tag{3.3.1}$$

Note that the decomposition of the identity matrix, \mathbf{I}_p , into $\mathbf{I}_p \mathbf{V}_r \mathbf{V}'_r$ and $\mathbf{I}_p - \mathbf{I}_p \mathbf{V}_r \mathbf{V}'_r$ exhibits Type A orthogonality since the \mathbf{V}_r is an orthonormal matrix. This orthogonality property underlies the validity of the adequacy of the representation of the k th variable as defined in equation (3.3.1) as a quality measure. In the remainder of this chapter, the adequacy of the representation of the k th measured variable will

also be referred to as the adequacy of the representation of the k th Cartesian axis, the adequacy of the k th biplot axis or simply the k th adequacy.

Being defined as the summation of r squared values, the adequacy of a biplot axis can only take on non-negative values and is a non-decreasing function of the dimension of the PCA biplot space, r . The adequacy of the k th biplot axis, γ_k , will therefore necessarily attain its maximum value when $r = p$. Since \mathbf{V} is an orthogonal matrix, this implies that the maximum value of γ_k is one. This is also evident from the Pythagoras' theorem, according to which the length of the vector $\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r$, that is the length of the orthogonal projection of the vector \mathbf{e}_k onto the r -dimensional PCA biplot space, will always be smaller than or equal to the length of the vector \mathbf{e}_k . That is, one unit of the k th variable, x_k , will always be represented by a length smaller than or equal to one in the r -dimensional PCA biplot space, $r \in [1 : p]$. The greater the size of the angle between the vectors \mathbf{e}'_k and $\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r$, the angle being measured in the two-dimensional subspace spanned by these two vectors, the smaller the length of the orthogonal projection of \mathbf{e}_k onto $\mathcal{V}(\mathbf{V}_r)$ will be. That is, the more the k th biplot axis of the r -dimensional PCA biplot departs from the Cartesian axis representing x_k in the p -dimensional measurement space, the smaller the length representing one unit of x_k in the r -dimensional PCA biplot will be. This is shown below:

$$\begin{aligned}
 \gamma_k &= \frac{\|\mathbf{e}'_k \mathbf{V}_r\|^2}{\|\mathbf{e}'_k\|^2} \\
 &= \frac{\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r \mathbf{e}_k}{\|\mathbf{e}'_k\|^2} \\
 &= \left\{ \frac{(\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r \mathbf{e}_k)^{1/2}}{\|\mathbf{e}'_k\|} \right\}^2 \\
 &= \left\{ \frac{\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r \mathbf{e}_k}{\|\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r\| \|\mathbf{e}'_k\|} \right\}^2 \\
 &\rightarrow \gamma_k = \cos^2(\theta_k)
 \end{aligned} \tag{3.3.2}$$

where θ_k denotes the angle between the vectors \mathbf{e}'_k and $\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r$, as measured in the two-dimensional subspace spanned by these two vectors. For convenience the angle θ_k will henceforth simply be referred to as the angle between the vectors \mathbf{e}'_k and $\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r$ or equivalently, the angle between the k th biplot axis and the k th Cartesian axis in the measurement space. It is evident that γ_k decreases monotonically from one to zero as θ_k increases from 0° to 90° . This implies that the length representing one unit of x_k in the r -dimensional PCA biplot decreases monotonically from one to zero as the angle between the k th biplot axis of the r -dimensional PCA biplot and the k th Cartesian axis in the measurement space increases from 0° to 90° .

From the expression of γ_k in (3.3.2) it is evident that γ_k has a maximum value

of one which it will attain if and only if $\theta_k = 0^\circ$ i.e.

$$\begin{aligned}\gamma_k = 1 &\longleftrightarrow \mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r = \mathbf{e}'_k \\ \therefore \gamma_k = 1 &\longleftrightarrow \mathbf{e}_k \in \mathcal{L}\end{aligned}$$

and a minimum value of zero which it will attain if and only if $\theta_k = 90^\circ$ i.e.

$$\begin{aligned}\gamma_k = 0 &\longleftrightarrow \mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r = \mathbf{0}' \\ \therefore \gamma_k = 0 &\longleftrightarrow \mathbf{e}_k \in \mathcal{L}^\perp.\end{aligned}$$

That is, the k th adequacy will attain its maximum value of one if and only if the k th Cartesian axis lies in the PCA biplot space while it will attain its minimum value of zero if and only if the k th Cartesian axis lies orthogonal to the PCA biplot space. The fact that $\gamma_k = 1$ if and only if the k th Cartesian axis lies in the PCA biplot space can also be shown as follows:

$$\begin{aligned}\gamma_k &= \frac{\sum_{j=1}^r v_{kj}^2}{\sum_{j=1}^p v_{kj}^2} \\ \therefore \gamma_k = 1 &\longleftrightarrow \sum_{j=r+1}^p v_{kj}^2 = 0 \\ \therefore \gamma_k = 1 &\longleftrightarrow v_{kj} = 0 \quad \forall j \in [r+1 : p] \\ \therefore \gamma_k = 1 &\longleftrightarrow \mathbf{e}'_k [\mathbf{v}_{(r+1)} \quad \dots \quad \mathbf{v}_{(p)}] = \mathbf{0}' \\ \therefore \gamma_k = 1 &\longleftrightarrow \mathbf{e}_k \perp \mathcal{V}^\perp(\mathbf{V}_r) \\ \therefore \gamma_k = 1 &\longleftrightarrow \mathbf{e}_k \in \mathcal{V}(\mathbf{V}_r).\end{aligned}$$

Similarly, the fact that $\gamma_k = 0$ if and only if the k th Cartesian axis lies orthogonal to the PCA biplot space, can be shown as follows:

$$\begin{aligned}\gamma_k &= \frac{\sum_{j=1}^r v_{kj}^2}{\sum_{j=1}^p v_{kj}^2} \\ \therefore \gamma_k = 0 &\longleftrightarrow \sum_{j=1}^r v_{kj}^2 = 0 \\ \therefore \gamma_k = 0 &\longleftrightarrow \mathbf{e}'_k \mathbf{V}_r = \mathbf{0} \\ \therefore \gamma_k = 0 &\longleftrightarrow \mathbf{e}_k \in \mathcal{L}^\perp.\end{aligned}$$

Given that γ_k will equal one if and only if the vector \mathbf{e}_k lies in \mathcal{L} and since all p the unit vectors, $\{\mathbf{e}_k\}$, can only lie in \mathcal{L} if $r = p$, it follows that $r = p$ is a necessary and sufficient condition for all p the biplot axes to have unit adequacies. This fact is also evident from the properties of the matrix, \mathbf{V} . Since the \mathbf{V} is an orthogonal matrix, all p its row vectors have unit lengths and hence if $r = p$, all p biplot axes will have unit adequacies. When $r < p$, \mathbf{V}_r is an orthonormal matrix which means that all p its row vectors cannot have unit lengths and hence all p the biplot axes cannot have unit adequacies. It is important to note that this holds irrespective of the rank of \mathbf{X} . When \mathbf{X} is of rank q with $q < p$, then the p biplot axes of the q -dimensional PCA biplot will not all have unit adequacies even though the matrix \mathbf{X} is perfectly represented in the q -dimensional PCA biplot. This implies that the adequacy of a biplot axis is not an appropriate measure to quantify the quality of the biplot axis with respect to its ability to reproduce the true measurements on the corresponding variable. A measure of the predictive ability of a PCA biplot axis called axis predictivity, which was proposed by Gardner-Lubbe *et al.* (2008), will be discussed in Section 3.4.1. It will be shown in Section 3.4.1 that the adequacy of a biplot axis is a lower bound for the axis predictivity of that biplot axis. Consequently, the adequacy of a biplot axis can provide useful information regarding the predictive ability of the biplot axis in some circumstances.

Note that since the adequacy of the PCA biplot axis is a function of the coefficients of the right singular vectors of the matrix \mathbf{X} , which are scale dependent quantities, the adequacy measure is itself scale dependent. If x_k has unit sample variance, the adequacy of the k th biplot axis of the r -dimensional PCA biplot is equal to the ratio of the square of the length representing one standard deviation of x_k in the r -dimensional PCA biplot space to the square of the length representing one standard deviation of x_k in the p -dimensional PCA biplot space. An example illustrating the scale dependence of the adequacy measure (together with the scale dependence of the overall quality measure) will be given in Section 3.4.1.6.

As an example, consider the adequacies of the biplot axes of the two-dimensional PCA biplot constructed from the standardised measurements of the *University* data set provided in Table 3.2. The fact that all six the biplot axes have fairly low to very low adequacies implies that each of the six biplot axes departs substantially from the corresponding Cartesian axis in the measurement space. It is evident that of the six biplot axes, the biplot axis representing the variable *SAT* departs most from the corresponding Cartesian axis in the measurement space while the biplot axis representing the variable *Expenses* departs the least from the corresponding Cartesian axis in the measurement space. It is also evident that the four biplot axes representing the variables *SAT*, *Top10*, *Accept* and *SFRatio* lie at very similar angles to the corresponding Cartesian axes in the measurement space.

Table 3.2: *The adequacies of the biplot axes of the two-dimensional PCA biplot constructed from the standardised measurements of the University data set.*

<i>SAT</i>	<i>Top10</i>	<i>Accept</i>	<i>SFRatio</i>	<i>Expenses</i>	<i>Grad</i>
0.21	0.22	0.28	0.34	0.53	0.41

3.3.2 Visual representation

Given that the k th adequacy, γ_k , is an increasing function of $\|\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r\|$ with a minimum value of zero and a maximum value of one, γ_k is a decreasing function of the Pythagorean distance between the endpoint of the vector $\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r$ emanating from the origin and the circumference of an r -dimensional unit sphere centred at the origin. It follows that the relative magnitudes of the distances between the endpoints of the vectors $\{\mathbf{e}'_k \mathbf{V}_r \mathbf{V}'_r\}$ and the circumference of the r -dimensional unit sphere centred at the origin, are representative of the relative magnitudes of the p adequacies of the r -dimensional PCA biplot. This is illustrated in Figure 3.4 at the hand of the *University* data set.

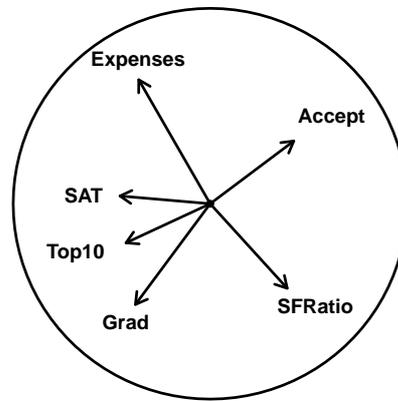


Figure 3.4: A unit circle in the two-dimensional PCA biplot space of the standardised *University* data set that is centred at the origin together with the projections of the six-dimensional unit vectors, $\{\mathbf{e}_k\}$, onto the biplot space.

Figure 3.4 shows a unit circle in the two-dimensional PCA biplot space of the standardised *University* data set that is centred at the origin with a vector emanating from the origin for each of the (standardised) measured variables. The vector representing the k th variable in Figure 3.4 is the projection of the six-dimensional unit vector \mathbf{e}_k that emanates from the origin onto the two-dimensional PCA biplot space, $k \in [1:6]$. Upon comparison of the lengths of the vectors representing the variables in Figure 3.4 to the adequacies in Table 3.2, it is evident that the relative magnitudes of the lengths of the vectors in Figure 3.4 represent the relative magnitudes of the corresponding adequacies. For example, the vector representing *Expenses*, which is represented by the biplot axis with the largest adequacy value, has the greatest length of the six vectors in Figure 3.4 while the vectors representing *SAT* and *Top10*, which are represented by the two biplot axes with the two smallest adequacy values, are the two vectors with the shortest lengths of the six vectors. The fact that the difference in the lengths of the vectors representing *SAT* and *Top10* is almost impossible to see, agrees with the fact that the adequacies of the two corresponding biplot axes are almost the same.

This graphical representation of the adequacies in a unit circle reminds of the correlation monoplot described in Gower *et al.* (2011). In the same way that the axes of the correlation monoplot are calibrated relative to a length of unity to read off the accuracy with which each of the variables approximates the unit correlation of an exact representation, the vectors emanating from the origin in Figure 3.4 can be extended and calibrated relative to a length of unity such that the adequacies can be read off directly.

Instead of representing the relative magnitudes of the adequacies of the p biplot axes of the r -dimensional PCA biplot using the unit circle (or sphere) approach as explained above, the relative magnitudes of the adequacies can be represented in the PCA biplot itself. This can be done, using either the interpolative or predictive r -dimensional PCA biplots, by imposing a thick straight line onto the k th biplot axis that stretches from the origin up to the point $\mathbf{e}'_k \mathbf{V}_r$ for $k \in [1 : p]$ - the relative magnitudes of the lengths of the thickened parts of the biplot axes represent the relative magnitudes of the adequacies. This yields exactly the same visual representation of the relative magnitudes of the adequacies as the unit circle (or sphere) approach. Given the way in which the interpolative biplot axes are calibrated, this is slightly easier to do using the interpolative PCA biplot. When the PCA biplot is constructed from the (centred and) standardised measurements, then the point $\mathbf{e}'_k \mathbf{V}_r$ in the r -dimensional interpolative PCA biplot will be calibrated with the value one if the biplot axes are calibrated in the same scales as elements of the matrix \mathbf{X} or with the value equal to the sum of the k th measured variable's observed mean and sample standard deviation if the biplot axes are calibrated in the scales of the original observed measurements.

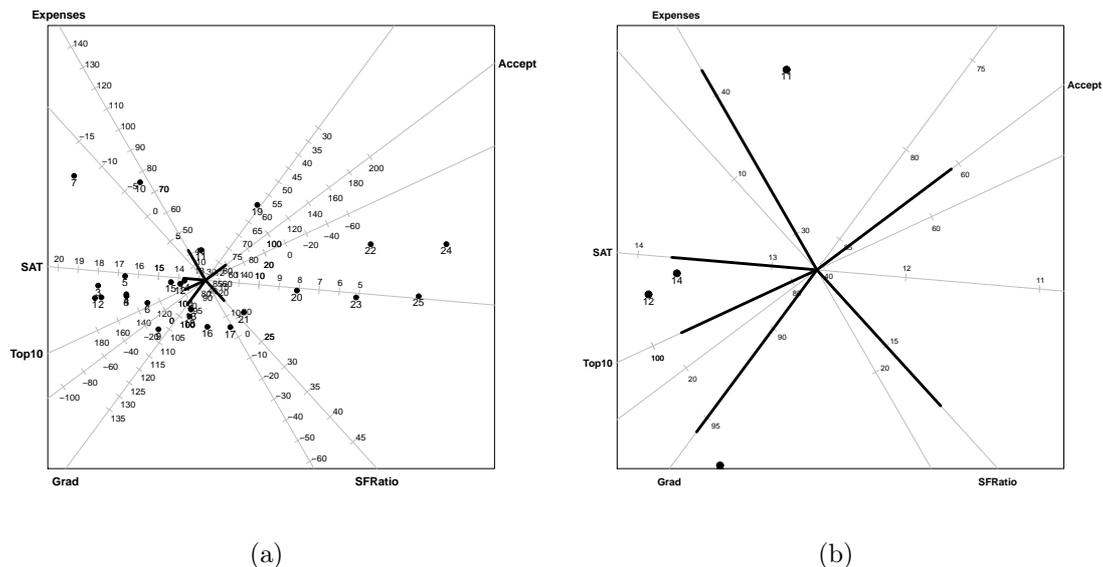


Figure 3.5: (a) The two-dimensional interpolative PCA biplot of the University data set with thick lines the relative lengths of which represents the relative magnitudes of the adequacies of the measured variables; (b) A small section of the interpolative PCA biplot in (a).

Figure 3.5(a) contains the two-dimensional interpolative PCA biplot constructed from the standardised measurements of the *University* data set with the biplot axes thickened in the way just explained. The biplot axes in Figure 3.5(a) are calibrated in the scales of the original observed measurements and hence each of the biplot axes were thickened from the origin up to the point calibrated with the value equal to the sum of the corresponding variable's sample mean and standard deviation. A small section of the PCA biplot has been enlarged and reproduced in Figure 3.5(b) to ease the comparison of the lengths of the thickened parts of the biplot axes. It is evident that the relative magnitudes of the lengths of the thickened parts of the biplot axes in Figures 3.5(a) and 3.5(b) exhibit the same pattern as the relative magnitudes of the adequacies in Table 3.2.

In conclusion, the adequacy of a biplot axis is not an appropriate measure to use to assess the predictive ability of the biplot axis. It will be explained in Section 3.4.1 that the adequacy of a biplot axis can however in some circumstances provide useful information about the biplot axis' ability to reproduce the true measurements of the corresponding variable.

3.4 Predictivities

3.4.1 Axis predictivities

3.4.1.1 Definition and properties

To assess the predictive ability of the individual PCA biplot axes, Gardner-Lubbe *et al.* (2008) proposed a quality measure which they called the axis predictivity of a biplot axis. The axis predictivity of a predictive biplot axis is a measure of the overall accuracy of the approximations which are read off from that particular predictive biplot axis.

Since the interpolative PCA biplot axes are only used to find the positions of the points representing the samples in the biplot space and not to read off approximations to the true values of the variables represented by the biplot axes, it is implied that it is the k th predictive biplot axis which is being referred to when reporting the axis predictivity of the k th biplot axis.

The axis predictivity of the k th biplot axis, or k th axis predictivity for short, is defined as the ratio of the sum of the squared predicted measurements on the k th variable to the sum of the squared true measurements (Gardner-Lubbe *et al.*, 2008). Let the k th axis predictivity be denoted by π_k , then

$$\begin{aligned} \pi_k &= \frac{\sum_{i=1}^n \hat{x}_{ik}^2}{\sum_{i=1}^n x_{ik}^2} \\ &= \frac{\hat{\mathbf{x}}'_{(k)} \hat{\mathbf{x}}_{(k)}}{\mathbf{x}'_{(k)} \mathbf{x}_{(k)}} \\ &= \frac{[\hat{\mathbf{X}}' \hat{\mathbf{X}}]_{kk}}{[\mathbf{X}' \mathbf{X}]_{kk}} \end{aligned} \tag{3.4.1}$$

$$= \frac{[\mathbf{V}_r \mathbf{D}_r^2 \mathbf{V}_r']_{kk}}{[\mathbf{V}_q \mathbf{D}_q^2 \mathbf{V}_q']_{kk}} \quad (3.4.2)$$

$$\longrightarrow \pi_k = \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^q d_j^2 v_{kj}^2} \quad (3.4.3)$$

where q denotes the rank of \mathbf{X} and r denotes the dimension of the PCA biplot as before. Since the decomposition of \mathbf{X} into $\widehat{\mathbf{X}} = \mathbf{X} \mathbf{V}_r \mathbf{V}_r'$ and $(\mathbf{X} - \widehat{\mathbf{X}})$ exhibits Type B orthogonality, that is

$$\mathbf{X}'\mathbf{X} = \widehat{\mathbf{X}}'\widehat{\mathbf{X}} + (\mathbf{X} - \widehat{\mathbf{X}})'(\mathbf{X} - \widehat{\mathbf{X}}) \quad (3.4.4)$$

π_k is meaningful as a quality measure, $r \in [1:p]$, $k \in [1:p]$. Note that unlike the adequacy of the k th biplot axis, the axis predictivity of the k th biplot axis depends on the true dimensionality of the configuration of points, $\{\mathbf{x}_j\}$, that is the rank of the matrix \mathbf{X} , q , and not simply on the number of measured variables, p . An axis predictivity can also be defined for a biplot axis representing a new variable that has been interpolated onto the PCA biplot (Gower *et al.*, 2011). This however does not lie within the scope of this thesis.

Let the p -component vector of axis predictivities with k th diagonal element equal to π_k for $k \in [1:p]$, be denoted by $\boldsymbol{\pi}$, then

$$\boldsymbol{\pi} = \frac{\text{diag}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})}{\text{diag}(\mathbf{X}'\mathbf{X})} \quad (3.4.5)$$

$$\longrightarrow \boldsymbol{\pi} = \frac{\text{diag}(\mathbf{V}_r \mathbf{D}_r^2 \mathbf{V}_r')}{\text{diag}(\mathbf{V}_q \mathbf{D}_q^2 \mathbf{V}_q')}.$$

It is evident from equation (3.4.4) that the k th axis predictivity can also be expressed as:

$$\pi_k = 1 - \frac{(\mathbf{x}_{(k)} - \widehat{\mathbf{x}}_{(k)})'(\mathbf{x}_{(k)} - \widehat{\mathbf{x}}_{(k)})}{\mathbf{x}_{(k)}'\mathbf{x}_{(k)}}.$$

Note that since $\frac{(\mathbf{x}_{(k)} - \widehat{\mathbf{x}}_{(k)})'(\mathbf{x}_{(k)} - \widehat{\mathbf{x}}_{(k)})}{\mathbf{x}_{(k)}'\mathbf{x}_{(k)}}$ is a ratio of sums of squared values, it can only take on non-negative values. It is evident that the smaller $(\mathbf{x}_{(k)} - \widehat{\mathbf{x}}_{(k)})'(\mathbf{x}_{(k)} - \widehat{\mathbf{x}}_{(k)})$ is, that is the more accurately $\widehat{\mathbf{x}}_{(k)}$ approximates $\mathbf{x}_{(k)}$ (in terms of least squares), the greater the k th axis predictivity will be. It is evident that the k th axis predictivity measures the overall accuracy of the approximations produced by the k th predictive

biplot axis.

Being defined as the ratio of sums of squared values, the axis predictivity measure can only take on non-negative values. It is evident from equation (3.4.3) that π_k is a non-decreasing function of the dimension of the PCA biplot space, r . The k th axis predictivity, π_k , will therefore necessarily equal its maximum value when $r = p$. It is evident from equation (3.4.3) that the maximum value that π_k can attain is one. Since π_k is a decreasing function of the sum of squared residuals, $(\mathbf{x}_{(k)} - \hat{\mathbf{x}}_{(k)})'(\mathbf{x}_{(k)} - \hat{\mathbf{x}}_{(k)})$, it will attain its maximum value if and only if the sum of squared residuals attains its minimum value. Since the sum of squared residuals, $(\mathbf{x}_{(k)} - \hat{\mathbf{x}}_{(k)})'(\mathbf{x}_{(k)} - \hat{\mathbf{x}}_{(k)})$, will attain its minimum value of zero if and only if

$$\hat{\mathbf{x}}_{(k)} = \mathbf{x}_{(k)}$$

π_k will attain its maximum value of one if and only if $\hat{\mathbf{x}}_{(k)} = \mathbf{x}_{(k)}$. It follows that all p biplot axes will have unit predictivity if and only if $\hat{\mathbf{X}} = \mathbf{X}$, or equivalently if and only if $r = q$. Note that $r = q$ is a sufficient but not a necessary condition for an individual biplot axis to attain unit axis predictivity. At this point it is important to note that the condition required for all p biplot axes to attain unit axis predictivities differs from that required for all p biplot axes to attain unit adequacies. Recall from Section 3.3 that the condition required for all p biplot axes to attain unit adequacies is $r = p$, irrespective of the true dimensionality of the data. This is the reason why the adequacy of a biplot axis is not a trustworthy measure of the biplot axis' quality with respect to its predictive ability. Note that $\pi_k = 1 \forall k \in [1 : p]$ when $r > q$, but since q is the true dimensionality of the data, it does not make sense to consider $r > q$. For this reason, only the two situations, namely $r < q$ and $r = q$ will be considered. It is evident from the expression of the k th axis predictivity given in (3.4.1) that π_k has a minimum value of zero which it will attain if and only if

$$\begin{aligned} \hat{\mathbf{x}}_{(k)}' \hat{\mathbf{x}}_{(k)} &= 0 \\ \therefore \pi_k = 0 &\longleftrightarrow \hat{\mathbf{x}}_{(k)} = \mathbf{0}. \end{aligned}$$

Due to the fact that \mathbf{X} is centred such that $\frac{1}{n}\mathbf{1}'\mathbf{X} = \mathbf{0}'$, the sum of squares $\mathbf{x}_{(k)}'\mathbf{x}_{(k)}$ is proportional to the sample variance of the variable x_k . Consequently the k th axis predictivity is equal to the proportion of the total sample variance of the k th variable, x_k , that is accounted for in the biplot.

Since a small axis predictivity means that overall, the measurements of the samples on that variable are poorly predicted by the biplot, no conclusions should be drawn about a variable that is represented by a biplot axis with low axis predictivity, based on the mere visual inspection of the biplot. In particular, no conclusions should be drawn regarding the relationships between samples with respect to that variable. Also, given two variables, when one or both of these variables have a low axis predictivity, then no conclusions should be drawn about the relationship be-

tween these two variables based on the visual inspection of the biplot. Only when the measurements of both variables are accurately predicted can it be trusted that the relationship between the variables which is suggested by the PCA biplot, is an accurate representation of reality.

3.4.1.2 The relationship between the axis predictivity and adequacy of a biplot axis

When \mathbf{X} is of full column rank and all p singular values of \mathbf{X} are identical i.e. when

$$\mathbf{D} = c\mathbf{I}$$

where c is any positive constant, the axis predictivity and adequacy of a biplot axis are identical:

$$\begin{aligned} \mathbf{D} = c\mathbf{I} &\longrightarrow \pi_k = \frac{[c\mathbf{V}_r\mathbf{V}'_r]_{kk}}{[c\mathbf{V}\mathbf{V}']_{kk}} \\ &= \frac{[\mathbf{V}_r\mathbf{V}'_r]_{kk}}{[\mathbf{V}\mathbf{V}']_{kk}} \\ \therefore \mathbf{D} = c\mathbf{I} &\longrightarrow \pi_k = \gamma_k . \end{aligned}$$

This means that when $\mathbf{D} = c\mathbf{I}$, the k th adequacy (like the k th axis predictivity) is equal to the proportion of the sample variance associated with the k th variable, x_k , that is accounted for in the PCA biplot. It is shown below that $\mathbf{D} = c\mathbf{I}$ implies that the p variables, $\{x_k\}$, are uncorrelated and have identical sample variances, that is, the configuration of points with coordinate vectors equal to the row vectors of \mathbf{X} is in the shape of a perfect spheroid:

$$\begin{aligned} \mathbf{D} &= \mathbf{I} \\ \longrightarrow \mathbf{X}'\mathbf{X} &= \mathbf{V}\mathbf{D}_p^2\mathbf{V}' \\ &= c^2\mathbf{V}\mathbf{V}' \\ &= c^2\mathbf{I} . \end{aligned}$$

The situation where some or all of the singular values of \mathbf{X} are identical is however not investigated any further in this thesis. More general relationships between the axis predictivity measure and adequacy measure will be investigated in what follows.

Consider the expression of the k th axis predictivity in (3.4.3). It is evident that

$$\pi_k = 0 \iff \sum_{j=1}^r d_j^2 v_{kj}^2 = 0 .$$

However, since

$$d_j^2 > 0 \quad \forall j \in [1 : q] ,$$

it follows that

$$\sum_{j=1}^r d_j^2 v_{kj}^2 = 0 \iff \mathbf{e}'_k \mathbf{V}_r = \mathbf{0}'$$

and hence that

$$\pi_k = 0 \iff \mathbf{e}'_k \mathbf{V}_r = \mathbf{0}' .$$

It follows that the predictivity of the k th biplot axis will equal zero if and only if the Cartesian axis representing the k th variable, x_k , in the measurement space, lies orthogonal to the biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$, that is

$$\pi_k = 0 \iff \mathbf{e}_k \perp \mathcal{L} .$$

Recall that

$$\gamma_k = 0 \iff \mathbf{e}_k \perp \mathcal{L} .$$

This means that the axis predictivity of the k th biplot axis will be equal to zero if and only if the adequacy of the k th biplot axis is equal to zero i.e.

$$\pi_k = 0 \iff \gamma_k = 0 .$$

The next derivation shows that the k th biplot axis will necessarily have unit axis predictivity if it has unit adequacy, irrespective of the rank of the matrix \mathbf{X} :

$$\begin{aligned} \gamma_k = 1 &\longrightarrow \sum_{j=r+1}^p v_{kj}^2 = 0 \\ &\longrightarrow \sum_{j=r+1}^q d_j^2 v_{kj}^2 = 0 \text{ for } q \leq p \\ &\longrightarrow \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^q d_j^2 v_{kj}^2} = 1 \\ \therefore \gamma_k = 1 &\longrightarrow \pi_k = 1 \\ \therefore \mathbf{e}_k \in \mathcal{L} &\longrightarrow \pi_k = 1. \end{aligned}$$

This is also evident upon consideration of the following expression of the prediction of a row vector \mathbf{x}' of \mathbf{X} , which is produced by the r -dimensional PCA biplot in the case where $\mathbf{e}_k \in \mathcal{V}(\mathbf{V}_r)$:

$$\begin{aligned} \hat{\mathbf{x}}' &= \mathbf{x}' \mathbf{V}_r \mathbf{V}_r' \\ &= \sum_{j=1}^p x_j \mathbf{e}_j' \mathbf{V}_r \mathbf{V}_r' \\ &= x_k \mathbf{e}_k' \mathbf{V}_r \mathbf{V}_r' + \sum_{j \neq k} x_j \mathbf{e}_j' \mathbf{V}_r \mathbf{V}_r' \\ \longrightarrow \hat{\mathbf{x}}' &= x_k \mathbf{e}_k' + \sum_{j \neq k} x_j \mathbf{e}_j' \mathbf{V}_r \mathbf{V}_r'. \end{aligned}$$

Consequently, the prediction of the k th element of \mathbf{x} , \hat{x}_k , can be expressed in the following way:

$$\begin{aligned} \hat{x}_k &= \hat{\mathbf{x}}' \mathbf{e}_k \\ &= x_k \mathbf{e}_k' \mathbf{e}_k + \sum_{j \neq k} x_j \mathbf{e}_j' \mathbf{V}_r \mathbf{V}_r' \mathbf{e}_k \\ &= x_k + \sum_{j \neq k} x_j \mathbf{e}_j' \mathbf{e}_k \\ \longrightarrow \hat{x}_k &= x_k. \end{aligned} \tag{3.4.6}$$

The second last step in the above derivation follows from the fact that $\mathbf{e}_k' \mathbf{V}_r \mathbf{V}_r' = \mathbf{e}_k'$ while the last step follows from the fact that $\mathbf{e}_j' \mathbf{e}_k = 0$ for all $j \in [1:p]$, $j \neq k$. Since the above derivation is true for all the row vectors of the matrix \mathbf{X} , it follows that when $\mathbf{e}_k \in \mathcal{L}$, all the elements of the k th column of \mathbf{X} are perfectly predicted by the r -dimensional PCA biplot and hence the axis predictivity of the k th biplot axis is

equal to unity:

$$\begin{aligned} \mathbf{e}_k \in \mathcal{L} &\longrightarrow \hat{x}_{ik} = x_{ik} \quad \forall i \in [1 : n] \\ \therefore \mathbf{e}_k \in \mathcal{L} &\longrightarrow \pi_k = \mathbf{X}'_{(k)} \mathbf{X}_{(k)} \mathbf{X}'_{(k)} \mathbf{X}_{(k)} \\ \therefore \mathbf{e}_k \in \mathcal{L} &\longrightarrow \pi_k = 1. \end{aligned}$$

Unit axis predictivity however only implies unit adequacy when \mathbf{X} is of full column rank. It is shown below that when the rank of \mathbf{X} , q , is less than p , unit axis predictivity does not necessarily imply unit adequacy:

$$\begin{aligned} \pi_k &= \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^q d_j^2 v_{kj}^2} \\ &= \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^r d_j^2 v_{kj}^2 + \sum_{j=r+1}^q d_j^2 v_{kj}^2} \\ \pi_k = 1 &\longrightarrow \sum_{j=r+1}^q d_j^2 v_{kj}^2 = 0 \\ \longrightarrow v_{kj}^2 &= 0 \quad \forall j \in [r+1 : q] \quad \text{since } d_j^2 > 0 \quad \forall j \in [1 : q] \\ \longrightarrow \gamma_k &= \frac{\sum_{j=1}^r v_{kj}^2}{\sum_{j=1}^p v_{kj}^2} \\ \therefore \gamma_k &= \frac{\sum_{j=1}^r v_{kj}^2}{\sum_{j=1}^r v_{kj}^2 + \sum_{j=r+1}^q v_{kj}^2 + \sum_{j=q+1}^p v_{kj}^2} \\ \therefore \gamma_k &= \frac{\sum_{j=1}^r v_{kj}^2}{\sum_{j=1}^r v_{kj}^2 + \sum_{j=q+1}^p v_{kj}^2}. \end{aligned}$$

The fact that the k th axis predictivity is equal to one does not imply anything about the values of the $(q+1)$ th to p th elements of the k th row vector of \mathbf{V} which means that $\sum_{j=q+1}^p v_{kj}^2 \geq 0$ and hence that $\gamma_k \leq 1$. This result is also evident from the fact that all p biplot axes of the q -dimensional PCA biplot, will have unit axis predictivities although not all p biplot axes will have unit adequacies when $q < p$. When however \mathbf{X} is of full column rank, unit axis predictivity does imply unit adequacy. It is shown below that when \mathbf{X} is of full column rank, the k th axis predictivity will be equal to one if and only if the k th adequacy is equal to one:

$$\begin{aligned} \pi_k &= \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^p d_j^2 v_{kj}^2} \\ &= \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^r d_j^2 v_{kj}^2 + \sum_{j=r+1}^p d_j^2 v_{kj}^2} \end{aligned}$$

$$\begin{aligned}
\therefore \pi_k = 1 &\iff \sum_{j=r+1}^p d_j^2 v_{kj}^2 = 0 \\
\therefore \pi_k = 1 &\iff v_{kj}^2 = 0 \quad \forall j \in [r+1:p] \quad \text{since } d_j^2 > 0 \quad \forall j \in [1:p] \\
\therefore \pi_k = 1 &\iff \mathbf{e}_k \perp \mathcal{V}^\perp(\mathbf{V}_r) \\
\therefore \pi_k = 1 &\iff \mathbf{e}_k \in \mathcal{V}(\mathbf{V}_r) \\
\therefore \pi_k = 1 &\iff \mathbf{e}_k \in \mathcal{L} \\
\therefore \pi_k = 1 &\iff \gamma_k = 1.
\end{aligned}$$

There exists a relationship between the adequacy and predictivity of a PCA biplot axis which always holds, irrespective of the rank of \mathbf{X} as well as whether the singular values of \mathbf{X} are all distinct or not, namely that in general, the adequacy of a biplot axis is a lower bound for the axis predictivity of the biplot axis (Gardner-Lubbe *et al.*, 2008). The derivation of this result provided below is the same as that given in Gardner-Lubbe *et al.* (2008) together with some additional steps. Consider the expressions of the k th adequacy, γ_k , and the k th axis predictivity, π_k , provided below:

$$\begin{aligned}
\gamma_k &= \sum_{j=1}^r v_{kj}^2 \\
&= \frac{\sum_{j=1}^r v_{kj}^2}{\sum_{j=1}^p v_{kj}^2} \\
\pi_k &= \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^p d_j^2 v_{kj}^2}.
\end{aligned}$$

In order to show that the k th axis predictivity is at least as great as the k th adequacy, it therefore needs to be shown that

$$\begin{aligned}
\frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^p d_j^2 v_{kj}^2} &\geq \frac{\sum_{j=1}^r v_{kj}^2}{\sum_{j=1}^p v_{kj}^2} \\
\longrightarrow \frac{\sum_{j=1}^r d_j^2 v_{kj}^2}{\sum_{j=1}^r v_{kj}^2} &\geq \frac{\sum_{j=1}^p d_j^2 v_{kj}^2}{\sum_{j=1}^p v_{kj}^2}.
\end{aligned}$$

Let S_r denote the summation,

$$S_r = \frac{\sum_{j=1}^r v_{kj}^2 d_j^2}{\sum_{j=1}^r v_{kj}^2}, \quad r \in [1:p].$$

It is evident that

$$\pi_k \geq \gamma_k \iff S_r \geq S_p.$$

It can be shown that $S_r \geq S_{r+1} \forall r \in [1 : (p-1)]$, and hence that $S_r \geq S_p \forall r \in [1 : p]$ (Gardner-Lubbe *et al.*, 2008). The derivation of this result is given below:

$$\begin{aligned} S_r - S_{r+1} &= \frac{\sum_{j=1}^r v_{kj}^2 d_j^2}{\sum_{j=1}^r v_{kj}^2} - \frac{\sum_{j=1}^{r+1} v_{kj}^2 d_j^2}{\sum_{j=1}^{r+1} v_{kj}^2} \\ &= \frac{(\sum_{j=1}^{r+1} v_{kj}^2) \sum_{j=1}^r v_{kj}^2 d_j^2 - (\sum_{j=1}^r v_{kj}^2) \sum_{j=1}^{r+1} v_{kj}^2 d_j^2}{(\sum_{j=1}^r v_{kj}^2) (\sum_{j=1}^{r+1} v_{kj}^2)} \\ &= \frac{(\sum_{j=1}^r v_{kj}^2) \sum_{j=1}^r v_{kj}^2 d_j^2 + v_{k(r+1)}^2 \sum_{j=1}^r v_{kj}^2 d_j^2}{(\sum_{j=1}^r v_{kj}^2) (\sum_{j=1}^{r+1} v_{kj}^2)} \\ &\quad - \frac{(\sum_{j=1}^r v_{kj}^2) \sum_{j=1}^r v_{kj}^2 d_j^2 + (\sum_{j=1}^r v_{kj}^2) v_{k(r+1)}^2 d_{r+1}^2}{(\sum_{j=1}^r v_{kj}^2) (\sum_{j=1}^{r+1} v_{kj}^2)} \\ &= \frac{v_{k(r+1)}^2 \sum_{j=1}^r v_{kj}^2 d_j^2 - (\sum_{j=1}^r v_{kj}^2) v_{k(r+1)}^2 d_{r+1}^2}{(\sum_{j=1}^r v_{kj}^2) (\sum_{j=1}^{r+1} v_{kj}^2)} \\ &= \frac{v_{k(r+1)}^2 (\sum_{j=1}^r v_{kj}^2 d_j^2 - \sum_{j=1}^r v_{kj}^2 d_{r+1}^2)}{(\sum_{j=1}^r v_{kj}^2) (\sum_{j=1}^{r+1} v_{kj}^2)} \\ \longrightarrow S_r - S_{r+1} &= \frac{v_{k(r+1)}^2 (\sum_{j=1}^r (d_j^2 - d_{r+1}^2) v_{kj}^2)}{(\sum_{j=1}^r v_{kj}^2) (\sum_{j=1}^{r+1} v_{kj}^2)}. \end{aligned}$$

Since the denominator in the last expression of $S_r - S_{r+1}$ is the product of two sums of squares, it can only take on non-negative values. The sign of the numerator depends only on the sign of $d_j^2 - d_{r+1}^2$ for $j \in [1 : r]$ since $v_{k(r+1)}^2$ and v_{kj}^2 are non-negative values, $j \in [1 : r]$. Since $d_j^2 \geq d_{r+1}^2$ when $j \in [1 : r]$, it follows that $d_j^2 - d_{r+1}^2 \geq 0 \forall j \in [1 : r]$ and hence that

$$\begin{aligned} S_r - S_{r+1} &\geq 0 \forall r \in [1 : (p-1)] \\ \longrightarrow S_r &\geq S_{r+1} \forall r \in [1 : (p-1)]. \end{aligned}$$

It follows that

$$S_1 \geq S_2 \geq \dots \geq S_p$$

Since the *University* data set is of full column rank, $r = 6$ is a necessary and sufficient condition for all six the biplot axes to have unit predictivities. This is confirmed by the values in Table 3.3. The fact that the biplot axis representing *Grad* in the five-dimensional PCA biplot has unit adequacy and unit axis predictivity demonstrates the fact that $r = p$ is not a necessary condition for an individual biplot axis to attain unit adequacy and axis predictivity. The adequacies and corresponding axis predictivities in Table 3.3 confirm that the adequacy of a biplot axis is a lower bound for the predictivity of that biplot axis. Consequently, when the adequacy of a biplot axis is very high, that necessarily implies that the axis predictivity of that biplot axis is also very high - consider for example the adequacy and axis predictivity of the biplot axis representing the variable *Expenses* in the four-dimensional PCA biplot. It is also evident from the values in the table that a very low adequacy does not provide much information on the corresponding axis predictivity. The biplot axis representing the variable *SAT* in the one-dimensional PCA biplot for example has a very low adequacy (0.210), but an extremely high predictivity (0.9664) whereas the biplot axis representing the variable *Expenses* has a very low adequacy (0.131) and only a moderately high axis predictivity (0.606). Table 3.3 also illustrates that an axis predictivity of one implies an adequacy of one when the data matrix under consideration is of full column rank.

It is evident from the adequacies of the biplot axes of the one-dimensional PCA biplot that all six biplot axes of the one-dimensional PCA biplot depart substantially from the corresponding Cartesian axes in the six-dimensional measurement space. Of the biplot axes of the two-dimensional PCA biplot, the axis representing the variable *Expenses* departs the least from the corresponding Cartesian axis in the measurement space. In the three-dimensional PCA biplot on the other hand it is the biplot axis representing the variable *SFRatio* that departs the least from the corresponding Cartesian axis in the measurement space. The biplot axes that represent the variable *SAT* in the PCA biplots of dimension less than or equal to five all depart substantially from the Cartesian axis that represents *SAT* in the measurement space. The large increase in the adequacy of the biplot axis representing *SAT* that results from increasing the dimension of the PCA biplot from five to six implies that the Cartesian axis representing *SAT* in the measurement space lies at a small angle to its projection onto the sixth dimension of the six-dimensional PCA biplot space - that is the one-dimensional space spanned by the sixth right singular vector of centred and standardised data matrix. On the other hand, the largest increase in the adequacy of the biplot axis representing the variable *Accept* occurs as a result of increasing the dimension of the PCA biplot from four to five, implying that the Cartesian axis representing *Accept* lies at a relatively small angle to its projection onto the fifth dimension of the PCA biplot space.

Upon consideration of the axis predictivities corresponding to the one-dimensional PCA biplot, it is clear that the biplot axis representing *SAT* produces the most accurate predictions while the biplot axis representing *Expenses* produces the least accurate predictions. The axis predictivities of the biplot axes representing *SAT* and *Expenses* in the one-dimensional PCA biplot can be interpreted as follows: 96.64 percent of the total sample variance of *SAT* is accounted for in the one-dimensional biplot while only 60.61 percent of the total sample variance of *Expenses* is accounted

for in the one-dimensional PCA biplot. Note that the quality of the predictions produced by the predictive biplot axis representing *Expenses* is substantially higher for the two-dimensional PCA biplot than for the one-dimensional PCA biplot while the quality of the predictions produced by the biplot axis representing *SAT* is almost the same for the one and two-dimensional PCA biplots. The increase in the quality of the predictions produced by the biplot axes resulting from increasing the dimension of the PCA biplot from one to two, is relatively small for *Top10* but very large for *Accept*, *SFRatio* and *Grad*. Four of the six biplot axes of the two-dimensional PCA biplot therefore produce predictions of much higher accuracy than the corresponding biplot axes of the one-dimensional PCA biplot. The fact that the predictions produced by the two-dimensional PCA biplot are more accurate on average is also reflected in the fact the overall quality of the two-dimensional PCA biplot is much greater than that of the one-dimensional PCA biplot (0.90 versus 0.769%, see Table 3.1). Except for the increase in the quality of the predictions produced by the biplot axis representing *SFRatio*, the increase in the quality of the predictions produced by the biplot axes resulting from increasing the PCA biplot's dimension from two to three is quite small, especially when compared to the increase in quality resulting from increasing the dimension from one to two. The small increase in the quality of the predictions produced by the PCA biplot resulting from increasing its dimension from two to three is also evident from the small resulting increase in the overall quality of the PCA biplot - the overall quality increases from 0.9 to 0.94 as the result of increasing the dimension of the biplot from two to three.

The increased accuracy of the predictions produced by the biplot axes resulting from increasing the dimension of the PCA biplot beyond three, is very small. Furthermore, it is not possible to visualise a biplot with dimension greater than three. Whether the *University* data set should be represented by a two or three dimensional PCA biplot depends on the investigator - both the two and three-dimensional PCA biplots are of high overall quality and in both all six biplot axes produce accurate predictions. The investigator should decide which is more important to him/her - the ease of the interpretation of the two-dimensional PCA biplot or the slightly higher overall quality of the three-dimensional PCA biplot.

Since the increase in the quality of the predictions produced by the biplot axes, resulting from the addition of the fourth dimension to the PCA biplot is very small and since the three-dimensional PCA biplot can be visualised while the four-dimensional biplot cannot be visualised, the biplot chosen to represent the *University* data set is the three-dimensional PCA biplot. Recall that this is the same conclusion as was drawn from the scree plot of the *University* data set (see Figure 3.2).

3.4.1.3 The relationship between the axis predictivities and the overall quality

The similarity between the expression of the overall quality of the PCA biplot,

$$\Omega = \frac{\text{tr}(\hat{\mathbf{X}}'\hat{\mathbf{X}})}{\text{tr}(\mathbf{X}'\mathbf{X})}$$

and the expression of the p -component vector of axis predictivities,

$$\boldsymbol{\pi} = \frac{\text{diag}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})}{\text{diag}(\mathbf{X}'\mathbf{X})}$$

indicates that there may be some connection between the p axis predictivities and the overall quality of the PCA biplot. It can in fact be shown that the overall quality of the PCA biplot equals the weighted average of the p axis predictivities, the weight assigned to each axis predictivity being proportional to the corresponding variable's sum of squares (Gardner-Lubbe *et al.*, 2008). A proof similar to that provided in the appendix of Gardner-Lubbe *et al.* (2008), is given below:

$$\begin{aligned} \pi_k &= \frac{[\widehat{\mathbf{X}}'\widehat{\mathbf{X}}]_{kk}}{[\mathbf{X}'\mathbf{X}]_{kk}} \\ \longrightarrow \pi_k [\mathbf{X}'\mathbf{X}]_{kk} &= [\widehat{\mathbf{X}}'\widehat{\mathbf{X}}]_{kk} \\ \longrightarrow \sum_{k=1}^p \pi_k [\mathbf{X}'\mathbf{X}]_{kk} &= \sum_{k=1}^p [\widehat{\mathbf{X}}'\widehat{\mathbf{X}}]_{kk} \\ \longrightarrow \sum_{k=1}^p \pi_k [\mathbf{X}'\mathbf{X}]_{kk} &= \text{tr} \{ \widehat{\mathbf{X}}'\widehat{\mathbf{X}} \} . \end{aligned}$$

Substituting $\sum_{k=1}^p \pi_k [\mathbf{X}'\mathbf{X}]_{kk}$ for $\text{tr}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})$ in the formula for the overall quality yields the following formula for the overall quality:

$$\begin{aligned} \frac{\text{tr}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})}{\text{tr}(\mathbf{X}'\mathbf{X})} &= \sum_{k=1}^p \pi_k \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})} \\ &= \sum_{k=1}^p \pi_k w_k \end{aligned} \tag{3.4.7}$$

where

$$\begin{aligned} w_k &= \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})} \\ &= \frac{[\mathbf{V}_p \mathbf{D}_p^2 \mathbf{V}_p']_{kk}}{\text{tr} \{ \mathbf{D}_p^2 \}} \\ &= \frac{c [\mathbf{V}_p \mathbf{D}_p^2 \mathbf{V}_p']_{kk}}{c \text{tr} \{ \mathbf{D}_p^2 \}} \end{aligned}$$

and where c is any constant. It is evident from equation (3.4.7) that the overall quality of the PCA biplot is equal to the weighted average of the p axis predictivities where the weight assigned to a specific axis predictivity is given by the ratio of the corresponding variable's sum of squares to the total sum of squares associated with the measured vector variable, $\underline{\mathbf{x}}$. Note that when $c = \frac{1}{n}$ and $c = \frac{1}{n-1}$, w_k equals the plug-in estimate (that is, the biased maximum likelihood estimate) and the unbiased estimate of the variance associated with x_k , respectively. It follows that the weight of π_k in the overall quality of the r -dimensional PCA biplot is equal to the ratio of the sample variance associated with the variable, x_k , to the total sample variance associated with the vector variable, $\underline{\mathbf{x}}$. It is evident that the contribution of the k th biplot axis to the overall quality of the PCA biplot is equal to

$$\pi_k w_k .$$

The relative contribution of the k th biplot axis to the overall quality follows as

$$\frac{\pi_k w_k}{\sum_{j=1}^p \pi_j w_j} .$$

It follows that if a PCA biplot is constructed from the unstandardised measurements of a data set and one of the variables has a much larger sample variance than the other variables, then that variable's relative contribution to the overall quality will be large even if the corresponding biplot axis does not have a high axis predictivity. When a variable with a sample variance which is relatively large is represented by a biplot axis with high predictivity, then the variable's relative contribution to the overall quality of the PCA biplot will also be very large. If this variable's large sample variance is only due to the fact that the variable has a large measurement unit and not due to the fact that the samples differ more with respect to this variable than with respect to the other variables, then its large contribution to the overall quality of the biplot will result in the value of the overall quality being overly optimistic.

Let α_r denote the overall quality of the r -dimensional PCA biplot and $\alpha_{(r,r+1)}$ the increase in the overall quality of the PCA biplot resulting from increasing the dimensionality of the PCA biplot from r to $r+1$, that is,

$$\alpha_{(r,r+1)} = \alpha_{r+1} - \alpha_r ,$$

where $r \in [0 : p - 1]$ and $\alpha_0 = 0$. Also, let $\pi_{k,r}$ denote the k th axis predictivity corresponding to the r -dimensional PCA biplot and $\pi_{k,(r,r+1)}$ the increase in the k th axis predictivity resulting from increasing the dimensionality of the PCA biplot from

r to $r + 1$, that is

$$\pi_{k,(r,r+1)} = \pi_{k,r+1} - \pi_{k,r}$$

where $r \in [0:p-1]$ and $\pi_{k,0} = 0$. Note that when the second subscript in the notation of the axis predictivity measure is omitted, as has been done up to now, the dimension of the PCA biplot to which the axis predictivity corresponds will be clear from the context. Similarly, if the subscript in the notation of the overall quality is omitted, the corresponding dimension will be evident from the context. Substituting $\sum_{k=1}^p \pi_{k,r} \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})}$ for α_r in the expression of $\alpha_{(r,r+1)}$ yields the following expression for $\alpha_{(r,r+1)}$:

$$\begin{aligned} \alpha_{(r,r+1)} &= \sum_{k=1}^p \pi_{k,r+1} \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})} - \sum_{k=1}^p \pi_{k,r} \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})} \\ &= \sum_{k=1}^p (\pi_{k,r+1} - \pi_{k,r}) \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})} \\ \longrightarrow \alpha_{(r,r+1)} &= \sum_{k=1}^p \pi_{k,(r,r+1)} \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})} \\ &= \sum_{k=1}^p \pi_{k,(r,r+1)} w_k \text{ where } w_k = \frac{[\mathbf{X}'\mathbf{X}]_{kk}}{\text{tr}(\mathbf{X}'\mathbf{X})}. \end{aligned}$$

It is evident that the increase in the overall quality resulting from increasing the dimensionality of the PCA biplot from r to $r + 1$ can be expressed as the weighted average of the increase in the p axis predictivities resulting from the increase in dimensionality. The contribution of the k th variable $\alpha_{(r,r+1)}$ is $\pi_{k,(r,r+1)} w_k$, while the relative contribution of the k th variable is given by

$$\frac{\pi_{k,(r,r+1)} w_k}{\sum_{k=1}^p \pi_{k,(r,r+1)} w_k}.$$

3.4.1.4 The relationship of the axis predictivities with the overall quality when the PCA biplot is constructed from the standardised measurements

When the PCA biplot is constructed from the standardised measurements, the sample variance of the x_k is equal to one for all $k \in [1:p]$, and hence each of the p axis predictivities will be assigned a weight equal to $\frac{1}{p}$. Consequently, when the PCA biplot is constructed from the standardised measurements, the overall quality of the

PCA biplot will be equal to the arithmetic average of the p axis predictivities,

$$\frac{\text{tr}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})}{\text{tr}(\mathbf{X}'\mathbf{X})} = \sum_{k=1}^p \frac{1}{p} \pi_k$$

and the relative magnitude of a biplot axis' contribution to the overall quality of the biplot will be equal to the relative magnitude of that biplot axis' predictivity,

$$\frac{\frac{1}{p} \pi_k}{\sum_{j=1}^p \frac{1}{p} \pi_j} = \frac{\pi_k}{\sum_{j=1}^p \pi_j} .$$

The magnitude of a biplot axis' predictivity indicates the quality of the axis' contribution to the overall quality of the biplot.

When the PCA biplot is constructed from the standardised measurements,

$$w_k = \frac{1}{p}$$

and hence the increase in the overall quality of the biplot resulting from an increase in the dimension of the biplot is equal to the arithmetic average of the increases in the p axis predictivities resulting from the increase in dimension. It follows that when the PCA biplot is constructed from the standardised measurements, a plot showing the overall quality of the biplot as well as each of the p axis predictivities against the possible dimensionalities of the biplot is useful in that it allows for the visual assessment of both the relative contributions of the biplot axes to the overall quality as well as the qualities of those contributions. When the dimension of a PCA biplot is increased from r to $r + 1$, the contribution of the k th variable to the resulting increase in the overall quality is

$$\frac{1}{p} \pi_{k,(r,r+1)}$$

while its relative contribution is

$$\frac{\pi_{k,(r,r+1)}}{\sum_{k=1}^p \pi_{k,(r,r+1)}} .$$

Hence, when the PCA biplot is constructed from the standardised measurements, the relative magnitude of the k th variable's contribution to $\alpha_{(r,r+1)}$ is equal to the relative magnitude of $\pi_{k,(r,r+1)}$. Notice that the gradient of the line connecting the points corresponding to $\pi_{k,r}$ to $\pi_{k,r+1}$ in the plot showing the overall quality and p axis predictivities against the dimension of the PCA biplot is equal to $\pi_{k,(r,r+1)}$. The plot of the overall quality and axis predictivities against the dimension of the biplot therefore also allows for the visual assessment of the relative contributions of the biplot axes to $\alpha_{(r,r+1)}$ for $r \in [1 : p - 1]$ and the qualities of these contributions. A plot such as the one just described is illustrated in Figure 3.6 for the *University* data set. It is evident from Figure 3.6 that the variable whose relative contribution to the overall quality of the one-dimensional PCA biplot of the *University* data set is the greatest, is the variable, *SAT*, while the variable whose relative contribution to the overall quality of the one-dimensional PCA biplot is the smallest, is the variable, *Expenses*.

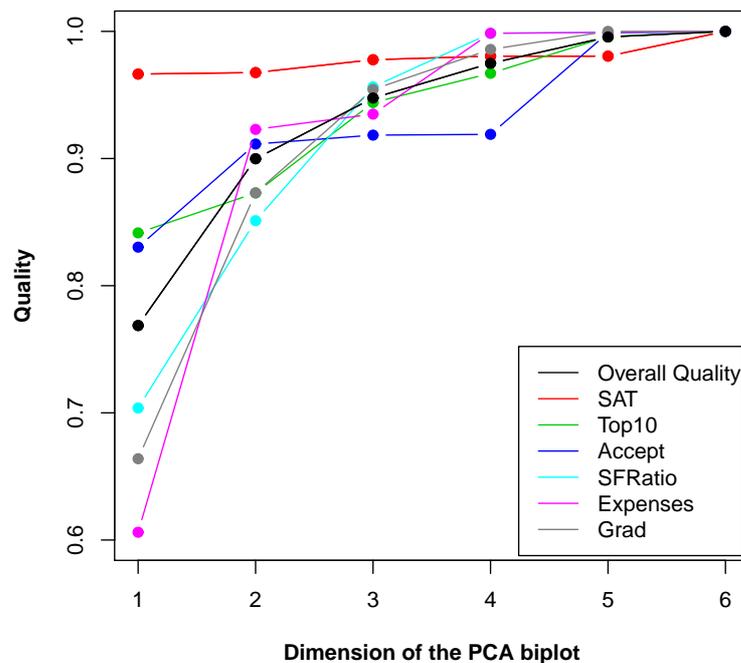


Figure 3.6: The overall quality and axis predictivities of the PCA biplot constructed from the standardised measurements of the *University* data set.

When the construction of the PCA biplot is based on the unstandardised data matrix, the interpreter of the biplot needs to take both the axis predictivity as well as the weight assigned to that axis predictivity into account when assessing the axis' relative contribution to the overall quality of the biplot. A plot showing the overall quality and each of the p axis predictivities against the possible dimensions of the PCA biplot constructed from the unstandardised measurements of a data

set therefore does not allow for the visualisation of the relative contributions of the various biplot axes to the overall quality of the biplot or the relative contributions of the biplot axes to $\alpha_{(r,r+1)}$ - it only allows for the visualisation of the qualities of these contributions. When the PCA biplot is constructed from the unstandardised measurements, a plot of the overall quality together with the contributions of the p biplot axes to the overall quality can be made, but this plot would be uninformative as to what the qualities of the contributions are concerned.

Consider the slopes of the six straight lines in Figure 3.6 illustrating the increase in the six axis predictivities resulting from increasing the dimension of the PCA biplot from one to two. It is evident that the increase in the overall quality of the PCA biplot resulting from the increased dimension is mainly due to the increased accuracy with which the variables, *Expenses*, *SFRatio* and *Grad* are predicted, or put differently, the addition of the second dimension to the PCA biplot contributed mainly to the increased accuracy with which the variables, *Expenses*, *SFRatio* and *Grad* are predicted. This means that the second principal component accounts for a large proportion of the sample variances of these three variables. The relative contribution of the variable, *Expenses*, to the increase in the overall quality is the greatest of all the variables. When the dimension of the biplot is increased from two to three, it is the variable *SFRatio* whose relative contribution to the increase in the overall quality of the biplot which is the greatest.

3.4.1.5 Axis predictivities and the interpretation of the PCA biplot

The fact that the overall quality of the PCA biplot is equal to an average of the p axis predictivities implies that a high overall quality does not necessarily suggest that the measurements of all p the variables are accurately approximated in the biplot. Similarly, a low overall quality does not necessarily suggest that the measurements of all p the variables are poorly approximated in the biplot. The average (weighted or arithmetic) of a number of very high axis predictivities together with a few low axis predictivities, can still be very high and similarly, the average of a number of very low axis predictivities together with a few high axis predictivities, can still be very low. This is especially true when the PCA biplot is constructed from the unstandardised measurements - one or two very high axis predictivities accompanied by very large weights in the overall quality measure can cause the overall quality of the PCA biplot to be very high even if all the other axis predictivities are very low. Remembering that a low axis predictivity means that the measurements of the corresponding variable are poorly approximated by the PCA biplot, it is clear that no conclusions regarding a variable represented by a biplot axis with low predictivity should be drawn from the visual inspection of the PCA biplot. The predictivities of the individual biplot axes should therefore be considered before drawing conclusions from the visual inspection of the PCA biplot, even when the overall quality of the biplot is very high. Similarly the weighted or arithmetic average of a lot of very low axis predictivities together with a few high axis predictivities can still be relatively low. The axis predictivities of a PCA biplot with low overall quality should therefore be considered before discarding the biplot - useful information can be gathered regarding the variables whose measurements are accurately approximated in the biplot.

It is very important to consider the values of the axis predictivities prior to drawing conclusions regarding the relationships between the measured variables based on visual inspection of the biplot alone. Only when two biplot axes both have very high axis predictivities can it safely be concluded that the relationship between the approximations of the measurements on the two variables which are read off from the (predictive) biplot axes, accurately represents reality.

Recall that unlike the correlation biplot, the PCA biplot is not designed to optimally represent the correlations between the measured variables via the angles between the corresponding pairs of biplot axes. The angle between a pair of PCA biplot axes may therefore in some circumstances seem to suggest information about the correlation between the corresponding two variables that is misleading. When however the j th and k th biplot axes have very high axis predictivities and the angle between the increasing parts of the two axes is close to zero degrees or 180 degrees, the increasing part of an axis being the part that stretches from the origin in the direction in which the calibrations on the axis increase, then it is safe to conclude that the j th and k th measured variables are strongly correlated. The reason for this is that, when the angle between the increasing parts of the j th and k th biplot axes is very small, then an increase in the approximation read off from the j th biplot axis almost always implies an increase in the approximation read off from the k th biplot axis. Similarly, when the angle between the increasing parts of the j th and k th biplot axes is close to 180 degrees then an increase in the approximation read off from the j th biplot axis almost always implies a decrease in the approximation read off from the k th biplot axis. In these situations, if the axis predictivities of both the j th and the k th biplot axes are very high, then it is safe to conclude that the relationship between the approximated measurements is an accurate representation of reality. If it is of interest to the investigator to visualise the correlations between the measured variables, then he/she can construct a correlation biplot or a correlation monoplot (Gower *et al.*, 2011) which is designed specifically for that purpose.

As an example consider the two-dimensional PCA biplot constructed from the standardised measurements of the National Track data set provided in Figure 3.7. Consider the angles between the various pairs of biplot axes together with the corresponding correlation coefficients provided in Table 3.4 as well as the corresponding axis predictivities provided in Table 3.5.

Consider the large angle between the biplot axes representing the variables $200m$ and $5000m$ as well as the large angle between the axes representing the variables $200m$ and $10000m$. Note that the three biplot axes representing $200m$, $5000m$ and $10000m$ have very high axis predictivities. If the angles between these pairs of biplot axes were to be interpreted to be indicative of the strength of the correlation between the corresponding variables, then it would have been concluded the variables $200m$ and $5000m$ as well as the pair of variables $200m$ and $10000m$ are almost uncorrelated. It is however evident from the correlation coefficients in Table 3.4 that both these pairs of variables are quite strongly positively correlated. On the other hand, consider the small angle between the two biplot axes representing the variables $5000m$ and 10000 respectively. Since both these biplot axes have very high axis predictivities, it can be expected that the two variables are strongly positively correlated. This expectation is confirmed by the very high correlation coefficient be-

tween these two variables (0.97). The same can be said about the pair of variables *100m* and *200m*.

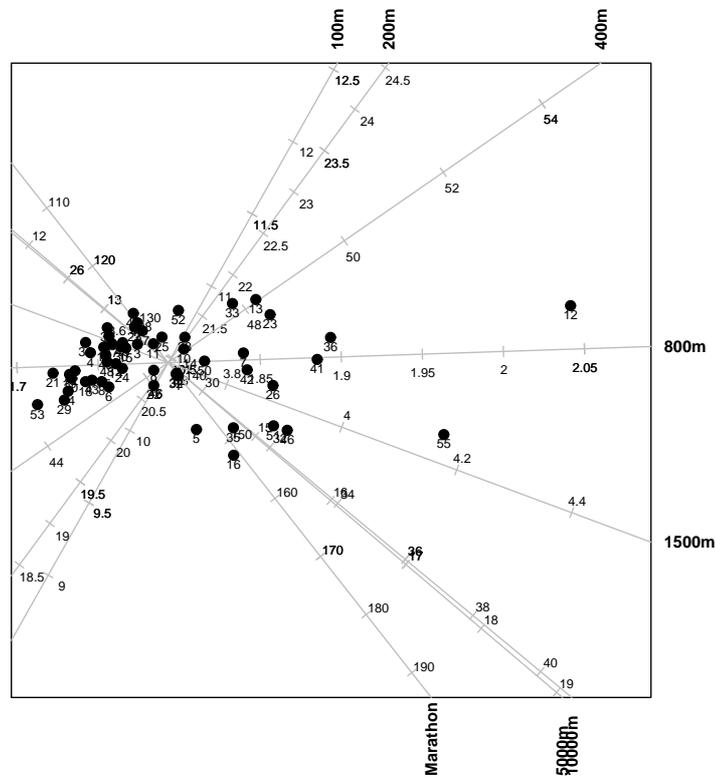


Figure 3.7: The two-dimensional predictive PCA biplot constructed from the standardised measurements of the National Track data set.

Table 3.4: The sample correlation matrix associated with the National Track data set.

	100m	200m	400m	800m	1500m	5000m	10000m	Marathon
100m	1.00	0.92	0.84	0.76	0.70	0.62	0.63	0.52
200m	0.92	1.00	0.85	0.81	0.77	0.70	0.70	0.60
400m	0.84	0.85	1.00	0.87	0.84	0.78	0.79	0.70
800m	0.76	0.81	0.87	1.00	0.92	0.86	0.87	0.81
1500m	0.70	0.77	0.84	0.92	1.00	0.93	0.93	0.87
5000m	0.62	0.70	0.78	0.86	0.93	1.00	0.97	0.93
10000m	0.63	0.70	0.79	0.87	0.93	0.97	1.00	0.94
Marathon	0.52	0.60	0.70	0.81	0.87	0.93	0.94	1.00

Table 3.5: *The axis predictivities corresponding to the two-dimensional PCA biplot constructed from the standardised measurements of the National Track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
0.950	0.939	0.892	0.900	0.938	0.965	0.974	0.943

3.4.1.6 The scale dependence of the PCA biplot, overall quality, axis predictivities and adequacies: an illustrative example

Consider the *National track* data set provided in Table 8.6 in Johnson and Wichern (2002). This data set provides information on the national track records of men in 55 countries. For each country, measurements on eight variables namely 100m, 200m, 400m, 800m, 1500m, 5000m, 10000m and *marathon*, are reported. The first three variables (100m, 200m and 400m) are measured in seconds while the other five variables are measured in minutes.

Table 3.6 contains the standard deviations of the eight measured variables. Due to the fact that the variables have widely differing standard deviations it is expected that the PCA biplots constructed from the standardised and unstandardised measurements of the data set will differ greatly. As these differences can be expected to be more substantial for the lower dimensional biplots, the axis predictivities, adequacies and overall qualities of the one-dimensional PCA biplots constructed from the standardised and unstandardised measurements of the *National Track* data respectively, will be compared to illustrate some of the concepts discussed thus far.

Table 3.6: *The standard deviations of the eight measured variables of the National Track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
0.3514	0.6446	1.4570	0.0637	0.1559	0.8012	1.8077	9.2270

Table 3.7: *The axis predictivities corresponding to the one-dimensional PCA biplot constructed from the unstandardised measurements of the National Track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
0.2873	0.3736	0.5184	0.6676	0.7656	0.8811	0.9024	0.9994

Table 3.7 contains the axis predictivities of the biplot axes of the one-dimensional PCA biplot constructed from the unstandardised measurements of the *National Track* data set. The very high axis predictivity of the biplot axis representing the variable *Marathon* is attributable to the very large relative magnitude of that variable's standard deviation. Due to the extremely large standard deviation of the variable *Marathon* relative to those of the other variables, the *Marathon* dominates the first principal component, that is the coefficient vector of the first principal component lies very close to the Cartesian axis which represents *Marathon* in the

eight-dimensional measurement space. This is evident from the very large relative magnitude of the coefficient associated with *Marathon* in the first principal component (Table 3.8) as well as from the very high adequacy of the biplot axis representing *Marathon* relative to those of the other biplot axes (Table 3.9). Due to the fact that the first principal component lies so close to the Cartesian axis which represents the variable *Marathon* in the eight-dimensional measurement space, the representation of the measurements on the variable *Marathon* in the one-dimensional PCA biplot will be very similar to the exact one-dimensional representation of the measurements on the variable *Marathon*.

Table 3.8: *The coefficients of the first principal component of the unstandardised national track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
-0.02	-0.042	-0.111	-0.006	-0.014	-0.079	-0.181	-0.973

Table 3.9: *The adequacies of the eight biplot axes in the one-dimensional PCA biplot constructed from the unstandardised measurements of the National track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
0.000	0.002	0.012	0.000	0.000	0.006	0.033	0.946

Note that the zero adequacies in Table 3.9 are not truly zero - they only appear to be zero due to rounding. Recall that a truly zero adequacy implies a zero axis predictivity (see Section 3.4.1). The fact that none of the axis predictivities in Table 3.9 are equal to zero confirms that none of the adequacies are exactly zero. If the first principal axis and the Cartesian axis representing the variable *Marathon* in the eight-dimensional measurement space were collinear, then all seven the other Cartesian axes would have been orthogonal to the one-dimensional PCA biplot space, and hence the adequacies of the corresponding seven biplot axes would have been zero. Since the first principal axis does not lie exactly in the direction of the Cartesian axis representing the variable *Marathon* in the eight-dimensional measurement space, the other seven Cartesian axes only lie at angles close to, but not exactly, 90° to the one-dimensional biplot space and hence have adequacies close to, but not exactly, zero.

Table 3.10: *The weights of the axis predictivities in the expression of the overall quality of the PCA biplot constructed from the unstandardised measurements of the National Track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
0.0013	0.0045	0.0231	0.0000	0.0003	0.0070	0.0356	0.9281

Upon consideration of the weights of the axis predictivities in the calculation of the overall quality of the PCA biplot provided in Table 3.10, it is evident that the

variable *Marathon* has a very large weight compared to the other variables. This is the result of the large relative magnitude of the standard deviation of the variable *Marathon*. The very large relative magnitude of the weight of the axis predictivity of the biplot axis representing the variable *Marathon* in the calculation of the overall quality together with the fact that the axis predictivity of this biplot axis is much higher than those of most of the other biplot axes, implies that the overall quality in Table 3.11 is overly optimistic.

Table 3.11: *The overall qualities corresponding to the one-dimensional PCA biplots constructed from the unstandardised and standardised measurements of the National Track data respectively.*

Unstandardised	Standardised
0.98	0.83

When the PCA biplot is constructed from the standardised measurements of the *National Track* data set, the first principal component is not dominated by the variable *Marathon* - in fact, the eight variables contribute almost equally to the first principal component. This is evident from the coefficients of the first principal component provided in Table 3.12 as well as from the adequacies of the eight biplot axes provided in Table 3.13.

Table 3.12: *The coefficients of the first principal component of the standardised national track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
-0.318	-0.337	-0.356	-0.369	-0.373	-0.364	-0.367	-0.342

Table 3.13: *The adequacies of the eight biplot axes of the one-dimensional PCA biplot constructed from the standardised measurements of the National Track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
0.1008	0.1136	0.1265	0.1359	0.1390	0.1328	0.1345	0.1169

Table 3.14: *The axis predictivities of the eight biplot axes of the one-dimensional PCA biplot constructed from the standardised measurements of the National Track data set.*

100m	200m	400m	800m	1500m	5000m	10000m	Marathon
0.6678	0.7520	0.8376	0.9001	0.9204	0.8792	0.8908	0.7742

Table 3.14 contains the axis predictivities of the eight biplot axes of the one-dimensional PCA biplot constructed from the standardised measurements of the

National Track data set. Since the eight standardised variables carry equal sized weights in the calculation of the overall quality of the biplot and hence the overall quality provided in Table 3.11 is equal to the arithmetic average of the eight axis predictivities in Table 3.14.

3.4.1.7 Changing the PCA biplot scaffolding axes

Up to now the r -dimensional PCA biplot has been defined as the r -dimensional biplot constructed from the first r principal components, $r \in [1 : p]$. Any r of the principal components can however be used to construct an r -dimensional PCA biplot. In the situation where there is a variable (or set of variables) that is of particular interest to the investigator but is poorly predicted in the two (or three) dimensional PCA biplot constructed from the first two (or three) principal components, the investigator can in addition to the existing biplot, construct a two (or three) dimensional PCA biplot from a different subset of two (or three) principal components in which the variable (or set of variables) of interest is more accurately predicted. Consider the following expression of the axis predictivity of the k th biplot axis of the r -dimensional PCA biplot constructed from the first r principal components:

$$\begin{aligned}\pi_{k,r} &= \frac{\mathbf{x}'_{(k)} \mathbf{V}_r \mathbf{V}'_r \mathbf{x}_{(k)}}{\mathbf{x}'_{(k)} \mathbf{x}_{(k)}} \\ \longrightarrow \pi_{k,r} &= \frac{\sum_{i=1}^r \left(\mathbf{x}'_{(k)} \mathbf{v}_{(i)} \right)^2}{\mathbf{x}'_{(k)} \mathbf{x}_{(k)}}.\end{aligned}$$

It is evident that the contribution of the r th principal component to $\pi_{k,r}$ is

$$\frac{\left(\mathbf{x}'_{(k)} \mathbf{v}_{(r)} \right)^2}{\mathbf{x}'_{(k)} \mathbf{x}_{(k)}} = \pi_{k,r} - \pi_{k,r-1} \quad (3.4.8)$$

where $\pi_{k,r}$ denotes the k th axis predictivity of the r -dimensional PCA biplot constructed from the first r principal components as before. Note that the ratio in (3.4.8) is equal to the proportion of the sample variance of x_k that is accounted for by the r th principal component. It follows that the two-dimensional PCA biplot in which the measurements on the k th variable will be most accurately predicted is the biplot constructed from the two principal components with ranks given by the two values of r for which the difference $\pi_{k,r} - \pi_{k,r-1}$ is largest. If for example the j th and k th variables are of particular interest to the investigator and he/she wants to construct a two-dimensional PCA biplot in which these two variables are more accurately predicted than in the biplot constructed from the first two principal components, he/she can construct a PCA biplot from the two principal components with ranks given by the two values of r for which the differences $\pi_{j,r} - \pi_{j,r-1}$ and

$\pi_{k,r} - \pi_{k,r-1}$ are optimal. In the remainder of this chapter, whenever an r -dimensional PCA biplot is constructed from a set of r principal components other than the set of the first r principal components, the $p \times r$ orthonormal matrix with column vectors equal to the r right singular vectors corresponding to the r principal components that were used to construct the biplot, will be denoted by $\mathbf{V}_r^\#$.

The procedure used to construct an r -dimensional PCA biplot from a set of r principal components other than the set of the first r principal components is exactly the same as that stipulated in Chapter 2 with the $p \times r$ matrix $\mathbf{V}_r^\#$ being substituted for the matrix \mathbf{V}_r . The matrix $\mathbf{V}_r^\#$ is also substituted for \mathbf{V}_r in the definition of the axis predictivities corresponding to the resulting PCA biplot.

This concept will now be illustrated at the hand of a simulated data set consisting of 1000 samples, each measured on one of five variables with correlation structure provided in Table 3.15. Table 3.16 provides the contributions of each of the five principal components to the sample variances of the standardised variables of the simulated data set, that is, the (rk) th element of Table 3.16 is equal to $\pi_{k,r} - \pi_{k,r-1}$, $r, k \in [1 : p]$.

Table 3.15: *The sample correlation matrix corresponding to the simulated data set.*

	X1	X2	X3	X4	X5
X1	1.0000	0.8649	-0.0028	0.0492	0.0724
X2	0.8649	1.0000	0.0952	0.2889	0.1323
X3	-0.0028	0.0952	1.0000	0.7947	0.0905
X4	0.0492	0.2889	0.7947	1.0000	0.1160
X5	0.0724	0.1323	0.0905	0.1160	1.0000

Table 3.16: *The contributions of the principal components (PCs) to the sample variances of the standardised variables of the simulated data set.*

	X1	X2	X3	X4	X5
PC 1	0.4785	0.6683	0.3693	0.5098	0.0726
PC 2	0.4493	0.2721	0.5161	0.3886	0.0010
PC 3	0.0101	0.0050	0.0108	0.0092	0.9262
PC 4	0.0376	0.0241	0.0973	0.0792	0.0002
PC 5	0.0245	0.0305	0.0065	0.0132	0.0000

Upon inspection of the values in Table 3.16, it is evident that the first two principal components are the two principal components that account for most of the variation in the variables X_1 , X_2 , X_3 and X_4 , while the first and third principal components are the two principal components that account for most of the variation in the variable X_5 . It follows that the investigator can use the two-dimensional PCA biplot constructed from the first and second principal components shown in Figure 3.8(a) to visualise more accurate information on the variables X_1 , X_2 , X_3 and

X_4 and the PCA biplot constructed from the first and third principal components shown in Figure 3.8(b) to visualise more accurate information on the variable X_5 . The axis predictivities corresponding to these two PCA biplots are provided in the rows of Table 3.17 labeled a and b respectively. It is evident that while the measurements on the variables X_1 , X_2 , X_3 and X_4 are very accurately predicted in the PCA biplot constructed from the first two principal components, the measurements on the variable X_5 are very poorly predicted. On the other hand, in the PCA biplot constructed from the first and third principal components, the measurements on the variable X_5 are very accurately predicted while those of X_2 are predicted only moderately accurately and those of X_1 , X_3 and X_4 are predicted poorly.

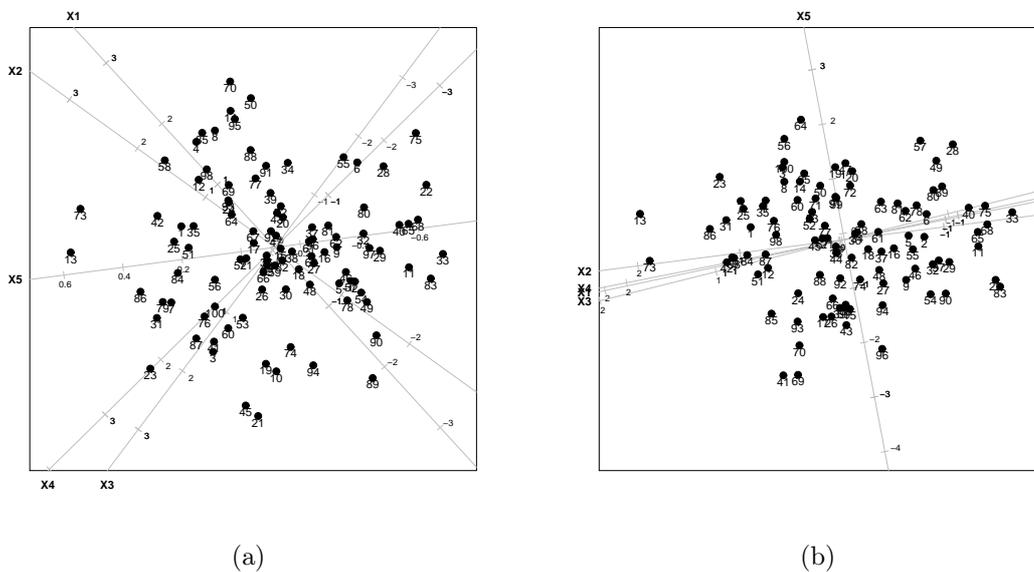


Figure 3.8: (a) The two-dimensional PCA biplot constructed from the first two principal components of the standardised simulated data set; (b) The two-dimensional PCA biplot constructed from the first and third principal components of the standardised simulated data set.

Table 3.17: (a) The axis predictivities corresponding to the two-dimensional PCA biplot constructed from the first two principal components of the standardised measurements of the simulated data set; (b) The axis predictivities corresponding to the two-dimensional PCA biplot constructed from the first and third principal components of the standardised measurements of the simulated data set.

	X1	X2	X3	X4	X5
a	0.9278	0.9404	0.8854	0.8984	0.0736
b	0.4886	0.6733	0.3801	0.5190	0.9988

In the remainder of this chapter it will be assumed that unless explicitly stated otherwise, the r -dimensional PCA biplot is constructed from the first r principal

components.

3.4.2 Sample predictivities

3.4.2.1 Definition and properties

The sample predictivity of the i th sample (or i th sample predictivity for short) measures the overall accuracy of the approximations of that sample's p measurements which are read off from the p predictive PCA biplot axes. It is defined as the ratio of the sum of the squared predicted measurements to the sum of the squared true measurements. Denoting the i th sample predictivity by ψ_i , then the i th sample predictivity corresponding to the r -dimensional PCA biplot can be expressed as:

$$\psi_i = \frac{\hat{\mathbf{x}}_i' \hat{\mathbf{x}}_i}{\mathbf{x}_i' \mathbf{x}_i} \quad (3.4.9)$$

$$\begin{aligned} &= \frac{[\hat{\mathbf{X}}\hat{\mathbf{X}}']_{ii}}{[\mathbf{X}\mathbf{X}']_{ii}} \\ &= \frac{[\mathbf{U}_r \mathbf{D}_r^2 \mathbf{U}_r']_{ii}}{[\mathbf{U}_q \mathbf{D}_q^2 \mathbf{U}_q']_{ii}} \\ \rightarrow \psi_i &= \frac{\sum_{j=1}^r d_j^2 u_{ij}^2}{\sum_{j=1}^q d_j^2 u_{ij}^2}. \end{aligned} \quad (3.4.10)$$

Given that the decomposition of \mathbf{X} into $\hat{\mathbf{X}}$ and $(\mathbf{X} - \hat{\mathbf{X}})$ exhibits Type A orthogonality, the i th sample predictivity as defined in equation (3.4.9) is meaningful as a quality measure. The n -component vector of sample predictivities with i th element equal to ψ_i , can be expressed as:

$$\begin{aligned} \boldsymbol{\psi} &= \text{diag}(\hat{\mathbf{X}}\hat{\mathbf{X}}') [\text{diag}(\mathbf{X}\mathbf{X}')]^{-1} \\ \rightarrow \boldsymbol{\psi} &= \text{diag}(\mathbf{U}_r \mathbf{D}_r^2 \mathbf{U}_r') [\text{diag}(\mathbf{U}_q \mathbf{D}_q^2 \mathbf{U}_q')]^{-1}. \end{aligned}$$

Given that the sample predictivity measure is defined as a ratio of sums of squares, it can only take on non-negative values. From the expression of ψ_i in (3.4.10) it is evident that ψ_i is a non-decreasing function of the dimension of the biplot space, r with a maximum value which is necessarily obtained when $r = q$ and a minimum value of zero which is obtained if and only if $\hat{\mathbf{x}}_i = \mathbf{0}$ i.e. if and only if $\mathbf{x}_i \in \mathcal{L}^\perp$. It is shown below that the i th sample predictivity is a decreasing function of the size of the angle between the vector \mathbf{x}_i emanating from the origin and the

projection of this vector onto the biplot space:

$$\begin{aligned}
 \psi_i &= \frac{\hat{\mathbf{x}}_i' \hat{\mathbf{x}}_i}{\mathbf{x}_i' \mathbf{x}_i} \\
 &= \frac{\mathbf{x}_i \mathbf{V}_r \mathbf{V}_r' \mathbf{V}_r \mathbf{V}_r' \mathbf{x}_i}{\|\mathbf{x}_i\|^2} \\
 &= \frac{\mathbf{x}_i \mathbf{V}_r \mathbf{V}_r' \mathbf{x}_i}{\|\mathbf{x}_i\|^2} \\
 &= \left(\frac{(\mathbf{x}_i \mathbf{V}_r \mathbf{V}_r' \mathbf{x}_i)^{1/2}}{\|\mathbf{x}_i\|} \right)^2 \\
 &= \left(\frac{(\mathbf{x}_i \mathbf{V}_r \mathbf{V}_r' \mathbf{x}_i)}{(\mathbf{x}_i \mathbf{V}_r \mathbf{V}_r' \mathbf{x}_i)^{1/2} \|\mathbf{x}_i\|} \right)^2 \\
 &= \left(\frac{(\mathbf{x}_i \mathbf{V}_r \mathbf{V}_r' \mathbf{x}_i)}{\|\mathbf{x}_i \mathbf{V}_r \mathbf{V}_r'\| \|\mathbf{x}_i\|} \right)^2 \\
 \longrightarrow \psi_i &= \cos^2(\theta_{\mathbf{x}_i, \hat{\mathbf{x}}_i}) \tag{3.4.11}
 \end{aligned}$$

where $\cos^2(\theta_{\mathbf{x}_i, \hat{\mathbf{x}}_i})$ is the angle between the vectors \mathbf{x}_i and $\hat{\mathbf{x}}_i$ both emanating from the origin. The expression of ψ_i in (3.4.11) shows that

$$\psi_i = 1 \iff \mathbf{x}_i \in \mathcal{L}$$

and confirms that

$$\psi_i = 0 \iff \mathbf{x}_i \in \mathcal{L}^\perp.$$

It is shown below that when $r = q$, the biplot space is identical to the row space of \mathbf{X} , $\mathcal{V}(\mathbf{X}')$:

$$\begin{aligned}
 \mathcal{V}(\mathbf{X}') &= \mathcal{V}(\mathbf{X}'\mathbf{X}) \\
 &= \mathcal{V}(\mathbf{V}_q \mathbf{D}_q^2 \mathbf{V}_q') \\
 &= \mathcal{V}(\mathbf{V}_q \mathbf{D}_q \mathbf{D}_q \mathbf{V}_q') \\
 &= \mathcal{V}(\mathbf{V}_q \mathbf{D}_q) \\
 \longrightarrow \mathcal{V}(\mathbf{X}') &= \mathcal{V}(\mathbf{V}_q).
 \end{aligned}$$

It follows that when $r = q$ each of the n row vectors of \mathbf{X} is contained in the biplot

space and hence each of the n sample predictivities will equal unity.

Due to the fact that the decomposition of \mathbf{X} into $\widehat{\mathbf{X}}$ and $\mathbf{X} - \widehat{\mathbf{X}}$ exhibits Type A orthogonality, the i th sample predictivity can be expressed as a decreasing function of the Pythagorean distance between the points representing the i th sample in the measurement space and the biplot space respectively:

$$\psi_i = 1 - \frac{(\mathbf{x}_i - \widehat{\mathbf{x}}_i)'(\mathbf{x}_i - \widehat{\mathbf{x}}_i)}{\mathbf{x}_i' \mathbf{x}_i}. \quad (3.4.12)$$

Since γ_i is a non-decreasing function of the dimension of the biplot space, it follows that the Pythagorean distance between the points representing the i th sample in the measurement space and the biplot space respectively, as well as the angle between those two points, are non-increasing functions of the dimension of the biplot space.

It is evident from equation (3.4.11) that when $\psi_i > \psi_j$, the angle between the points representing the i th sample in the measurement space and the biplot space respectively, is necessarily smaller than the angle between the corresponding two points associated with the j th sample. Equation (3.4.12) shows that it is only when lengths of the vectors \mathbf{x}_i and \mathbf{x}_j are identical that $\psi_i > \psi_j$ necessarily implies that the Pythagorean distance between the two points representing the i th sample in the measurement space and the biplot space respectively is smaller than the Pythagorean distance between the corresponding two points associated with the j th sample.

The position of a sample with a low sample predictivity in the biplot is meaningless and therefore no conclusions about it or its relationships to the other samples should be drawn based on its position in the biplot. When for example a sample with low predictivity seems to be an outlier based on its position in the biplot, it cannot be concluded that the sample is in fact an outlier. A possible method to detect outliers using sample predictivities is discussed in Section 3.4.2.2. The sample predictivities of the individual samples should therefore always be considered prior to drawing conclusions about the samples and the relationships between the samples based on visual inspection of the biplot alone.

In Section 2.7.1.1 it was explained how to interpolate a new sample onto the PCA biplot. The predictivity of the new interpolated sample can be calculated in exactly the same way as the predictivities of the 'old' samples (that is the samples upon which the construction of the PCA biplot was based). Recall that, before a new observed sample, \mathbf{x}^* , can be interpolated onto an existing biplot, the measurements of that sample need to be transformed to the same scales as those of the elements in the centred matrix, \mathbf{X} upon which the construction of the biplot is based. This is achieved by subtracting from each element of the new sample, the mean of the observed measurements of the samples used in the construction of the biplot on the corresponding variable. The transformed sample is then projected orthogonally onto the biplot space based on the samples contained in \mathbf{X} , that is $\mathcal{V}(\mathbf{V}_r)$. It follows

that the sample predictivity of the new observed sample, \mathbf{x}^* is given by

$$\frac{(\mathbf{x}^{*'}\mathbf{V}_r\mathbf{V}_r' - \bar{\mathbf{x}}'\mathbf{V}_r\mathbf{V}_r')(\mathbf{x}^{*'}\mathbf{V}_r\mathbf{V}_r' - \bar{\mathbf{x}}'\mathbf{V}_r\mathbf{V}_r')'}{(\mathbf{x}^* - \bar{\mathbf{x}})'(\mathbf{x}^* - \bar{\mathbf{x}})}.$$

3.4.2.2 Using sample predictivities to detect outliers

Numerous methods have been proposed to detect outliers in data sets. One of these is to use plots of the first few principal components that account for most of the variability in the data set (i.e. the major principal components) as well as plots of the last few principal components that account for a negligible proportion of the variability in the data set (i.e. the minor principal components) to detect outliers. Samples which are outliers on the first few principal components are typically outliers with respect to one or more of the individual measured variables (Gnanadesikan and Kettinger (1972), Jolliffe (2002)). Such outliers can also be detected from plots of the measured variables - individually or in pairs. Samples that are outliers on the last few principal components are usually samples which deviate substantially from the correlation structure of the bulk of the data (Jolliffe (2002)). The remainder of this section will focus on the detection of samples that do not conform with the correlation structure of the bulk of the data set. Since the coordinates of the points that represent the samples in a PCA biplot constructed from the i th and j th principal components are given by the i th and j th principal component scores of those samples, PCA biplots constructed from the last few principal components can be used to detect samples that deviate substantially from the correlation structure of the bulk of the data set.

If a sample has a very low sample predictivity corresponding to the PCA biplot constructed from the major principal components but a moderately high or very high sample predictivity corresponding to the PCA biplot constructed from the minor principal components, then that sample most likely deviates substantially from the correlation structure of the bulk of the data set.

The i th sample predictivity corresponding to the r -dimensional PCA biplot constructed from the first r principal components, $\psi_{i,r}$, can be partitioned into the individual contributions of the first r principal components:

$$\begin{aligned}\psi_{i,r} &= \frac{\mathbf{x}_i'\mathbf{V}_r\mathbf{V}_r'\mathbf{x}_i}{\mathbf{x}_i'\mathbf{x}_i} \\ \rightarrow \psi_{i,r} &= \sum_{j=1}^r \frac{(\mathbf{x}_i'\mathbf{v}_j)^2}{\mathbf{x}_i'\mathbf{x}_i}.\end{aligned}$$

It is evident that the ratio

$$\begin{aligned} \frac{(\mathbf{x}'_i \mathbf{v}_{(r)})^2}{\mathbf{x}'_i \mathbf{x}_i} &= \sum_{j=1}^r \frac{(\mathbf{x}'_i \mathbf{v}_{(j)})^2}{\mathbf{x}'_i \mathbf{x}_i} - \sum_{j=1}^{r-1} \frac{(\mathbf{x}'_i \mathbf{v}_{(j)})^2}{\mathbf{x}'_i \mathbf{x}_i} \\ &= \psi_{i,r} - \psi_{i,r-1} \end{aligned} \quad (3.4.13)$$

is the contribution of the r th principal component to the i th sample predictivity corresponding to the r -dimensional PCA biplot constructed from the first r principal components. The ratio in (3.4.13) is also equal to the i th sample predictivity corresponding to the one-dimensional PCA biplot constructed from the r th principal component. Similarly,

$$\frac{(\mathbf{x}'_i \mathbf{v}_{(p)})^2}{\mathbf{x}'_i \mathbf{x}_i} + \frac{(\mathbf{x}'_i \mathbf{v}_{(p-1)})^2}{\mathbf{x}'_i \mathbf{x}_i} = \psi_{i,p} - \psi_{i,p-2}$$

is equal to the i th sample predictivity corresponding to the two-dimensional PCA biplot constructed from the last two principal components. In general, the i th sample predictivity corresponding to an r -dimensional PCA biplot constructed from a set of r principal components other than the set of the first r principal components, is defined as in (3.4.9) but with $\hat{\mathbf{x}}_i = \mathbf{x}'_i \mathbf{V}_r^\# (\mathbf{V}_r^\#)'$.

As an example, consider the individual contributions of the eight principal components to the sample predictivity associated with Greece provided in Table 3.18. The sample predictivity associated with Greece corresponding to the PCA biplot constructed from the first two principal components, which account for approximately 94% of the variation in the data set, is very low while the sample predictivity associated with Greece corresponding to the PCA biplot constructed from the last two principal components, which account for approximately 0.9% of the variation in the data set, is moderately high. This suggests that the measurement vector associated with Greece deviates substantially from the correlation structure of the rest of the data set. Upon consideration of the PCA biplot constructed from the last two principal components of the standardised *National Track* data set in Figure 3.9, it is evident that Greece is an outlier on the last two principal components, confirming that the measurement vector associated with Greece does not conform with the correlation structure of the rest of the data set.

Table 3.18: *The individual contributions of the eight principal components to the sample predictivity associated with Greece corresponding to the PCA biplot constructed from the standardised measurements of the National Track data set.*

Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8
0.0788	0.1982	0.0178	0.0327	0.0818	0.0092	0.3674	0.2141

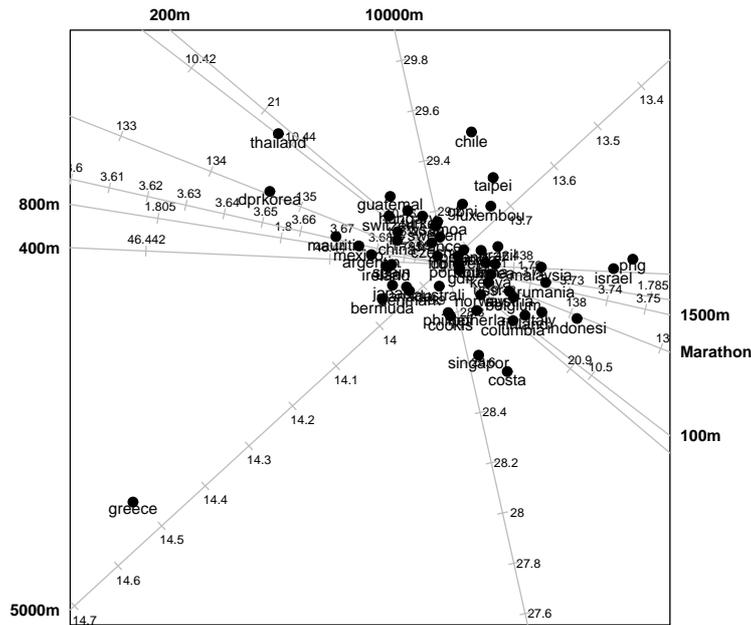


Figure 3.9: The two-dimensional PCA biplot constructed from the last two principal components of the standardised National Track data set.

3.4.2.3 The relationship between sample predictivities and the overall quality

Recall that the overall quality of the PCA biplot can be expressed as a weighted average of the p axis predictivities. It is shown below that the overall quality can also be expressed as the weighted average of the n sample predictivities, the weights being proportional to the respective samples' total sum of squares:

$$\begin{aligned} \psi_i &= \frac{[\widehat{\mathbf{X}}\widehat{\mathbf{X}}']_{ii}}{[\mathbf{X}\mathbf{X}']_{ii}} \\ \longrightarrow \psi_i [\mathbf{X}\mathbf{X}']_{ii} &= [\widehat{\mathbf{X}}\widehat{\mathbf{X}}']_{ii} \\ \longrightarrow \sum_{i=1}^n \psi_i [\mathbf{X}\mathbf{X}']_{ii} &= \sum_{i=1}^n [\widehat{\mathbf{X}}\widehat{\mathbf{X}}']_{ii} \\ \longrightarrow \sum_{i=1}^n \psi_i [\mathbf{X}\mathbf{X}']_{ii} &= \text{tr} \{ \widehat{\mathbf{X}}\widehat{\mathbf{X}}' \} . \end{aligned}$$

The overall quality measure can now be expressed as:

$$\frac{\text{tr}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})}{\text{tr}(\mathbf{X}'\mathbf{X})} = \sum_{i=1}^n \psi_i \frac{[\mathbf{X}\mathbf{X}']_{ii}}{\text{tr}(\mathbf{X}'\mathbf{X})}. \quad (3.4.14)$$

Unlike the expression of the overall quality in terms of the axis predictivities, the expression of the overall quality in terms of the sample predictivities will not simplify to that of an arithmetic average of the individual predictivities when the PCA biplot is constructed from the standardised measurements.

The fact that the overall quality of the PCA biplot is equal to a weighted average of the n sample predictivities implies that a high overall quality does not necessarily suggest that the measurements of all n the samples are accurately approximated in the biplot. Similarly, a low overall quality does not necessarily suggest that the measurements of all n the samples are poorly approximated in the biplot. The average (weighted or arithmetic) of a number of very high sample predictivities together with a few low sample predictivities, can still be very high and similarly, the average of a number of very low sample predictivities together with a few high sample predictivities, can still be very low. The sample predictivities (as well as the axis predictivities) of a PCA biplot with low overall quality should therefore be considered before discarding the biplot - useful information can be gathered regarding the samples (and variables) whose measurements are accurately approximated in the biplot.

Using equation (3.4.14), it can be shown that the increase in the overall quality of the PCA biplot resulting from an increase in its dimensionality can be expressed as the weighted average of the increase in the n sample predictivities resulting from the increase in dimensionality, the weights being proportional to the respective samples' total sum of squares. Consider for instance the increase in the overall quality of the PCA biplot resulting from increasing the dimension of the biplot from r to $r + 1$:

$$\begin{aligned} \alpha_{(r,r+1)} &= \alpha_{r+1} - \alpha_r \\ &= \sum_{i=1}^n \psi_{i,r+1} \frac{[\mathbf{X}\mathbf{X}']_{ii}}{\text{tr}(\mathbf{X}'\mathbf{X})} - \sum_{i=1}^n \psi_{i,r} \frac{[\mathbf{X}\mathbf{X}']_{ii}}{\text{tr}(\mathbf{X}'\mathbf{X})} \\ &= \sum_{i=1}^n (\psi_{i,r+1} - \psi_{i,r}) \frac{[\mathbf{X}\mathbf{X}']_{ii}}{\text{tr}(\mathbf{X}'\mathbf{X})}. \end{aligned}$$

Some of the above mentioned concepts regarding sample predictivities will now be illustrated using the *University* data set. The sample predictivities of three universities, namely the University of Chicago (UChicago), University of California, Berkeley (UCBerkeley) and Purdue University (Purdue) corresponding to each of the possible dimensionalities of the PCA biplot constructed from the standardised measurements of the *University* data set, are provided in Table 3.19. Table 3.20

contains the overall quality of the r -dimensional PCA biplot for each $r \in [1 : 6]$.

Table 3.19: *The sample predictivities of Yale University (Yale) University of Chicago (UChicago), University of California, Berkeley (UCBerkeley) and Purdue University (Purdue) corresponding to the PCA biplot of the University data set constructed from the standardised measurements.*

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
Yale	0.9386	0.9410	0.9524	0.9994	1.0000	1.0000
UChicago	0.0098	0.4422	0.4526	0.7082	0.9509	1
UCBerkeley	0.1744	0.3003	0.8258	0.9750	0.9837	1
Purdue	0.9694	0.9915	0.9968	0.9988	1	1

Table 3.20: *The overall qualities of the PCA biplot of the University data set constructed from the standardised measurements.*

Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
76.87	89.98	94.76	97.49	99.56	100.00

The very small sample predictivity of the University of California, Berkeley, corresponding to the one-dimensional PCA biplot implies that the vector emanating from the origin to the point representing this university in the measurement space lies at a very large angle to its projection onto the one-dimensional PCA biplot space. The almost zero sample predictivity of the University of Chicago corresponding to the one-dimensional PCA biplot implies that the vector emanating from the origin to the point representing this university in the six-dimensional measurement space lies almost orthogonal to the one-dimensional PCA biplot space. The very high sample predictivities of Yale University associated with the one-dimensional PCA biplot implies that the angle between the vector stretching from the origin to the point representing the university in the measurement space and the projection of this vector onto the biplot space is very small, suggesting that the point representing Yale University in the measurement space lies very close to the one-dimensional PCA biplot space. The same can be said about the point representing Purdue University. The overall quality of the one-dimensional PCA biplot is moderately high, accounting for 76.87% of the total sample variance associated with the vector of measured variables. The very low sample predictivities associated with the University of Chicago and the University of California, Berkeley and the very high sample predictivity associated with the Purdue University shows that it is possible that samples in a PCA biplot with moderately high overall quality are very poorly or very accurately approximated. The very small sample predictivity of the University of Chicago corresponding to the three-dimensional PCA biplot, which has an overall quality of 94.76%, confirms that it is possible that a sample in a PCA biplot with very high overall quality is very poorly represented.

Consider the two-dimensional PCA biplot constructed from the standardised

measurements of the *University* data set provided in Figure 3.10. The sample predictivities associated with the University of California, Berkeley and the University of Chicago corresponding to this PCA biplot are low while the sample predictivity associated with Yale University is very high. That is, the University of California, Berkeley and the University of Chicago are poorly represented in the two-dimensional PCA biplot in Figure 3.10 while Yale University is very accurately represented. Upon visual inspection of the biplot the University of California, Berkeley seems to differ substantially from Yale University with respect to the variable *Top10*. In reality however, these two universities have identical measurements on this variable. From the biplot it also seems as if the University of California, Berkeley and the University of Chicago are very similar with respect to the variable *Top10* whereas in reality these two universities differ substantially with respect to this variable. These examples confirm that the position of a sample relative to another in a PCA biplot is meaningless if one or both of these samples are associated with low sample predictivities.

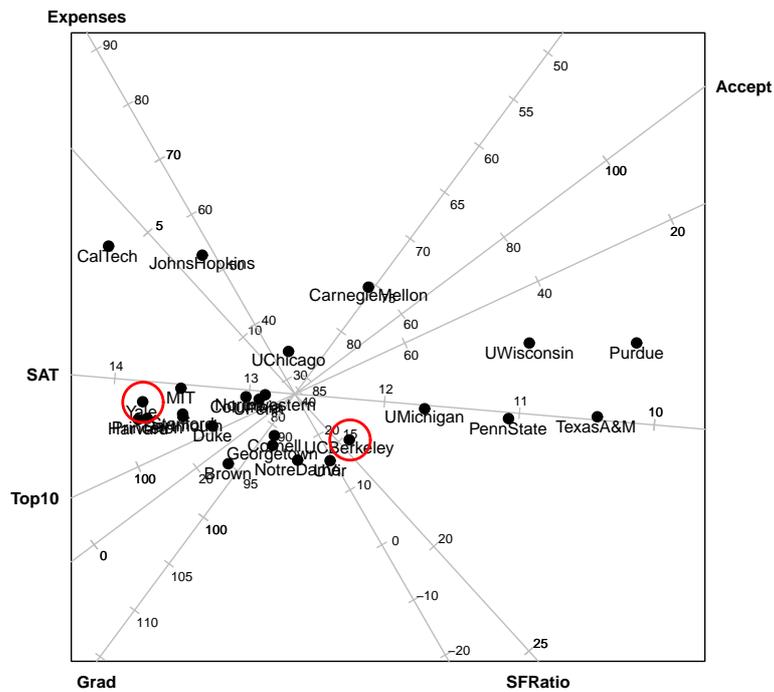


Figure 3.10: *The two-dimensional PCA biplot constructed from the standardised measurements of the University data set.*

3.5 Summary

The conclusions drawn from the PCA biplot are meaningless if the PCA biplot does not accurately represent reality. Measures of the quality of the different aspects of the PCA biplot are therefore required in order to evaluate to what extent the relationships and predictions suggested by a PCA biplot can be trusted to be representative of reality.

The overall quality of the PCA biplot measures the overall accuracy of the approximations of the elements of the matrix \mathbf{X} that are read off the predictive biplot axes. The overall quality is therefore a very crude quality measure which does not provide sufficient information regarding the quality of the representation of the individual samples and variables in the PCA biplot. It is for example possible that some samples (and/or variables) in a PCA biplot with low overall quality, are accurately predicted by the PCA biplot and similarly that some samples (and/or variables) in a PCA biplot with high overall quality, are poorly predicted by the PCA biplot. Since conclusions drawn about samples and variables that are poorly represented in the PCA biplot are likely to be erroneous, measures of the quality of the individual samples and variables are required so that those samples and variables that are poorly represented can be identified.

The sample predictivity of a sample is a measure that quantifies the overall accuracy of the approximations of that sample's measurements that are read off from the predictive biplot axes. The overall quality of the PCA biplot can be expressed as a weighted average of the sample predictivities. The sample predictivities corresponding to the PCA biplot constructed from the last few principal components, which account for a negligible proportion of the variability in the data set, can be used to identify samples that are likely to deviate substantially from the correlation structure of the bulk of the data set. The axis predictivity of a biplot axis quantifies the predictive ability of that individual biplot axis. The overall quality can also be expressed as a weighted average of the axis predictivities. The adequacy of a biplot axis on the other hand measures how much the biplot axis departs from the corresponding Cartesian axis in the measurement space - the less the biplot axis departs from the corresponding Cartesian axis, the higher the adequacy of that biplot axis. The adequacy of a biplot axis is not a trustworthy measure of the predictive ability of that biplot axis but it can in some circumstances provide useful information thereof due to the fact that it is a lower bound for the corresponding axis predictivity.

All four of the quality measures discussed in this chapter are scale dependent. When the measured variables have widely differing standard deviations and the PCA biplot is constructed from the unstandardised measurements, the first few principal components are usually dominated by the variables with standard deviations that are very large compared to those of the other variables. As a result, the biplot axes representing those variables typically have large adequacies and axis predictivities compared to the other biplot axes and the overall quality as a result tends to be overly optimistic.

Each of the quality measures associated with the PCA biplot is defined as a ratio of sums of squared values. The fact that the approximation to \mathbf{X} which is produced by the PCA biplot, $\hat{\mathbf{X}}$, is the orthogonal projection of \mathbf{X} onto the subspace spanned

by the first r right singular vectors of \mathbf{X} , ensures that the decomposition of \mathbf{X} into $\widehat{\mathbf{X}}$ and $\mathbf{X} - \widehat{\mathbf{X}}$ exhibits both Type A and Type B orthogonality. These two orthogonality properties validates all four the quality measures that were discussed in this chapter as quality measures.

Chapter 4 - CVA and the CVA biplot

4.1 Introduction

In Chapters 2 and 3 data sets were graphically represented by means of PCA biplots. The PCA biplot is however not designed to represent the group structure underlying data sets comprising samples partitioned into a number of predefined groups - in fact, the group membership of the individual samples does not play any role in the construction of the PCA biplot. At most the PCA biplot can suggest possible differences between the groups by using different plotting characters and/or colours to represent samples belonging to different groups and imposing an α -bag or convex hull for each of the groups as in the example provided at the end of Chapter 2. When a graphical representation of the group structure underlying a data set is desired, it would be more appropriate to represent the data set by means of a Canonical Variate Analysis (CVA) biplot (Gower and Hand (1996), Gardner-Lubbe *et al.* (2008) and Gower *et al.* (2011)) which is designed specifically for this purpose.

As its name indicates, the CVA biplot is based on the statistical analysis, CVA. CVA is a linear dimension reduction technique which is used to investigate the dissimilarities (and similarities) amongst groups as measured by the Mahalanobis distance metric and is concerned with both discrimination between the groups as well as classification of new observations of unknown origin. CVA, like PCA, is an MDS technique, but differs from PCA in that the distance metric which stands at the centre of the technique is not the Pythagorean distance metric, but rather the Mahalanobis distance metric.

In this chapter, CVA will be discussed from the viewpoint of three different perspectives, namely (1) as equivalent to Linear Discriminant Analysis (LDA) for the multigroup case; (2) as a special case of Canonical Correlation Analysis (CCA) and (3) as a two-stage procedure where the first stage consist of the transformation of the measurement vectors to a space in which the group centroids are optimally separated and the second stage consist of a least squares approximation in that space. Each of these three methods of defining CVA will be discussed in detail in this chapter. The construction of the CVA biplot is however easiest to understand from the point of view of the two-step approach to CVA.

CVA can be either weighted or unweighted depending on whether the different sizes of the groups are taken into account in the analysis or not. Accordingly the CVA biplot can also be either weighted or unweighted. The construction of the weighted and two different types of unweighted CVA biplots will be discussed in this chapter.

An important assumption of CVA is that the within-group covariance matrices of

all the groups are identical. It is very important to check whether this assumption is appropriate for the data to be investigated prior to performing CVA or constructing a CVA biplot. If tests suggest that the assumption of equal covariance matrices is not appropriate for the data at hand, the data should not be analysed by means of CVA. A more appropriate analysis for such data is Analysis of Distance (AOD). The group structure of the data can then be graphically represented by means of an AOD biplot (Gardner *et al.*, 2005). If the assumption of identical within-group covariance matrices is appropriate for the data set at hand, it is important to test whether all the prespecified groups are in fact different prior to performing CVA or constructing a CVA biplot. The appropriate hypotheses to be tested prior to performing CVA or constructing a CVA biplot will be discussed in Section 4.2.5. The close relationship between CVA and Multivariate Analysis of Variance (MANOVA) is also highlighted in this section.

4.2 CVA is equivalent to LDA for the multi-group case

CVA is equivalent to the generalisation of LDA for the two-group case, as proposed by Fisher (1936), to the multi-group case. Fisher (1936) arrived at LDA by searching for the linear combination of the p measured variables that maximally separates two groups. Rao (1948, 1952) generalised Fisher's result so as to make it applicable for the discrimination of several groups, by searching for a set of uncorrelated linear combinations that maximally separates the groups. Note that the more separated groups are, the easier it is to correctly classify an observation of unknown origin. Hence, maximal separation of the groups not only leads to optimal discrimination between the groups but also to optimal classification performance.

The ratio of a linear combination's total unconditional variance to its within-group variance is a measure of the separation of the groups obtained when each observation is represented by the value of the linear combination for that observation - the larger the value of the ratio, the more separated the groups are.

4.2.1 Weighted CVA

4.2.1.1 Discrimination using weighted CVA

Consider a population of observations, each measured on the same p variables and structured into J groups. Let Σ denote the population unconditional (on the groups) covariance matrix associated with the vector variable \mathbf{x} and $\Sigma_{\mathbf{W}}^i$ denote the population within-group covariance matrix of the i th group, $i \in [1 : J]$. An important assumption which is made when performing CVA is that the J groups have identical population within-group covariance matrices, that is

$$\Sigma_{\mathbf{W}}^1 = \Sigma_{\mathbf{W}}^2 = \dots = \Sigma_{\mathbf{W}}^J .$$

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 151

Hence, it must be true that each group's within-group covariance matrix equals the weighted average of the J within-group covariance matrices, that is

$$\Sigma_W = \sum_{i=1}^J p_i \Sigma_W^i$$

where p_i denotes the prior probability of the i th group. Since all the possible observations of $\underline{\mathbf{x}}$ are structured into J mutually exclusive groups, each of which is described by a distribution with a different mean but with the same covariance matrix, the distribution of $\underline{\mathbf{x}}$ is a mixture distribution. Letting f_i denote the probability density function of the i th group, the unconditional (on the group) probability density function associated with $\underline{\mathbf{x}}$ is given by

$$f = \sum_{i=1}^J p_i f_i.$$

The expressions for the unconditional population mean of $\underline{\mathbf{x}}$ and the unconditional covariance matrix associated with $\underline{\mathbf{x}}$ are given by

$$\begin{aligned} \boldsymbol{\mu} &= \sum_{i=1}^J p_i \boldsymbol{\mu}_i \\ \text{and } \boldsymbol{\Sigma} &= \sum_{i=1}^J p_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' + \sum_{i=1}^J p_i \Sigma_W^i \end{aligned} \quad (4.2.1)$$

respectively (Flury, 1997). It is evident from equation (4.2.1) that the unconditional covariance matrix, $\boldsymbol{\Sigma}$, can be partitioned into a between-groups covariance matrix, $\boldsymbol{\Sigma}_B$, and a within-group covariance matrix, $\boldsymbol{\Sigma}_W$, in the following way:

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W \\ \text{where } \boldsymbol{\Sigma}_B &= \sum_{i=1}^J p_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' \\ \text{and } \boldsymbol{\Sigma}_W &= \sum_{i=1}^J p_i \Sigma_W^i. \end{aligned}$$

It follows that the total (unconditional) variance association with a linear combination $\mathbf{m}'\underline{\mathbf{x}}$, that is $\mathbf{m}'\boldsymbol{\Sigma}\mathbf{m}$, can be partitioned into a between-groups variance part,

$\mathbf{m}'\Sigma_B\mathbf{m}$, and a within-group variance part, $\mathbf{m}'\Sigma_W\mathbf{m}$:

$$\mathbf{m}'\Sigma\mathbf{m} = \mathbf{m}'\Sigma_B\mathbf{m} + \mathbf{m}'\Sigma_W\mathbf{m}.$$

The ratio of the total variance to the within-group variance associated with the linear combination $\mathbf{m}'\mathbf{x}$ can therefore be expressed as:

$$\begin{aligned} \frac{\mathbf{m}'\Sigma\mathbf{m}}{\mathbf{m}'\Sigma_W\mathbf{m}} &= \frac{\mathbf{m}'\Sigma_B\mathbf{m} + \mathbf{m}'\Sigma_W\mathbf{m}}{\mathbf{m}'\Sigma_W\mathbf{m}} \\ \rightarrow \frac{\mathbf{m}'\Sigma\mathbf{m}}{\mathbf{m}'\Sigma_W\mathbf{m}} &= \frac{\mathbf{m}'\Sigma_B\mathbf{m}}{\mathbf{m}'\Sigma_W\mathbf{m}} + 1. \end{aligned}$$

It is evident that a vector \mathbf{m} that maximises the ratio of total to within-group variance, $\frac{\mathbf{m}'\Sigma\mathbf{m}}{\mathbf{m}'\Sigma_W\mathbf{m}}$, also maximises the between-to-within-groups variance ratio,

$$\frac{\mathbf{m}'\Sigma_B\mathbf{m}}{\mathbf{m}'\Sigma_W\mathbf{m}}. \quad (4.2.2)$$

Using the between-to-within-groups variance ratio as the measure of the separation between the groups is therefore equivalent to using the ratio of the total variance to the within-group variance as the measure of the separation. The linear combination $\mathbf{x}'\mathbf{m}$ which maximises equation (4.2.2) is called the first linear discriminant function or first canonical variable and is the complete LDA solution for the two-group case.

When dealing with samples, the true population parameters are unknown and need to be estimated from the available samples. A popular choice is to estimate each population parameter of interest by its plug-in estimate (Efron and Tibshirani, 1993). Let n_i denote the number of samples belonging to the i th group, n denote the total number of samples, that is

$$n = \sum_{i=1}^J n_i$$

and let \mathbf{x}_j^i denote the measurement vector of the j th sample belonging to the i th group, $j \in [1 : n_i]$, $i \in [1 : J]$. The plug-in estimate of the population mean of the i th group, $\boldsymbol{\mu}_i$, is given by

$$\bar{\mathbf{x}}^i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_j^i \quad (4.2.3)$$

for $i \in [1 : J]$, while that of the unconditional population mean, $\boldsymbol{\mu}$, is given by

$$\begin{aligned}\bar{\mathbf{x}} &= \sum_{i=1}^J \frac{n_i}{n} \bar{\mathbf{x}}^i \\ \longrightarrow \bar{\mathbf{x}} &= \sum_{i=1}^J \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_j^i \\ \longrightarrow \bar{\mathbf{x}} &= \frac{1}{n} \sum_{i=1}^J \sum_{j=1}^{n_i} \mathbf{x}_j^i.\end{aligned}$$

The plug-in estimate in (4.2.3) will henceforth be referred to as the centroid of the i th group, or i th group centroid for short. The plug-in estimate of the common within-group covariance matrix is given by:

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}} &= \sum_{i=1}^J \frac{n_i}{n} \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}}^i \\ \longrightarrow \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}} &= \sum_{i=1}^J \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \bar{\mathbf{x}}^i) (\mathbf{x}_j^i - \bar{\mathbf{x}}^i)' \\ \longrightarrow \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}} &= \frac{1}{n} \sum_{i=1}^J \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \bar{\mathbf{x}}^i) (\mathbf{x}_j^i - \bar{\mathbf{x}}^i)' .\end{aligned}$$

It is important to test whether the within-group (population) covariance matrices of the J groups are identical prior to performing CVA. When each of the groups has a multivariate normal distribution, equality of the J covariance matrices can be tested using Box's M test (Box, 1949). If the distribution of the data deviates from normality then the test proposed by Tiku and Balakrishnan (1985) can be used.

The total sums of squares and cross-products matrix associated with the measured vector variable, that is

$$\mathbf{T} = \left(\mathbf{X} - \frac{1}{n} \mathbf{1}' \mathbf{X} \right) \left(\mathbf{X} - \frac{1}{n} \mathbf{1}' \mathbf{X} \right)'$$

can be partitioned into a between-groups sums of squares and cross-products matrix, \mathbf{B} , and a within-group sums of squares and cross-products matrix, \mathbf{W} :

$$\mathbf{T} = \sum_{j=1}^J \sum_{i=1}^{n_i} (\mathbf{x}_i^j - \bar{\mathbf{x}}) (\mathbf{x}_i^j - \bar{\mathbf{x}})' = \sum_{j=1}^J n_j (\bar{\mathbf{x}}^j - \bar{\mathbf{x}}) (\bar{\mathbf{x}}^j - \bar{\mathbf{x}})' + \sum_{j=1}^J \sum_{i=1}^{n_i} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j) (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \quad (4.2.4)$$

$$\longrightarrow \mathbf{T} = \mathbf{B} + \mathbf{W}$$

where

$$\mathbf{B} = \sum_{i=1}^J n_i (\bar{\mathbf{x}}^i - \bar{\mathbf{x}}) (\bar{\mathbf{x}}^i - \bar{\mathbf{x}})' \quad (4.2.5)$$

$$\text{and } \mathbf{W} = \sum_{i=1}^J \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \bar{\mathbf{x}}^i) (\mathbf{x}_j^i - \bar{\mathbf{x}}^i)' . \quad (4.2.6)$$

The plug-in estimate of the within-group covariance matrix can therefore be expressed as:

$$\widehat{\Sigma}_{\mathbf{W}} = \frac{1}{n} \mathbf{W} .$$

The plug-in estimate of the between-groups covariance matrix, $\Sigma_{\mathbf{B}}$, is given by

$$\widehat{\Sigma}_{\mathbf{B}} = \sum_{i=1}^J \frac{n_i}{n} (\bar{\mathbf{x}}^i - \bar{\mathbf{x}}) (\bar{\mathbf{x}}^i - \bar{\mathbf{x}})'$$

$$\widehat{\Sigma}_{\mathbf{B}} = \frac{1}{n} \mathbf{B} .$$

The plug-in estimate of the unconditional population covariance matrix, Σ , follows as:

$$\widehat{\Sigma} = \sum_{i=1}^J \frac{n_i}{n} (\bar{\mathbf{x}}^i - \bar{\mathbf{x}}) (\bar{\mathbf{x}}^i - \bar{\mathbf{x}})' + \sum_{i=1}^J \frac{n_i}{n} \widehat{\Sigma}_{\mathbf{W}}^i$$

$$\longrightarrow \widehat{\Sigma} = \widehat{\Sigma}_{\mathbf{B}} + \widehat{\Sigma}_{\mathbf{W}}$$

$$\longrightarrow \widehat{\Sigma} = \frac{1}{n} \mathbf{B} + \frac{1}{n} \mathbf{W}$$

$$\longrightarrow \widehat{\Sigma} = \frac{1}{n} \mathbf{T} .$$

Let \mathbf{X} denote the $n \times p$ matrix of individual observations. It will be assumed throughout the rest of this chapter, as well as Chapter 5, that the $n \times p$ matrix \mathbf{X} is centred such that $\mathbf{1}'\mathbf{X} = \mathbf{0}'$, that $p \leq n - 1$ and that \mathbf{X} is of full column rank. The plug-in estimate of the unconditional population mean, $\boldsymbol{\mu}$, is therefore equal to $\mathbf{0}$ and hence the expressions for the between-groups sums of squares and cross

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 155

products matrix and the plug-in estimates of the between-groups covariance matrix and the unconditional covariance matrix, simplify to

$$\begin{aligned}\mathbf{B} &= \sum_{i=1}^J n_i \bar{\mathbf{x}}^i (\bar{\mathbf{x}}^i)' \\ \widehat{\Sigma}_B &= \sum_{i=1}^J \frac{n_i}{n} \bar{\mathbf{x}}^i (\bar{\mathbf{x}}^i)' \\ \text{and } \widehat{\Sigma} &= \frac{1}{n} \sum_{i=1}^J \sum_{j=1}^{n_i} \mathbf{x}_j^i (\mathbf{x}_j^i)'\end{aligned}$$

respectively.

Let \mathbf{N} denote the $J \times J$ diagonal matrix with i th diagonal element equal to the size of the i th group, n_i , $i \in [1 : J]$, so that

$$[\mathbf{N}]_{ij} = \begin{cases} n_i & \text{if } i = j \\ 0 & \text{otherwise} . \end{cases}$$

Let \mathbf{G} denote the $n \times J$ matrix which indicates the group membership of the n samples in the following way:

$$[\mathbf{G}]_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ belongs to the } j\text{th group} \\ 0 & \text{otherwise} . \end{cases} \quad (4.2.7)$$

The matrix \mathbf{G} will henceforth be referred to as the group indicator matrix. Given the definition of \mathbf{G} , the matrix \mathbf{N} can be expressed as:

$$\mathbf{N} = \mathbf{G}'\mathbf{G} .$$

In the remainder of this chapter, as well as throughout Chapter 5, it will be assumed that the rows of \mathbf{X} are ordered such that the rows corresponding to a particular group are consecutive rows in \mathbf{X} , with the rows corresponding to the first group appearing first and then the rows belonging to the second group and then the rows belonging to the third group etc. Hence, the first n_1 rows of \mathbf{X} will correspond to the first group, while rows $n_1 + 1$ to $n_1 + n_2$ will correspond to the second group etc. Denoting the $n_i \times p$ submatrix of \mathbf{X} consisting of rows that belong to the i th group by \mathbf{X}_i , \mathbf{X}

can be expressed as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_J \end{bmatrix}. \quad (4.2.8)$$

Note that since the matrix \mathbf{G} will account for any ordering of the rows of \mathbf{X} , the assumption of this particular ordering does not imply any loss of generality. The assumption is made merely because it allows for the convenient representation of \mathbf{X} in (4.2.8). Letting $\bar{\mathbf{X}}$ denote the matrix of (sample) group means, then

$$\bar{\mathbf{X}} = \mathbf{N}^{-1} \mathbf{G}' \mathbf{X}.$$

The between-groups sums of squares and cross-products matrix, \mathbf{B} , and the within-group sums of squares and cross-products matrix, \mathbf{W} , can now be expressed as follows:

$$\begin{aligned} \mathbf{B} &= \bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \\ &= \mathbf{X}' \mathbf{G} \mathbf{N}^{-1} \mathbf{N} \mathbf{N}^{-1} \mathbf{G}' \mathbf{X} \\ \rightarrow \mathbf{B} &= \mathbf{X} \mathbf{G} (\mathbf{G}' \mathbf{G})^{-1} \mathbf{G}' \mathbf{X} \\ \mathbf{W} &= \mathbf{X}' \mathbf{X} - \bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \\ \rightarrow \mathbf{W} &= \mathbf{X}' (\mathbf{I} - \mathbf{G} (\mathbf{G}' \mathbf{G})^{-1} \mathbf{G}') \mathbf{X}. \end{aligned}$$

Since \mathbf{X} is centred such that the (unweighted) centroid of its row vectors is given by the origin, the weighted centroid of the row vectors of $\bar{\mathbf{X}}$, the weights being equal to the corresponding group sizes, is given by the origin:

$$\begin{aligned} \mathbf{1}' \mathbf{X} &= \mathbf{0}' \\ \rightarrow \sum_{i=1}^J n_i (\bar{\mathbf{x}}^i)' &= \mathbf{0}' \\ \rightarrow \mathbf{1}' \mathbf{N} \bar{\mathbf{X}} &= \mathbf{0}'. \end{aligned}$$

This implies that in general, the rank of the matrix $\bar{\mathbf{X}}$ is less than or equal to $\min(J-1, p)$. It will however be assumed in this chapter, as well as in Chapter 5, that the rank of $\bar{\mathbf{X}}$ is equal to $\min(J-1, p)$.

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 157

When dealing with samples, it is the ratio of the total to within-group sample variance corresponding to a linear combination which measures the extent to which the groups are separated if each observation is represented by the value of the linear combination for that observation. The linear combination which optimally separates the observations from the different groups with respect to the n samples at hand is therefore defined by the coefficient vector \mathbf{m} which maximises the sample variance ratio,

$$\frac{\mathbf{m}'\widehat{\Sigma}\mathbf{m}}{\mathbf{m}'\widehat{\Sigma}_W\mathbf{m}}. \quad (4.2.9)$$

Since $\widehat{\Sigma}$ and $\widehat{\Sigma}_W$ are proportional to $\mathbf{X}'\mathbf{X}$ and \mathbf{W} respectively, any vector \mathbf{m} that maximises the ratio

$$\frac{\mathbf{m}'\mathbf{X}'\mathbf{X}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} \quad (4.2.10)$$

also maximises the ratio in (4.2.9). Furthermore, given that

$$\mathbf{X}'\mathbf{X} = \mathbf{B} + \mathbf{W}$$

a vector that maximises the ratio

$$\frac{\mathbf{m}'\mathbf{B}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} \quad (4.2.11)$$

also maximises the ratio in (4.2.9):

$$\begin{aligned} \frac{\mathbf{m}'\widehat{\Sigma}\mathbf{m}}{\mathbf{m}'\widehat{\Sigma}_W\mathbf{m}} &= \frac{\mathbf{m}'\mathbf{X}'\mathbf{X}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} \\ \frac{\mathbf{m}'\mathbf{X}'\mathbf{X}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} &= \frac{\mathbf{m}'\mathbf{B}\mathbf{m} + \mathbf{m}'\mathbf{W}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} \\ &= \frac{\mathbf{m}'\mathbf{B}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} + 1 \\ \rightarrow \operatorname{argmax}_{\mathbf{m}} \left\{ \frac{\mathbf{m}'\widehat{\Sigma}\mathbf{m}}{\mathbf{m}'\widehat{\Sigma}_W\mathbf{m}} \right\} &= \operatorname{argmax}_{\mathbf{m}} \left\{ \frac{\mathbf{m}'\mathbf{X}'\mathbf{X}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} \right\} \\ \rightarrow \operatorname{argmax}_{\mathbf{m}} \left\{ \frac{\mathbf{m}'\widehat{\Sigma}\mathbf{m}}{\mathbf{m}'\widehat{\Sigma}_W\mathbf{m}} \right\} &= \operatorname{argmax}_{\mathbf{m}} \left\{ \frac{\mathbf{m}'\mathbf{B}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} \right\}. \end{aligned}$$

It is evident that any one of the ratios in (4.2.9), (4.2.10) and (4.2.11) can be used as a measure of the separation amongst the groups. In the remainder of this chapter as well as in Chapter 5 the ratio in (4.2.11) will be used for this purpose.

It is important to note that the vector \mathbf{m} that maximises the ratio in (4.2.11) is only uniquely defined up to a scalar multiple. Imposing a constraint on \mathbf{m} will confine the search space. A popular choice is the constraint,

$$\mathbf{m}'\mathbf{W}\mathbf{m} = 1. \quad (4.2.12)$$

The vector that maximises the ratio in (4.2.11) and satisfies the constraint in (4.2.12) is uniquely defined only up to multiplication by minus one. The vector \mathbf{m} that maximises the ratio in (4.2.11) while satisfying the constraint in (4.2.12), is a non-zero solution of the following equation:

$$\frac{d}{d\mathbf{m}} \{ \mathbf{m}'\mathbf{B}\mathbf{m} - \lambda (\mathbf{m}'\mathbf{W}\mathbf{m} - 1) \} = \mathbf{0} \quad (4.2.13)$$

$$\longrightarrow 2\mathbf{B}\mathbf{m} - 2\lambda\mathbf{W}\mathbf{m} = \mathbf{0} \quad (4.2.14)$$

$$\longrightarrow \mathbf{B}\mathbf{m} = \lambda\mathbf{W}\mathbf{m}. \quad (4.2.15)$$

Note that since \mathbf{B} is a $p \times p$ symmetric matrix and \mathbf{W} is a $p \times p$ positive definite matrix, equation (4.2.15) is a two-sided eigenvalue problem (Section 1.6.12). Hence, the vector \mathbf{m} which maximises the between-to-within-groups sums of squares ratio in (4.2.11) is an eigenvector of the two-sided eigenvalue problem in (4.2.15). The two-sided eigenvalue problem in (4.2.15) has p eigenvectors and p eigenvalues. Only the case where all the non-zero eigenvalues of the two-sided eigenvalue problem in (4.2.15) are distinct will be considered in this thesis. Let λ_i denote the i th largest eigenvalue of the two-sided eigenvalue problem in (4.2.15) and $\mathbf{m}_{(i)}$ denote the eigenvector of that two-sided eigenvalue problem that corresponds to the eigenvalue λ_i . For convenience, λ_i will henceforth be referred to as the i th eigenvalue of the two-sided eigenvalue problem in (4.2.15) and $\mathbf{m}_{(i)}$ as the i th eigenvector of the two-sided eigenvalue problem, $i \in [1 : p]$. The p eigenvectors and eigenvalues of the two-sided eigenvalue problem in (4.2.15) can be simultaneously represented by

$$\begin{aligned} \mathbf{B}\mathbf{M} &= \mathbf{W}\mathbf{M}\mathbf{\Lambda} \\ \therefore \bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\mathbf{M} &= \mathbf{W}\mathbf{M}\mathbf{\Lambda} \end{aligned} \quad (4.2.16)$$

where $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix with i th diagonal element equal to λ_i and \mathbf{M} is the $p \times p$ matrix with i th column vector equal to $\mathbf{m}_{(i)}$, $i \in [1 : p]$. The i th eigenvector of (4.2.15) maximises the ratio in (4.2.11) under the constraint in (4.2.12) and conditional on being orthogonal to $\mathbf{m}_{(j)}$ in the metric \mathbf{W} , that is $\mathbf{m}'_{(j)}\mathbf{W}\mathbf{m}_{(i)} = 0$,

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 159

for all $j < i$ (Rao, 1952), $i \in [2 : p]$. The fact that the p eigenvectors are orthogonal in the metric \mathbf{W} together with the set of constraints

$$\mathbf{m}'_{(i)} \mathbf{W} \mathbf{m}_{(i)} = 1 \quad \forall i \in [1 : p] \quad (4.2.17)$$

implies that

$$\mathbf{M}' \mathbf{W} \mathbf{M} = \mathbf{I}. \quad (4.2.18)$$

Note that equation (4.2.18) implies that the $p \times p$ matrix \mathbf{M} is of full column rank:

$$\begin{aligned} \mathbf{M}' \mathbf{W} \mathbf{M} &= \mathbf{I} \\ \longrightarrow \text{rank}(\mathbf{M}' \mathbf{W} \mathbf{M}) &= \text{rank}(\mathbf{I}) \\ \longrightarrow \text{rank}(\mathbf{M}' \mathbf{W} \mathbf{M}) &= p \\ \longrightarrow \text{rank}\left(\left(\mathbf{W}^{1/2} \mathbf{M}\right)' \mathbf{W}^{1/2} \mathbf{M}\right) &= p \\ \longrightarrow \text{rank}\left(\mathbf{W}^{1/2} \mathbf{M}\right) &= p \\ \longrightarrow \text{rank}(\mathbf{M}) &= p. \end{aligned}$$

The last step in the above derivation follows from the fact that the matrix $\mathbf{W}^{1/2}$ is non-singular.

It is shown below that λ_i is equal to the value of the between-to-within-groups sums of squares ratio corresponding to the linear combination with coefficient vector $\mathbf{m}_{(i)}$, $i \in [1 : p]$:

$$\begin{aligned} \mathbf{B} \mathbf{M} &= \mathbf{W} \mathbf{M} \mathbf{\Lambda} \\ \longrightarrow \mathbf{M}' \mathbf{B} \mathbf{M} &= \mathbf{M}' \mathbf{W} \mathbf{M} \mathbf{\Lambda} \\ \longrightarrow \frac{\mathbf{m}'_{(i)} \mathbf{B} \mathbf{m}_{(i)}}{\mathbf{m}'_{(i)} \mathbf{W} \mathbf{m}_{(i)}} &= \lambda_i \quad \forall i \in [1 : p]. \end{aligned}$$

This means that λ_i quantifies the extent to which the linear combination $\mathbf{m}'_{(i)} \mathbf{x}$ separates the groups as measured by the between-to-within-groups sums of squares ratio in (4.2.11). It follows that the maximum value of the between-to-within-groups sums of squares ratio in (4.2.11) is equal to the largest eigenvalue of the two-sided eigenvalue problem in equation (4.2.15), that is λ_1 , identifying the linear combination maximising the ratio as that linear combination with coefficient vector $\mathbf{m}_{(1)}$. The linear combination of the measured variables with coefficient vector $\mathbf{m}_{(1)}$ is called

the first sample linear discriminant function or the first sample canonical variable. In general, the linear combination of the measured variables with coefficient vector $\mathbf{m}_{(i)}$ is called the i th sample canonical variable. It will however be explained shortly that the last $p - K$ sample canonical variables are not uniquely defined. Since only sample canonical variables will be considered in the remainder of this chapter as well as throughout Chapter 5, the sample canonical variables will henceforth be referred to simply as canonical variables.

Using the results in Section 1.6.12 it is evident that the eigenvalues of the two-sided eigenvalue problem in (4.2.15) are identical to the eigenvalues of the matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, where $\mathbf{W}^{-1/2}$ is the square root matrix of the positive definite matrix \mathbf{W} . Also, the eigenvectors of the two-sided eigenvalue problem that are orthonormal in the metric \mathbf{W} are given by the column vectors of the matrix

$$\mathbf{M} = \mathbf{W}^{-1/2}\mathbf{F}$$

where \mathbf{F} is the matrix with j th column vector given by the eigenvector of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ that corresponds to the j th largest eigenvalue of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ and has unit length, $j \in [1 : p]$. Since both $\mathbf{W}^{-1/2}$ and \mathbf{F} are $p \times p$ matrices of rank p , the matrix \mathbf{M} is also a $p \times p$ matrix of rank p and hence is non-singular. Since the eigenvalues of the two-sided eigenvalue problem in (4.2.15) are identical to the eigenvalues of the matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, the number of non-zero eigenvalues of the two-sided eigenvalue problem is equal to the rank of the matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$. It is shown below that the rank of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ is less than or equal to K , where

$$K = \min(J - 1, p)$$

and hence that $p - K$ of the eigenvalues of the two-sided eigenvalue problem in (4.2.15) are equal to zero:

$$\begin{aligned} \text{rank}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}) &= \text{rank}(\mathbf{B}) \\ &= \text{rank}(\overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}}) \\ &= \text{rank}(\mathbf{N}^{1/2}\overline{\mathbf{X}}) \\ &= \text{rank}(\overline{\mathbf{X}}) \\ \longrightarrow \text{rank}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}) &\leq K \end{aligned}$$

where $K = \min(J - 1, p)$. It will be assumed throughout the rest of this chapter as well as Chapter 5 that the rank of $\overline{\mathbf{X}}$ is equal to K . It follows that only the first K canonical variables contribute to the separation of the groups. It follows that in the case where the samples of a data set are only partitioned into two groups,

the first canonical variable is the only canonical variable that contributes to the separation of the groups and hence forms the complete CVA (and LDA) solution for the two-group scenario. When however the number of groups is more than two and $p > 1$, more canonical variables contribute to the separation of the groups. When $j > i$ and $i, j \in [1 : K]$, the j th canonical variable separates the groups to a lesser extent than the i th canonical variable. However, since $\mathbf{m}_{(j)}$ is orthogonal to $\mathbf{m}_{(i)}$ in terms of the metric \mathbf{W} , the j th canonical variable separates the groups in a direction different than that in which the i th canonical variable separates the groups. Note that since each of the last $p - K$ eigenvalues of the two-sided eigenvalue problem in (4.2.15) is equal to zero, the last $p - K$ canonical variables are not uniquely defined. Note that in the case where $p < J - 1$, $K = p$ and hence all p eigenvalues of the two-sided eigenvalue problem in (4.2.15) are non-zero which means that all p canonical variables contribute to the separation of the groups.

When $K > 1$, increasing the dimension of the space in which the group structure of the data set is graphically represented by using, in addition to $\mathbf{m}'_{(1)}\mathbf{x}$, linear combinations which separate the groups in directions different to that in which $\mathbf{m}'_{(1)}\mathbf{x}$ separates the groups, will ensure greater separation of the groups and more information regarding the nature of the group structure. The greater the extent to which the groups are separated, the easier it is to discriminate amongst the groups. Also, the more separated the groups are, the easier it is to correctly classify an observation of unknown origin and hence the smaller the misclassification rate of the classifier will be. CVA can provide a solution of any dimension less than or equal to $K = \min(J - 1, p)$. The r -dimensional CVA solution comprises the first r canonical variables, $r \in [1 : K]$. Actually CVA can provide a solution of any dimension less than or equal to p , however the $(K + 1)$ th to p th dimensions of the solution are not uniquely defined since the last $p - K$ eigenvalues of the two-sided eigenvalue problem in (4.2.15) are zero. Only CVA solutions of dimension less than or equal to K will be studied in this thesis. Gardner *et al.* (2006) showed that the trace,

$$\text{tr} \{ \mathbf{F}'_r \mathbf{B} \mathbf{F}_r (\mathbf{F}'_r \mathbf{W} \mathbf{F}_r)^{-1} \} \quad (4.2.19)$$

which is the r -dimensional form of the between-to-within-groups sums of squares ratio,

$$\frac{\mathbf{m}'\mathbf{B}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}}$$

is maximised over all $p \times r$ matrices of full column rank, \mathbf{F}_r , by $\mathbf{F}_r = \mathbf{M}_r$. That is, the r linear combinations defined by the r eigenvectors of the two-sided eigenvalue

problem

$$\mathbf{Bm} = \lambda \mathbf{Wm}$$

that correspond to the r largest eigenvalues, that is the first r canonical variables, are the r linear combinations that maximally separate the groups in r -dimensional space. That is, the groups are maximally separated in the r -dimensional space in which the j th group mean is represented by the point $(\bar{\mathbf{x}}^j)' \mathbf{M}_r$ and the i th sample belonging to the j th group is represented by the point $(\mathbf{x}_i^j)' \mathbf{M}_r$. The maximum value of the trace in equation (4.2.19) is given by

$$\begin{aligned} \text{tr} \{ \mathbf{M}'_r \mathbf{B} \mathbf{M}_r (\mathbf{M}'_r \mathbf{W} \mathbf{M}_r)^{-1} \} &= \text{tr} \{ \mathbf{M}'_r \mathbf{B} \mathbf{M}_r \} \\ &= \text{tr} \{ \mathbf{\Lambda}_r \} \\ &= \sum_{i=1}^r \lambda_i . \end{aligned}$$

Hence, the sum of the r largest eigenvalues of the two-sided eigenvalue problem is a measure of the extent to which the groups are separated by the first r canonical variables. The ratio $\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^K \lambda_i}$ is indicative of how much the first r canonical variables separate the groups compared to the first K canonical variables.

Let y_i denote the i th canonical variable, that is,

$$y_i = \mathbf{x}' \mathbf{m}_{(i)}, i \in [1 : p] .$$

Note that technically, y_i as defined above, is the i th sample canonical variable - the i th population canonical variable is the linear combination with coefficient vector which maximise the between-to-within-groups population variance ratio,

$$\frac{\mathbf{m}' \mathbf{\Sigma}_B \mathbf{m}}{\mathbf{m}' \mathbf{\Sigma}_W \mathbf{m}} .$$

Since it is the sample canonical variables which will be investigated in the rest of this chapter and the next, the sample canonical variables will for convenience simply be referred to as the canonical variables. The observed values of the p canonical variables for a sample with measurement vector \mathbf{x} are given by the elements of the

vector,

$$\mathbf{y}' = \mathbf{x}'\mathbf{M}.$$

It is evident that the canonical variables are obtained by applying a non-singular linear transformation to the vector of measured variables, \mathbf{x} . Consequently, the original measured variables can be expressed as functions of the canonical variables:

$$\mathbf{x}' = \mathbf{y}'\mathbf{M}^{-1}.$$

The space containing the p canonical variables is a p -dimensional space called the canonical space. The canonical space will henceforth be denoted by \mathbb{C}^p . The group means, given by the rows of the $J \times p$ matrix $\overline{\mathbf{X}}$ after transformation to the canonical space, that is the row vectors of the matrix $\overline{\mathbf{X}}\mathbf{M}$, are called the canonical means. Since the transformation from the set of x -variables to the set of canonical variables is a linear one, the canonical means can also be calculated as the group means of the canonical variables and hence are measures of the central localities of the J groups of canonical observations:

$$\begin{aligned} \overline{\mathbf{Y}} &= \overline{\mathbf{X}}\mathbf{M} \\ &= \mathbf{N}^{-1}\mathbf{G}'\mathbf{X}\mathbf{M} \\ \longrightarrow \overline{\mathbf{Y}} &= \mathbf{N}^{-1}\mathbf{G}'\mathbf{Y} \text{ where } \mathbf{Y} = \mathbf{X}\mathbf{M}. \end{aligned}$$

Note that since each row vector of $\overline{\mathbf{X}}\mathbf{M}$ is a linear combination of row vectors of $\mathbf{X}\mathbf{M}$, the row space of $\overline{\mathbf{X}}\mathbf{M}$ is a subspace of the row space of $\mathbf{X}\mathbf{M}$. That is, the J canonical means are perfectly contained in a subspace of the p -dimensional canonical space. It can be shown that the last $p - K$ elements of the J canonical means are identical, that is

$$(\overline{\mathbf{x}}^1)' \mathbf{m}_{(i)} = (\overline{\mathbf{x}}^2)' \mathbf{m}_{(i)} = \dots = (\overline{\mathbf{x}}^J)' \mathbf{m}_{(i)} \quad \forall i \in [K + 1 : p].$$

This means that the J canonical means are perfectly contained in a K -dimensional subspace of the p -dimensional canonical space. This K -dimensional subspace of the p -dimensional canonical space will henceforth be denoted by \mathbb{C}^K . A derivation of this result, which is similar to that of the same result for the population canonical means provided in Gardner (2001), is provided in the appendix at the end of this chapter.

Due to the fact that the J groups have identical population covariance matri-

ces, the population within-group covariance matrix associated with the vector of canonical variables, $\underline{\mathbf{y}}$, is also identical for the J groups. This common population within-group covariance matrix is given by

$$\text{cov}_W(\underline{\mathbf{y}}, \underline{\mathbf{y}}') = \mathbf{M}'\widehat{\Sigma}_W\mathbf{M}.$$

The plug-in estimate of this covariance matrix is given by

$$\begin{aligned} \widehat{\text{cov}}_W(\underline{\mathbf{y}}, \underline{\mathbf{y}}') &= \mathbf{M}'\widehat{\Sigma}_W\mathbf{M} \\ &= \frac{1}{n}\mathbf{M}'\mathbf{W}\mathbf{M} \\ \longrightarrow \widehat{\text{cov}}_W(\underline{\mathbf{y}}, \underline{\mathbf{y}}') &= \frac{1}{n}\mathbf{I}. \end{aligned}$$

It is evident that within each group, the p canonical variables are uncorrelated and have identical sample variances. This means that the n_j points representing the individual canonical observations belonging to the j th group in the canonical space, take on a spherical structure, $j \in [1 : J]$. It follows that the i th canonical variable maximally separates the groups conditional on being uncorrelated with the previous $i - 1$ canonical variables. It is shown below that the unconditional sample covariance matrix associated with the vector of canonical variables is also a diagonal matrix, indicating that the canonical variables are also unconditionally uncorrelated:

$$\begin{aligned} \widehat{\text{cov}}(\underline{\mathbf{y}}, \underline{\mathbf{y}}') &= \widehat{\text{cov}}(\mathbf{M}'\underline{\mathbf{x}}, (\mathbf{M}'\underline{\mathbf{x}})') \\ &= \mathbf{M}'\widehat{\Sigma}\mathbf{M} \\ &= \mathbf{M}'\widehat{\Sigma}_B\mathbf{M} + \mathbf{M}'\widehat{\Sigma}_W\mathbf{M} \\ &= \frac{1}{n}\mathbf{M}'\mathbf{B}\mathbf{M} + \frac{1}{n}\mathbf{M}'\mathbf{W}\mathbf{M} \\ \longrightarrow \widehat{\text{cov}}(\underline{\mathbf{y}}, \underline{\mathbf{y}}') &= \frac{1}{n}\mathbf{\Lambda} + \frac{1}{n}\mathbf{I}. \end{aligned}$$

Recall that the groups are maximally separated in the r -dimensional space in which the j th group mean is represented by the point $(\bar{\mathbf{x}}^j)' \mathbf{M}_r$, that is the point with i th coordinate equal to the value of the i th canonical variable for the j th group mean, $j \in [1 : J]$, $i \in [1 : r]$. The r -dimensional display that has r calibrated orthogonal axes representing the first r canonical variables and contains the J points representing the group centroids, that is $\left\{ (\bar{\mathbf{x}}^j)' \mathbf{M}_r \right\}$, will henceforth be referred to as the r -dimensional CVA display, $r \in [1 : K]$. The $(K + 1)$ th to $(p - 1)$ th CVA displays are not uniquely defined since the last $p - K$ canonical variables are not uniquely defined. The p -dimensional CVA display space is identical to the p -dimensional

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 165

canonical space and contains the J points $\{(\bar{\mathbf{x}}^j)' \mathbf{M}\}$. The use of CVA displays dates back to the work of Rao (1948). The r -dimensional CVA display of a data set allows for visualisation of the group structure underlying the data set. Representation of the group means alone however provides a very limited view of the group structure underlying the data set. Interpolating the individual samples onto the CVA display will allow for visualisation of the within-group dispersion and yield a more informative representation of the group structure. The i th observation belonging to the j th group can be represented in the r -dimensional CVA display by the point $(\mathbf{x}_i^j)' \mathbf{M}_r$, $i \in [1 : n_j]$, $j \in [1 : J]$.

The CVA display will be illustrated at the hand of a simulated data set comprising 400 samples structured into four groups. Group 1 contains 40 samples while each of Group 2, Group 3 and Group 4 contains 120 samples. The observations from the four different groups were drawn (at random) from four multivariate normal distributions with different centroids and identical covariance matrices. The population means and the common within-group correlation matrix is provided in Tables 4.1 and 4.2 respectively.

Table 4.1: *The (population) group means of the four groups of the simulated data set.*

	Var1	Var2	Var3	Var4	Var5
Group 1	4	7	15	4	5
Group 2	7	8	17	6	7
Group 3	10	9	15	5	9
Group 4	9	9	15	9	11

Table 4.2: *The (population) correlation matrix associated with each of the four five-variate normal distributions from which the samples of the simulated data set were drawn.*

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.445	0.160	0.447	0.462
Var2	0.445	1.000	0.579	0.333	0.396
Var3	0.160	0.579	1.000	-0.029	-0.148
Var4	0.447	0.333	-0.029	1.000	0.672
Var5	0.462	0.396	-0.148	0.672	1.000

The two-dimensional CVA display of the simulated data set is provided in Figure 4.1. To aid in the visualisation of the degree of overlap amongst the four groups, a 50% bag has been superimposed onto the CVA display for each of the four groups. The four groups are also represented by different colours - the observations (solid circles) and group centroids (triangles) of Group 1 are black, those of Group 2 are red, those of Group 3 are green while those of Group 4 are blue. Groups 2, 3

and 4 seem to be well separated in the two-dimensional CVA display whereas Group 1 overlaps substantially with Groups 2 and 3. Due to the lack of information on the original measured variables in the CVA display, it cannot be inferred from the CVA display with respect to which variables Groups 2, 3 and 4 differ substantially, causing them to be well separated in the display or with respect to which variable(s) Group 1 is similar to Group 2 and Group 3, causing Group 1 to substantially overlap with these two groups.

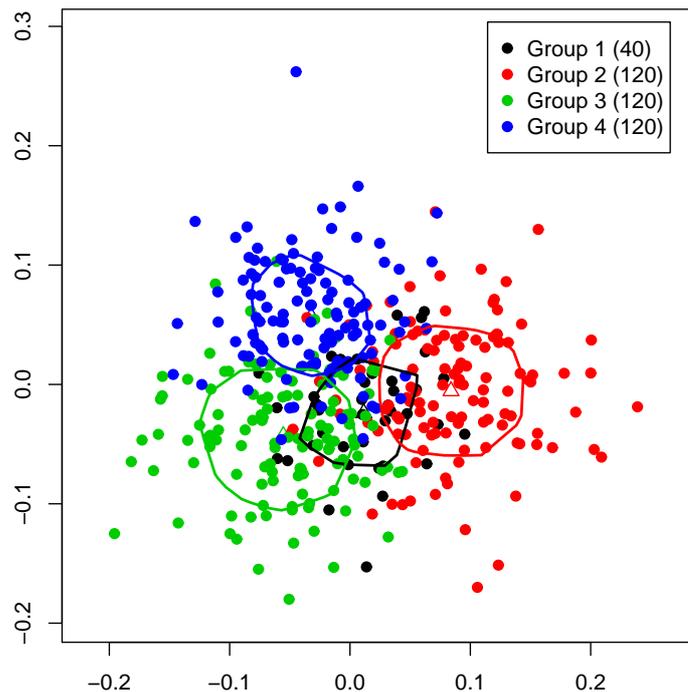


Figure 4.1: *The two-dimensional CVA display of the simulated data set.*

4.2.1.2 Classification using weighted CVA

Up to now, the focus has mainly been on the separation of the groups. CVA is however also designed to classify observations of unknown origin. When the group membership of an observation is unknown, it is desirable to classify that observation to the group the samples of which it is most similar to, or equivalently, least dissimilar to. Given that the centroid of a group is a measure of the group's central locality, it seems natural to use the dissimilarity between an observation and a group's centroid to quantify the dissimilarity between the observation and that group. It therefore makes sense to classify an observation to the group, the centroid of which it is least dissimilar to. Since within each group, the canonical variables are uncorrelated and have identical standard deviations, the Pythagorean distance metric is an appropriate distance metric to quantify the dissimilarity between two canonical observations or between a canonical observation and a canonical mean.

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 167

An observation of unknown origin should therefore be classified to the group, the canonical mean of which lies nearest to the corresponding canonical observation in terms of Pythagorean distance.

Consider the squared Pythagorean distance between a canonical observation in the p -dimensional canonical space, $\mathbf{x}'\mathbf{M}$, and the j th canonical mean, $(\bar{\mathbf{x}}^j)'\mathbf{M}$:

$$\begin{aligned} \|\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^i)'\mathbf{M}\|^2 &= (\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^i)'\mathbf{M})(\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^i)'\mathbf{M})' \\ &= (\mathbf{x} - \bar{\mathbf{x}}^i)'\mathbf{M}\mathbf{M}'(\mathbf{x} - \bar{\mathbf{x}}^i) \\ &= (\mathbf{x} - \bar{\mathbf{x}}^i)'\mathbf{W}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^i) \\ &= \frac{1}{n}(\mathbf{x} - \bar{\mathbf{x}}^i)'\widehat{\Sigma}_{\mathbf{W}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^i) \\ \rightarrow \left\{ (\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^i)'\mathbf{M})(\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^i)'\mathbf{M})' \right\}^{1/2} &= \frac{1}{\sqrt{n}} \left\{ (\mathbf{x} - \bar{\mathbf{x}}^i)'\widehat{\Sigma}_{\mathbf{W}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^i) \right\}^{1/2}. \end{aligned}$$

Note that

$$\left\{ (\mathbf{x} - \bar{\mathbf{x}}^i)'\widehat{\Sigma}_{\mathbf{W}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^i) \right\}^{1/2}$$

is the sample Mahalanobis distance between \mathbf{x} and the i th group mean, $\bar{\mathbf{x}}^i$. It is evident that, as a result of the fact that $\mathbf{M}'\mathbf{W}\mathbf{M} = \mathbf{I}$, or equivalently that

$$\mathbf{M}\mathbf{M}' = \mathbf{W}^{-1},$$

the Pythagorean distance between a canonical observation and canonical mean in the p -dimensional canonical space is proportional to the sample Mahalanobis distance between the corresponding observation and group mean in the p -dimensional measurement space. This implies that if an observation of unknown origin, \mathbf{x} , is such that the corresponding canonical observation, $\mathbf{M}'\mathbf{x}$, is closer to the k th canonical mean, $\bar{\mathbf{y}}^k = \mathbf{M}'\bar{\mathbf{x}}^k$, than to any of the other canonical means in terms of Pythagorean distance, then \mathbf{x} is necessarily closer to the k th group mean, $\bar{\mathbf{x}}^k$, than to any of the other group means in terms of (sample) Mahalanobis distance, that is,

$$\begin{aligned} \min_j \left\{ \left\{ (\mathbf{x} - \bar{\mathbf{x}}^j)'\widehat{\Sigma}_{\mathbf{W}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^j) \right\}^{1/2} \right\} &= \left\{ (\mathbf{x} - \bar{\mathbf{x}}^k)'\widehat{\Sigma}_{\mathbf{W}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^k) \right\}^{1/2} \\ \leftrightarrow \min_j \left\{ \left\{ (\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^j)'\widehat{\Sigma}_{\mathbf{W}}^{-1}(\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^j) \right\}^{1/2} \right\} &= \left\{ (\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^k)'\widehat{\Sigma}_{\mathbf{W}}^{-1}(\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^k) \right\}^{1/2}. \end{aligned}$$

An observation of unknown origin should therefore be classified to the group, the centroid of which it is closest to in terms of (sample) Mahalanobis distance. Hence, the fact that the dissimilarity between a canonical observation and a canonical mean is measured by the Pythagorean distance metric implies that the dissimilarity between an observation and a group mean in the measurement space is measured by the Mahalanobis distance metric. It is evident that in the context of CVA, the distance metric which is used to quantify the dissimilarity between two measurement vectors is the Mahalanobis distance metric. This seems appropriate since the measured variables are usually correlated with different standard deviations and unlike the Pythagorean distance metric, the Mahalanobis distance metric takes the correlations between the measured variables as well as the different standard deviations of the measured variables into account.

A more heuristic reasoning behind classifying an observation of unknown origin to the group with the nearest mean in terms of Mahalanobis distance, is provided in what follows for the case where the observations of the J groups are distributed according to multivariate normal distributions. Consider J groups of individuals, where the observations in the j th group is described by a multivariate normal distribution with covariance matrix Σ_W , and mean $\boldsymbol{\mu}^j$, $j \in [1 : J]$. The likelihood that an observation, \mathbf{x} , is an observation belonging to the j th group is given by the value of the density function of the multivariate normal distribution describing the j th group, in the point, \mathbf{x} , that is,

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_W|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}^j)' \Sigma_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^j)\right). \quad (4.2.20)$$

An observation of unknown origin, \mathbf{x} , should be classified to the group for which the likelihood of the observation, \mathbf{x} , is the greatest. Notice that the (population) Mahalanobis distance between \mathbf{x} and $\boldsymbol{\mu}^j$, that is,

$$(\mathbf{x} - \boldsymbol{\mu}^j)' \Sigma_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^j), \quad (4.2.21)$$

is contained in the density function describing the j th group provided in equation (4.2.20). As a result of this form of the density function of the multivariate normal distribution, the group for which the likelihood of the observation, \mathbf{x} , is greatest, is the group the centroid of which is closest to \mathbf{x} in terms of Mahalanobis distance. This is shown below assuming that the density function of the j th group takes on the greatest value in the point, \mathbf{x} :

$$\begin{aligned} f_j(\mathbf{x}) &> f_i(\mathbf{x}) \forall i \neq j \\ \iff \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}^j)' \Sigma_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^j)\right) &> \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}^i)' \Sigma_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^i)\right) \forall i \neq j \end{aligned}$$

$$\begin{aligned}
 &\longleftrightarrow (\mathbf{x} - \boldsymbol{\mu}^j)' \boldsymbol{\Sigma}_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^j) < (\mathbf{x} - \boldsymbol{\mu}^i)' \boldsymbol{\Sigma}_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^i) \quad \forall i \neq j \\
 &\quad \therefore f_j(\mathbf{x}) = \max_i \{f_i(\mathbf{x})\} \\
 &\longleftrightarrow (\mathbf{x} - \boldsymbol{\mu}^j)' \boldsymbol{\Sigma}_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^j) = \min_i \left\{ (\mathbf{x} - \boldsymbol{\mu}^i)' \boldsymbol{\Sigma}_W^{-1} (\mathbf{x} - \boldsymbol{\mu}^i) \right\}.
 \end{aligned} \tag{4.2.22}$$

It is evident that the classifier that classifies the observation of unknown origin, \mathbf{x} , to the group described by the distribution with the highest density in the point, \mathbf{x} , (or equivalently the group with the highest likelihood in the point, \mathbf{x}) is identical to the classifier that classifies the observation of unknown origin to the group the centroid of which it is nearest to in terms of Mahalanobis distance.

Replacing $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\mu}^j$ by their respective plug-in estimates in equation (4.2.20) and equation (4.2.21) yields the estimated probability density function of the distribution describing group j in the point \mathbf{x} ,

$$\hat{f}_i(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_W|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}}^i)' \hat{\boldsymbol{\Sigma}}_W^{-1} (\mathbf{x} - \bar{\mathbf{x}}^i)\right),$$

and the sample Mahalanobis distance between the point \mathbf{x} and the centroid of the j th group,

$$(\mathbf{x}_i - \mathbf{x}_j)' \hat{\boldsymbol{\Sigma}}_W^{-1} (\mathbf{x}_i - \mathbf{x}_j).$$

By the same argument as before, it is evident that

$$\begin{aligned}
 &\hat{f}_j(\mathbf{x}) = \max_i \{ \hat{f}_i(\mathbf{x}) \} \\
 &\longleftrightarrow (\mathbf{x} - \bar{\mathbf{x}}^j)' \hat{\boldsymbol{\Sigma}}_W^{-1} (\mathbf{x} - \bar{\mathbf{x}}^j) = \min_i \left\{ (\mathbf{x} - \bar{\mathbf{x}}^i)' \hat{\boldsymbol{\Sigma}}_W^{-1} (\mathbf{x} - \bar{\mathbf{x}}^i) \right\} \\
 &\longleftrightarrow (\mathbf{x} - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x} - \bar{\mathbf{x}}^j) = \min_i \left\{ (\mathbf{x} - \bar{\mathbf{x}}^i)' \mathbf{W}^{-1} (\mathbf{x} - \bar{\mathbf{x}}^i) \right\} \\
 &\longleftrightarrow (\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^j)' (\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^j) = \min_i \left\{ (\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^i)' (\mathbf{M}'\mathbf{x} - \mathbf{M}'\bar{\mathbf{x}}^i) \right\}.
 \end{aligned}$$

It is evident that the classifier that classifies an observation of unknown origin \mathbf{x} to the group described by the distribution with the highest estimated density in the point \mathbf{x} is identical to (1) the classifier that classifies \mathbf{x} to the group the centroid of which it is nearest to in terms of sample Mahalanobis distance and (2) the classifier that classifies \mathbf{x} to the group the canonical mean of which is nearest to the corresponding canonical observation, $\mathbf{M}'\mathbf{x}$, in terms of Pythagorean distance.

It seems natural that classifying an observation to the group for which the likelihood of observing that observation is greatest, should yield a small misclassification

rate relative to other classification rules. It can in fact be shown that when the observations from two groups are distributed according to two multivariate normal distributions with identical covariance matrices and differing means, the Mahalanobis distance between the two group means is an increasing function of the probability of misclassification (Rao, 1952). Note that if two populations have identical prior probabilities, then classifying an observation of unknown origin to the group for which the likelihood of observing that observation is greatest, is equivalent to classifying the observation to the group with the greatest posterior probability for that observation. Hence the CVA classifier is identical to the Bayesian classifier in the case where the J groups have identical prior probabilities.

In the rest of this thesis, only analyses and biplots based on a finite number of samples drawn from some population will be discussed and therefore, for convenience, sample Mahalanobis distances will simply be referred to as Mahalanobis distances. The phrase ‘approximate Mahalanobis distance’ will be used to refer to an approximation of the sample Mahalanobis distance in a lower dimensional space. Similarly, sample group means will from this point onwards simply be referred to as group means.

The canonical space can be partitioned into J convex regions (each of which is associated with one of the J groups) separated by hyperplanes, which are such that a canonical observation will lie in a particular region if and only if that canonical observation is closer, in terms of Pythagorean distance, to the canonical mean of the group associated with that region than to any of the other canonical means (Gardner, 2001; Gardner-Lubbe *et al.*, 2008; Gower *et al.*, 2011). The J regions which are constructed in this way are called classification regions as they act as a classification rule - an observation is classified as belonging to the j th group if and only if that observation lies in the classification region associated with the j th group. The classification region consisting of all the canonical observations which are closer, in terms of Pythagorean distance, to the j th canonical mean than to any of the other canonical means, will from now on be referred to as the j th classification region - any observation which lies in this region will be classified as belonging to the j th group. Superimposing the J classification regions onto the CVA display therefore allows for the visualisation of the classification process. Let the j th classification region in the p -dimensional canonical space be denoted by C_j^p . The definition of C_j^p is given by:

$$C_j^p = \{ \mathbf{y} \in \mathbb{C}^p : \|\mathbf{y} - \bar{\mathbf{y}}^j\| < \|\mathbf{y} - \bar{\mathbf{y}}^h\| \forall h \neq j \} \quad j \in [1 : J] . \quad (4.2.24)$$

Given that the Pythagorean distance between two points in the canonical space is proportional to the Mahalanobis distance between the corresponding two points in the p -dimensional measurement space, an observation in \mathbb{C}^p will be classified as belonging to the j th group by the classification regions defined in equation (4.2.24) if and only if the corresponding observation in the measurement space is closer to the j th group mean in terms of Mahalanobis distance than to any of the other $(J - 1)$ group means. The classification regions as defined in equation (4.2.24) are therefore

referred to as exact classification regions.

Consider the boundary between the j th and h th classification regions, that is the set of canonical observations, \mathbf{y} , which are such that

$$\|\mathbf{y} - \bar{\mathbf{y}}^j\| = \|\mathbf{y} - \bar{\mathbf{y}}^h\| .$$

It is shown below that this boundary is in fact a hyperplane:

$$\begin{aligned} & \|\mathbf{y} - \bar{\mathbf{y}}^j\| = \|\mathbf{y} - \bar{\mathbf{y}}^h\| \\ \longrightarrow & (\mathbf{y} - \bar{\mathbf{y}}^j)'(\mathbf{y} - \bar{\mathbf{y}}^j) - (\mathbf{y} - \bar{\mathbf{y}}^h)'(\mathbf{y} - \bar{\mathbf{y}}^h) = 0 \\ \longrightarrow & \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\bar{\mathbf{y}}^j + (\bar{\mathbf{y}}^j)'\bar{\mathbf{y}}^j - \mathbf{y}'\mathbf{y} + 2\mathbf{y}'\bar{\mathbf{y}}^h - (\bar{\mathbf{y}}^h)'\bar{\mathbf{y}}^h = 0 \\ & \longrightarrow \mathbf{y}'(2\bar{\mathbf{y}}^h - 2\bar{\mathbf{y}}^j) = (\bar{\mathbf{y}}^h)'\bar{\mathbf{y}}^h - (\bar{\mathbf{y}}^j)'\bar{\mathbf{y}}^j \\ & \longrightarrow \mathbf{y}'\boldsymbol{\beta} = \mathbf{d} \\ & \text{where } \boldsymbol{\beta} = 2\bar{\mathbf{y}}^h - 2\bar{\mathbf{y}}^j \text{ and } \mathbf{d} = (\bar{\mathbf{y}}^h)'\bar{\mathbf{y}}^h - (\bar{\mathbf{y}}^j)'\bar{\mathbf{y}}^j . \end{aligned}$$

It follows that when the J groups are perfectly linearly separable in the canonical space, the classification regions defined in equation (4.2.24) will perfectly classify the n observations from which $\bar{\mathbf{X}}$ was calculated.

Recall that the J canonical means are perfectly contained in a K -dimensional subspace of the canonical space, \mathbb{C}^K , in which the j th canonical mean in \mathbb{C}^K is represented by the point, $(\bar{\mathbf{z}}^j)' = (\bar{\mathbf{x}}^j)'\mathbf{M}_K$. Just like the p -dimensional canonical space, \mathbb{C}^K can be partitioned into J convex classification regions. Like the j th classification region in \mathbb{C}^p , the j th classification region in \mathbb{C}^K , which will be denoted by C_j^K , is defined as the region consisting of all the points which are such that they are closer to the point representing the j th canonical mean than to any of the points representing the other $(J - 1)$ canonical means in terms of Pythagorean distance, that is,

$$C_j^K = \{\mathbf{z} \in \mathbb{C}^K : \|\mathbf{z} - \bar{\mathbf{z}}^j\| < \|\mathbf{z} - \bar{\mathbf{z}}^h\| \quad \forall h \neq j\} , j \in [1 : J] .$$

Note that since the last $p - K$ components of the J canonical means are identical, it follows that when a canonical observation in the p -dimensional canonical space, $\mathbf{y} = \mathbf{M}'\mathbf{x}$, is closer to the j th canonical mean than to the h th canonical mean, then the point which represents that canonical observation in \mathbb{C}^K , that is $\mathbf{z} = \mathbf{M}'_K\mathbf{x}$, will necessarily be closer to the point representing the j th canonical mean in \mathbb{C}^K than to the point representing the h th canonical mean in \mathbb{C}^K , $j, h \in [1 : J]$. Hence the spread of the points representing the canonical observations in \mathbb{C}^K around the canonical

means is identical to that in \mathbb{C}^p . This is shown below:

$$\begin{aligned}
 & \left(\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^j)' \mathbf{M} \right) \left(\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^j)' \mathbf{M} \right)' < \left(\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^h)' \mathbf{M} \right)' \left(\mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^h)' \mathbf{M} \right)' \\
 & \qquad \qquad \qquad \longrightarrow \sum_{m=1}^p (y_m - \bar{y}_m^j)^2 < \sum_{m=1}^p (y_m - \bar{y}_m^h)^2 \\
 & \longrightarrow \sum_{m=1}^K (y_m - \bar{y}_m^j)^2 + \sum_{m=K+1}^p (y_m - \bar{y}_m^j)^2 < \sum_{m=1}^K (y_m - \bar{y}_m^h)^2 + \sum_{m=K+1}^p (y_m - \bar{y}_m^h)^2 \\
 & \qquad \qquad \qquad \longrightarrow \sum_{m=1}^K (y_m - \bar{y}_m^j)^2 < \sum_{m=1}^K (y_m - \bar{y}_m^h)^2 \\
 & \longrightarrow \left(\mathbf{x}'\mathbf{M}_K - (\bar{\mathbf{x}}^j)' \mathbf{M}_K \right) \left(\mathbf{x}'\mathbf{M}_K - (\bar{\mathbf{x}}^j)' \mathbf{M}_K \right)' < \left(\mathbf{x}'\mathbf{M}_K - (\bar{\mathbf{x}}^h)' \mathbf{M}_K \right)' \left(\mathbf{x}'\mathbf{M}_K - (\bar{\mathbf{x}}^h)' \mathbf{M}_K \right)' .
 \end{aligned}$$

This implies that

$$\left\| \mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^j)' \mathbf{M} \right\| = \min_i \left\{ \left\| \mathbf{x}'\mathbf{M} - (\bar{\mathbf{x}}^i)' \mathbf{M} \right\| \right\} \quad (4.2.25)$$

$$\longrightarrow \left\| \mathbf{x}'\mathbf{M}_K - (\bar{\mathbf{x}}^j)' \mathbf{M}_K \right\| = \min_i \left\{ \left\| \mathbf{x}'\mathbf{M}_K - (\bar{\mathbf{x}}^i)' \mathbf{M}_K \right\| \right\} . \quad (4.2.26)$$

That is, if the point representing an observation \mathbf{x} in \mathbb{C}^p lies in the j th classification region, \mathbb{C}_j^p , then the point representing \mathbf{x} in \mathbb{C}^K will lie in the j th classification region, \mathbb{C}_j^K . Hence, like the classification regions of \mathbb{C}^p , the classification regions of \mathbb{C}^K are exact classification regions. It also implies that if the J groups are perfectly linearly separable in \mathbb{C}^p , they will also be perfectly linearly separable in \mathbb{C}^K . It is evident that the K -dimensional subspace of the canonical space, \mathbb{C}^K , contains all the information contained in the data set which is useful for classification purposes.

The fact that the last $p-K$ elements of the canonical means are identical implies that the Pythagorean distance between two points representing two canonical means in \mathbb{C}^K is equal to the Pythagorean distance between the two points representing the two canonical means in \mathbb{C}^p and hence proportional to the Mahalanobis distance between the corresponding two group means in the p -dimensional measurement space. It follows that the separation between the J canonical means in \mathbb{C}^K is identical to the separation between the J canonical means in \mathbb{C}^p . This together with the fact that the order of the magnitudes of the Pythagorean distances between a canonical observation and the J canonical means in \mathbb{C}^K is identical to that in \mathbb{C}^p , (see equation (4.2.25) and equation (4.2.26)) implies that the separation between the groups in \mathbb{C}^K is identical to that in \mathbb{C}^p and hence \mathbb{C}^K contains all of the information useful for discrimination purposes. This fact is also evident from the result that only the first K canonical variables yield non-zero between-to-within-groups sums of squares ratios.

Note that since the last $p-K$ elements of the individual canonical observations are in general not identical, the Pythagorean distance between two points representing

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 173

two individual observations in \mathbb{C}^K will in general not be equal to the Pythagorean distance between corresponding two points in \mathbb{C}^p . Since the matrix $\mathbf{X}\mathbf{M}$ is centred such that $\mathbf{1}'\mathbf{X}\mathbf{M} = \mathbf{0}'$ and the matrix

$$\mathbf{M}'\mathbf{X}'\mathbf{X}\mathbf{M} = \frac{1}{n}(\mathbf{\Lambda} + \mathbf{I})$$

is a diagonal matrix, the coordinates given by the rows of $\mathbf{X}\mathbf{M}$ are referred to the principal axes of the set of points, $\{\mathbf{x}_i^j\mathbf{M}\}$. Since $[\mathbf{\Lambda}]_{ii} = \lambda_i$ is the i th largest diagonal element of $\mathbf{\Lambda}$,

$$\left[\frac{1}{n}(\mathbf{\Lambda} + \mathbf{I})\right]_{ii} = \frac{1}{n}(\lambda_i + 1)$$

is the i th largest eigenvalue of $\widehat{\text{cov}}(\mathbf{y}, \mathbf{y}')$. This means that the i th coordinate of $(\mathbf{x}_i^j)'\mathbf{M}$ is referred to the i th principal component axis of the set of np points $\{(\mathbf{x}_i^j)'\mathbf{M}\}$, $i \in [1:p]$. It follows that the coordinates of the point representing $(\mathbf{x}_i^j)'\mathbf{M}$ in \mathbb{C}^K are referred to the first K principal component axes of the set of points $\{(\mathbf{x}_i^j)'\mathbf{M}\}$. It follows that the Pythagorean distances between the points representing the individual observations in \mathbb{C}^K optimally approximate the Pythagorean distances between the corresponding points in \mathbb{C}^p .

Since only the first K canonical variables are important for discrimination and classification purposes, transformation of the original measurement vectors and group means to the canonical space results in a significant reduction in the dimensionality of the problem - it 'transforms' the p -dimensional discrimination and classification problem into a K -dimensional problem, where $K = \min(J - 1, p)$. An advantage of this reduction in dimensionality is that if $J - 1$ is equal to or less than three and p is greater than J , then the data can be graphically represented together with the exact classification regions. Another advantage of the reduction in dimensionality is the fact that the reduced space may be more stable than the full space, resulting in better classification performance Hastie *et al.* (1994).

Discrimination and classification can be performed in any r -dimensional space, where $r < K$, however, the J groups will be separated to a lesser extent than in \mathbb{C}^K and the classification regions will not be exact. Hence, the misclassification rate based on the observations from which the matrix $\overline{\mathbf{X}}$ was calculated, will be greater than that associated with the p and K -dimensional CVA display. Recall that out of all possible linear combinations, the first r canonical variables are the r linear combinations of the x -variables that separate the J groups to the greatest extent in r -dimensional space. This implies that out of all possible sets of r linear combinations, using the first r canonical variables to classify the observations from which $\overline{\mathbf{X}}$ was calculated, will ensure the smallest misclassification rate.

Like \mathbb{C}^p and \mathbb{C}^K , \mathbb{C}^r can be partitioned into J convex classification regions sep-

arated by hyperplanes. The j th classification region in \mathbb{C}^r is defined as

$$\mathbb{C}_j^r = \{ \mathbf{z} \in \mathbb{C}^r : \|\mathbf{z} - \bar{\mathbf{z}}^j\| < \|\mathbf{z} - \bar{\mathbf{z}}^h\| \forall h \neq j \}, j \in [1 : J]$$

where $\bar{\mathbf{z}}^i = \mathbf{M}'_r \bar{\mathbf{x}}^i$ is the point representing the i th canonical mean in \mathbb{C}^r , $i \in [1 : J]$. As mentioned before, the classification regions of \mathbb{C}^r are not exact. This means that for an observation, $\mathbf{y} = \mathbf{M}'_r \mathbf{x}$, in \mathbb{C}^r which is such that

$$\|\mathbf{M}'_r \mathbf{x} - \mathbf{M}'_r \bar{\mathbf{x}}^j\| = \min_i \{ \|\mathbf{M}'_r \mathbf{x} - \mathbf{M}'_r \bar{\mathbf{x}}^i\| \}, i \in [1 : J]$$

it is not necessarily true that

$$\|\mathbf{M}'_{K'} \mathbf{x} - \mathbf{M}'_{K'} \bar{\mathbf{x}}^j\| = \min_i \{ \|\mathbf{M}'_{K'} \mathbf{x} - \mathbf{M}'_{K'} \bar{\mathbf{x}}^i\| \}, i \in [1 : J].$$

The reason for this is that in general the $(r + 1)$ th to K th elements of the J canonical means are not identical. This implies that the spread of the individual canonical observations around the canonical means in \mathbb{C}^r is not the same as the spread of the individual canonical observations around the canonical means in \mathbb{C}^K (and \mathbb{C}^p). For the same reason the Pythagorean distances between the pairs of points representing pairs of canonical means in \mathbb{C}^r are not equal to the Pythagorean distances between the pairs of points representing the pairs of canonical means in \mathbb{C}^K and hence not proportional to the Mahalanobis distances between the pairs of group means in the p -dimensional measurement space. This implies that the separation between the groups in \mathbb{C}^r is not the same as the separation between the groups in \mathbb{C}^K or \mathbb{C}^p .

Since the coordinates of $(\mathbf{x}_i^j)' \mathbf{M}_r$ are referred to the first r principal component axes associated with the set of np points $\{(\mathbf{x}_i^j)' \mathbf{M}\}$, the Pythagorean distances between the points representing the individual observations in \mathbb{C}^r optimally approximate the Pythagorean distances between the corresponding points in \mathbb{C}^p . Consequently, the relative magnitudes of the Mahalanobis distances between the points representing the individual observations in the measurement space are optimally approximated by the corresponding Pythagorean distances between the points representing the individual observations in \mathbb{C}^r .

Note that the matrix $\mathbf{M}' \bar{\mathbf{X}}' (\mathbf{I} - \frac{1}{J} \mathbf{1} \mathbf{1}') \bar{\mathbf{X}} \mathbf{M}$ is not a diagonal matrix. Hence, the coordinates of the points $\{(\bar{\mathbf{x}}^j)' \mathbf{M}\}$ are not referred to the principal axes associated with the set of J points $\{(\bar{\mathbf{x}}^j)' \mathbf{M}\}$. Hence, the Pythagorean distances between the points representing the canonical means in \mathbb{C}^r do not optimally approximate the Pythagorean distances between the canonical means in \mathbb{C}^p . Consequently, the relative magnitudes of the Mahalanobis distances between the points representing

the group centroids in the measurement space are optimally approximated by the corresponding Pythagorean distances between the points representing the group centroids in \mathbb{C}^r .

Note that multiplying the matrix \mathbf{N} by an arbitrary constant c has no effect on the matrix \mathbf{M} which determines the positions of the points representing the group means and individual samples in the CVA display and hence also determines the classification regions:

$$\bar{\mathbf{X}}' (c\mathbf{N}) \bar{\mathbf{X}}\mathbf{M} = \mathbf{W}\mathbf{M} (c\mathbf{A}) .$$

The classification regions of the weighted CVA display are therefore determined by the matrix of group means, $\bar{\mathbf{X}}$, and the relative group sizes corresponding to the set of samples from which $\bar{\mathbf{X}}$ is calculated. It follows that in order for the classification regions of the r -dimensional weighted CVA display space to perform similarly as a classification rule for a new set of samples that have been interpolated onto the CVA biplot as it does for the set of samples from which $\bar{\mathbf{X}}$ is calculated, the group means as well as the relative group sizes associated with the new set of samples need to be similar to those associated with the set of samples from which $\bar{\mathbf{X}}$ is calculated, $r \in [1 : p]$. The more the relative group sizes associated with the new set of samples differ from those associated with the set of samples from which $\bar{\mathbf{X}}$ is calculated, the worse the classification regions of the weighted CVA display constructed from $\bar{\mathbf{X}}$ are likely to perform as a classifier for the new set of samples.

4.2.2 Unweighted CVA

4.2.2.1 Discrimination using unweighted CVA

Up to now, the (possibly) different sizes of the J groups have been accounted for in the CVA and the construction of the corresponding CVA displays. If instead of using the ratio

$$\frac{\mathbf{m}'\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}}$$

as maximisation criterion, the ratio

$$\frac{\mathbf{m}'\bar{\mathbf{X}}'\bar{\mathbf{X}}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}} \tag{4.2.27}$$

is used, the (possibly) different sizes of the J groups will not be taken into account in the analysis and CVA displays. Note that since the matrices $\bar{\mathbf{X}}'\bar{\mathbf{X}}$ and \mathbf{W} do not

add up to $\mathbf{X}'\mathbf{X}$, a vector \mathbf{m} that maximises the ratio in (4.2.27) will not necessarily maximise the ratio of total-to-within-groups sums of squares, $\frac{\mathbf{m}'\overline{\Sigma}\mathbf{m}}{\mathbf{m}'\overline{\Sigma}_W\mathbf{m}}$. The ratio in (4.2.27) is the maximisation criterion proposed by Rao (1952) in his generalisation of LDA for the two-group case as proposed by Fisher (1936), to the multi-group scenario. The CVA that corresponds with the maximisation criterion in (4.2.27) is referred to as unweighted CVA.

Substituting the matrix $\overline{\mathbf{X}}'\overline{\mathbf{X}}$ for the matrix \mathbf{B} in equations (4.2.13)-(4.2.15) shows that a vector \mathbf{m} that maximises the ratio in (4.2.27) under the constraint $\mathbf{m}'\mathbf{W}\mathbf{m} = 1$ satisfies the two-sided eigenvalue problem,

$$\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{m} = \lambda\mathbf{W}\mathbf{m}. \quad (4.2.28)$$

For convenience the two-sided eigenvalue problem in (4.2.28) will henceforth be referred to as the unweighted two-sided eigenvalue problem, while

$$\overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}}\mathbf{m} = \lambda\mathbf{W}\mathbf{m}$$

will be referred to as the weighted two-sided eigenvalue problem. Letting the i th largest eigenvalue of the two-sided eigenvalue problem in (4.2.28) be denoted by λ_i and the eigenvector corresponding to λ_i be denoted by $\mathbf{m}_{(i)}$, the p solutions of (4.2.28) can be simultaneously represented by:

$$\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda} \quad (4.2.29)$$

where $\mathbf{\Lambda}$ is the $p \times p$ diagonal matrix with i th diagonal element equal to λ_i and \mathbf{M} is the $p \times p$ matrix with i th column vector equal to $\mathbf{m}_{(i)}$, $i \in [1 : p]$. The eigenvalue, λ_i , and the i th column vector of \mathbf{M} will from now on be referred to as the i th eigenvalue and i th eigenvector of the unweighted two-sided eigenvalue problem respectively. By the same argument as before, it can be shown that $\mathbf{m}_{(1)}$ maximises the ratio in (4.2.27) subject to $\mathbf{m}'\mathbf{W}\mathbf{m} = 1$ and $\mathbf{m}_{(j)}$ maximises the ratio in (4.2.27) subject to $\mathbf{m}'\mathbf{W}\mathbf{m} = 1$ and conditional on being orthogonal to $\mathbf{m}_{(i)}$ in the metric \mathbf{W} for all $i < j$, or equivalently, subject to the linear combination $\mathbf{x}'\mathbf{m}_{(j)}$ being uncorrelated with the previous $j - 1$ linear combinations. It is shown below that the i th eigenvalue of the unweighted two-sided eigenvalue problem, λ_i , is the value of the ratio in (4.2.28) corresponding to the linear combination of the measured variables with coefficient vector $\mathbf{m}_{(i)}$:

$$\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda}$$

$$\begin{aligned} &\longrightarrow \mathbf{M}'\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{M} = \mathbf{M}\mathbf{W}\mathbf{M}\mathbf{A} \\ &\longrightarrow \frac{\mathbf{m}'_{(j)}\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{m}_{(j)}}{\mathbf{m}'_{(j)}\mathbf{W}\mathbf{m}_{(j)}} = \lambda_j \quad \forall j \in [1 : p] . \end{aligned}$$

The linear combinations, $\{\mathbf{x}'\mathbf{m}_{(i)}\}$ are, as before, called the canonical variables, with $\mathbf{x}'\mathbf{m}_{(i)}$ being referred to as the i th canonical variable and denoted by y_i . Since each of the canonical means is a linear combination of the canonical variables, the J canonical means are contained in a subspace of the p -dimensional canonical space. It is shown below that the rank of the matrix $\mathbf{W}^{-1/2}\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{W}^{-1/2}$ is equal to $K = \min(J - 1, p)$ and hence that the unweighted two-sided eigenvalue problem has K non-zero eigenvalues:

$$\begin{aligned} \text{rank}\left(\mathbf{W}^{-1/2}\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{W}^{-1/2}\right) &= \text{rank}\left(\overline{\mathbf{X}}'\overline{\mathbf{X}}\right) \\ &= \text{rank}\left(\overline{\mathbf{X}}\right) \\ \longrightarrow \text{rank}\left(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\right) &= K \end{aligned}$$

where $K = \min(J - 1, p)$. The fact that the last $p - K$ eigenvalues of the unweighted two-sided eigenvalue problem are all equal to zero implies that the last $p - K$ canonical variables are uninformative for discrimination and classification purposes as in the case of weighted CVA.

Upon substituting \mathbf{I} for \mathbf{N} in equation (4.10.1) and 1 for n_i in equations (4.10.2)-(4.10.5) and equation (4.10.6) in the appendix at the end of this chapter, it is evident that as a consequence of the structure of \mathbf{A} in the unweighted two-sided eigenvalue problem, the last $p - K$ elements of the J canonical means are identical, that is

$$\left(\overline{\mathbf{x}}^1\right)' \mathbf{m}_{(j)} = \left(\overline{\mathbf{x}}^2\right)' \mathbf{m}_{(j)} = \cdots = \left(\overline{\mathbf{x}}^J\right)' \mathbf{m}_{(j)}, j \in [K + 1 : p] .$$

This indicates that the J canonical means are perfectly contained in a K -dimensional subspace of the p -dimensional canonical space.

Due to the fact that $\overline{\mathbf{X}}'\overline{\mathbf{X}}$ and \mathbf{W} do not add up to $\mathbf{X}'\mathbf{X}$, the unconditional sample covariance matrix associated with the vector of canonical variables,

$$\text{cov}(\underline{\mathbf{y}}, \underline{\mathbf{y}}') = \frac{1}{n}\mathbf{M}'\mathbf{X}'\mathbf{X}\mathbf{M}$$

is not a diagonal matrix (as is the case when $\mathbf{B} = \overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}}$). This implies that the coordinates given by the rows of $\mathbf{X}\mathbf{M}$ are not referred to the principal component

axes of the set of points $\{(\mathbf{x}_i^j)' \mathbf{M}\}$. Hence, the Pythagorean distances between the points representing the individual observations in the r -dimensional unweighted CVA display do not optimally approximate the Pythagorean distances between the corresponding points in the p -dimensional canonical space. Since the matrix $\bar{\mathbf{X}}$ is not centered and the matrix $\bar{\mathbf{X}}' (\mathbf{I} - \frac{1}{j} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}}$ is not a diagonal matrix, the coordinates given by the rows of $\bar{\mathbf{X}}\mathbf{M}$ are not referred to the principal component axes of the set of points $\{(\bar{\mathbf{x}}^j)' \mathbf{M}\}$ and hence the Pythagorean distances between the points representing the group centroids in the r -dimensional unweighted CVA display do not optimally approximate the Pythagorean distances between the corresponding points in the p -dimensional canonical space.

The unweighted CVA display will now be illustrated at the hand of the simulated data set introduced in Section 4.2.1.1. The two-dimensional unweighted CVA display of the simulated data set is provided in Figure 4.2. In the display a 50% bag is provided for each of the four groups to aid in the visualisation of the group structure.

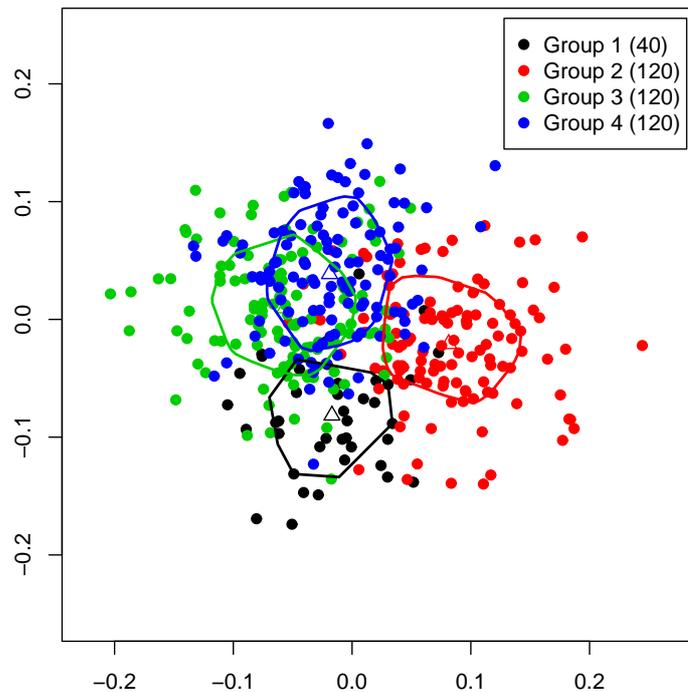


Figure 4.2: *The two-dimensional unweighted CVA display of the simulated data set.*

Unlike in the weighted CVA display in Figure 4.1, Group 1 is well separated from Groups 2 and 3 while groups 3 and 4 overlap substantially. Group 2 is well separated from both Group 3 and Group 4 as in the weighted CVA display. As for the weighted CVA display, it is not possible to infer with respect to which variables the groups are similar or dissimilar due to the lack of information on the original measured variables in the CVA display. The display can only suggest that it is likely that there is at

least one variable with respect to which Group 1 differs substantially from Group 2 and Group 3. The substantial differences between the two-dimensional weighted and unweighted CVA displays of the simulated data is attributable to the large difference between the group size of Group 1 and the sizes of the other three groups. These group sizes were taken into account in the construction of the weighted CVA display in Figure 4.1 but ignored in the construction of the unweighted CVA display in Figure 4.2. The effect of taking the (possibly) different sizes of the groups in a data set into account in the construction of a CVA display is investigated further in Section 4.8.

4.2.2.2 Classification using unweighted CVA

The classification regions of the unweighted CVA display is defined in exactly the same way as those of the weighted CVA display. Denoting the j th classification region of the r -dimensional CVA display as C_r^j , its definition is given by

$$C_r^j = \{ \mathbf{z} \in \mathbb{C}^r : \|\mathbf{z} - \bar{\mathbf{z}}^j\| < \|\mathbf{z} - \bar{\mathbf{z}}^h\| \forall h \neq j \}, j \in [1 : J], r \in [1 : p]$$

where $\mathbf{z} = \mathbf{M}'_r \mathbf{x}$, with $\mathbf{x} \in \mathbb{R}^p$ and $\bar{\mathbf{z}}^j = \mathbf{M}'_r \bar{\mathbf{x}}^j$, $j \in [1 : J]$. Note that since $\mathbf{M}\mathbf{M}' = \mathbf{W}^{-1}$, for both \mathbf{M} satisfying equation (4.2.29) as well as the matrix \mathbf{M} that satisfies equation (4.2.16) the Pythagorean distances between any two points in the p -dimensional weighted CVA display is identical to the Pythagorean distance between the corresponding two points in the p -dimensional unweighted CVA display. When however $r < p$, the inner product matrix $\mathbf{M}'_r \mathbf{M}'_r$ differs for \mathbf{M} satisfying equation (4.2.16) and equation (4.2.29). It follows that in general, the Pythagorean distance between two points in the r -dimensional weighted CVA display will differ from the Pythagorean distance between the corresponding two points in the r -dimensional unweighted CVA display. However since the elements of each of the last $p - K$ columns of the matrix $\bar{\mathbf{X}}\mathbf{M}$ in both equations (4.2.16) and (4.2.29) are identical, the Pythagorean distance between two points representing two group centroids in the K -dimensional weighted and unweighted CVA displays, are identical to the Pythagorean distance between the corresponding two points in the p -dimensional weighted and unweighted CVA displays respectively. This implies that the Pythagorean distance between two points representing two group means in the K -dimensional weighted CVA display is identical to the Pythagorean distance between the corresponding two points in the K -dimensional unweighted CVA display. It also implies that a point representing a sample in the K -dimensional weighted CVA display will lie closer to the point representing the j th group mean than to the point representing the h th group mean if and only if the point representing that sample in the K -dimensional unweighted CVA display lies closer to the point representing the j th group mean than to the point representing the h th group mean. Hence, a point in the K -dimensional weighted CVA display will lie in the j th classification region if and only if the corresponding point in the K -dimensional unweighted CVA display lies in the j th classification region. It follows that when the dimension of the CVA display is equal to K or greater, the relative group sizes corresponding to the set of samples from which the

CVA display was constructed has no effect on the classifications made based on the classification regions.

4.2.3 The connection between weighted and unweighted CVA

When the sizes of the J groups are identical, that is when

$$\mathbf{N} = \frac{n}{J}\mathbf{I},$$

the canonical variables derived from the maximisation of $\frac{\mathbf{m}'\bar{\mathbf{X}}'\bar{\mathbf{X}}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}}$ are identical to the canonical variables derived from the maximisation of $\frac{\mathbf{m}'\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\mathbf{m}}{\mathbf{m}'\mathbf{W}\mathbf{m}}$. The reason for this is that when $\mathbf{N} = \frac{n}{J}\mathbf{I}$, the eigenvector of the two-sided eigenvalue problem,

$$\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\mathbf{m} = \lambda\mathbf{W}\mathbf{m},$$

that is associated with the i th largest eigenvalue of the two-sided eigenvalue problem is identical to the eigenvector of the two-sided eigenvalue problem,

$$\bar{\mathbf{X}}'\bar{\mathbf{X}}\mathbf{m} = \lambda\mathbf{W}\mathbf{m}$$

that is associated with the i th largest eigenvalue of the latter two-sided eigenvalue problem:

$$\begin{aligned}\bar{\mathbf{X}}'\left(\frac{n}{J}\mathbf{I}\right)\bar{\mathbf{X}}\mathbf{M} &= \mathbf{W}\mathbf{M}\mathbf{\Lambda} \\ \bar{\mathbf{X}}'\bar{\mathbf{X}}\mathbf{M} &= \mathbf{W}\mathbf{M}\left(\frac{J}{n}\mathbf{\Lambda}\right).\end{aligned}$$

Hence, when the sizes of the J groups are identical, the r -dimensional weighted CVA display and unweighted CVA display are identical, $r \in [1:p]$. When performing unweighted CVA it is therefore in effect assumed that the sizes of the J groups are identical. The more the group sizes differ, the more the matrix \mathbf{M} satisfying $\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda}$ will differ from the matrix \mathbf{M} satisfying $\bar{\mathbf{X}}'\bar{\mathbf{X}}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda}$ and hence the more the representation of the group structure in the r -dimensional weighted CVA display will differ from that in the r -dimensional unweighted CVA display. When it is known or believed that the relative sizes of the groups in the set

of new samples differ substantially from those in the set of original samples, it will likely be better to represent the new samples in the unweighted CVA display - their relative positions in the display as well as their classifications based on the (approximate) classification regions of the unweighted CVA display are likely to be more representative of reality than if these samples were represented in the weighted CVA display. Recall however that when $r \geq K$, the classifications based on the weighted and unweighted CVA displays are identical and hence the classification performance of the classification regions of both the weighted and unweighted CVA displays are independent of the difference between the relative group sizes corresponding to the set of new samples and that corresponding to the set of original samples.

4.2.4 The scale invariance of CVA

Unlike PCA, CVA performed on the unstandardised measurements of a data set is identical to CVA performed on the measurements standardised such that each column of the standardised data matrix has a unit mean square. This is shown below by highlighting the relationships between the matrices $\bar{\mathbf{X}}$, \mathbf{W} , \mathbf{B} , \mathbf{M} and \mathbf{A} and the corresponding matrices associated with the standardised measurements.

Let \mathbf{X}^* denote the standardised data matrix, obtained by standardising each column of \mathbf{X} to unit mean square. Letting \mathbf{A} denote the $p \times p$ diagonal matrix with j th diagonal element equal to the sample standard deviation of the j th measured variable, the standardised matrix, \mathbf{X}^* , can be expressed as

$$\mathbf{X}^* = \mathbf{X}\mathbf{A}^{-1}.$$

Note that since \mathbf{X} is centred such that $\mathbf{1}'\mathbf{X} = \mathbf{0}'$, the standardised matrix, \mathbf{X}^* , is centred such that

$$\mathbf{1}'\mathbf{X}^* = \mathbf{1}'\mathbf{X}\mathbf{A}^{-1} = \mathbf{0}'\mathbf{A}^{-1} = \mathbf{0}'.$$

Let the matrix of group means, the within-group sums of squares and cross products matrix and the between-groups sums of squares and cross products matrix associated with the standardised measurements, be denoted by $\bar{\mathbf{X}}^*$, \mathbf{W}^* and \mathbf{B}^* respectively. The relationships between the matrices, $\bar{\mathbf{X}}$, \mathbf{W} , $\bar{\mathbf{X}}^*$ and \mathbf{W}^* are provided below:

$$\begin{aligned}\bar{\mathbf{X}}^* &= \mathbf{N}^{-1}\mathbf{G}'\mathbf{X}^* \\ &= \mathbf{N}^{-1}\mathbf{G}'\mathbf{X}\mathbf{A}^{-1} \\ \longrightarrow \bar{\mathbf{X}}^* &= \bar{\mathbf{X}}\mathbf{A}^{-1} \\ \mathbf{W}^* &= (\mathbf{X}^*)'\mathbf{X}^* - (\bar{\mathbf{X}}^*)'\mathbf{N}\bar{\mathbf{X}}^* \\ &= \mathbf{A}^{-1}\mathbf{X}'\mathbf{X}\mathbf{A}^{-1} - \mathbf{A}^{-1}\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\mathbf{A}^{-1}\end{aligned}$$

$$\begin{aligned}
 &= \mathbf{A}^{-1} \left(\mathbf{X}'\mathbf{X} - \bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}} \right) \mathbf{A}^{-1} \\
 \longrightarrow \mathbf{W}^* &= \mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1}.
 \end{aligned} \tag{4.2.30}$$

If the matrix of between-groups sums of squares and cross products is defined in its weighted form, $\mathbf{B}^* = \left(\bar{\mathbf{X}}^*\right)' \mathbf{N}\bar{\mathbf{X}}^*$, while if the matrix of between-groups sums of squares and cross products is defined in its unweighted form, $\mathbf{B}^* = \left(\bar{\mathbf{X}}^*\right)' \bar{\mathbf{X}}^*$. It is shown below that $\mathbf{B}^* = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, irrespective of whether \mathbf{B}^* is defined in its weighted or unweighted form:

$$\begin{aligned}
 \left(\bar{\mathbf{X}}^*\right)' \mathbf{N}\bar{\mathbf{X}}^* &= \mathbf{A}^{-1}\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\mathbf{A}^{-1} \\
 \left(\bar{\mathbf{X}}^*\right)' \bar{\mathbf{X}}^* &= \mathbf{A}^{-1}\bar{\mathbf{X}}'\bar{\mathbf{X}}\mathbf{A}^{-1} \\
 \longrightarrow \mathbf{B}^* &= \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}.
 \end{aligned} \tag{4.2.31}$$

Let \mathbf{M}^* denote the matrix with first column vector equal to the vector \mathbf{m}^* that maximises the ratio,

$$\frac{(\mathbf{m}^*)' \mathbf{B}^* \mathbf{m}^*}{(\mathbf{m}^*)' \mathbf{W}^* \mathbf{m}^*} \tag{4.2.32}$$

under the constraint $(\mathbf{m}^*)' \mathbf{W}^* \mathbf{m}^* = 1$ and j th column vector equal to the vector \mathbf{m}^* that maximises the ratio in (4.2.32) under the constraint $(\mathbf{m}^*)' \mathbf{W}^* \mathbf{m}^* = 1$ and conditional on it being orthogonal to $\mathbf{m}_{(i)}^*$, the i th column vector of \mathbf{M}^* , in the metric \mathbf{W}^* for all $i < j$, that is

$$\left(\mathbf{m}_{(j)}^*\right)' \mathbf{W}^* \mathbf{m}_{(i)}^* = 0 \quad \forall i < j$$

for $j \in [2:p]$. This matrix \mathbf{M}^* satisfies the two-sided eigenvalue problem,

$$\mathbf{B}^* \mathbf{M}^* = \mathbf{W}^* \mathbf{M}^* \mathbf{\Lambda}^* \tag{4.2.33}$$

$$\longrightarrow \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{M}^* = \mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1}\mathbf{M}^* \mathbf{\Lambda}^*. \tag{4.2.34}$$

Now, since it is known that

$$\mathbf{B}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda} \quad (4.2.35)$$

and hence that

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{A}\mathbf{M} = \mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1}\mathbf{A}\mathbf{M}\mathbf{\Lambda}, \quad (4.2.36)$$

it follows that

$$\mathbf{B}^*\mathbf{M}^* = \mathbf{W}^*\mathbf{M}^*\mathbf{\Lambda}^* \quad (4.2.37)$$

where \mathbf{B}^* and \mathbf{W}^* are defined in (4.2.31) and (4.2.30) respectively and

$$\begin{aligned} \mathbf{M}^* &= \mathbf{A}\mathbf{M} \\ \text{and } \mathbf{\Lambda}^* &= \mathbf{\Lambda}. \end{aligned}$$

Consequently the matrix of canonical means calculated from the standardised measurements, $\bar{\mathbf{Y}}^*$, and the matrix of canonical observations calculated from the standardised measurements, \mathbf{Y}^* , are identical to the corresponding matrices calculated from the unstandardised measurements, \mathbf{X} :

$$\begin{aligned} \bar{\mathbf{Y}}^* &= \bar{\mathbf{X}}^*\mathbf{M}^* \\ &= \bar{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{M} \\ &= \bar{\mathbf{X}}\mathbf{M} \\ \longrightarrow \bar{\mathbf{Y}}^* &= \bar{\mathbf{Y}} \\ \mathbf{Y}^* &= \mathbf{X}^*\mathbf{M}^* \\ &= \mathbf{X}\mathbf{A}^{-1}\mathbf{A}\mathbf{M} \\ &= \mathbf{X}\mathbf{M} \\ \longrightarrow \mathbf{Y}^* &= \mathbf{Y}. \end{aligned}$$

The coordinate vector of the point representing the j th group mean calculated from the standardised measurements, that is $(\bar{\mathbf{x}}^j)'\mathbf{A}^{-1}$, in the r -dimensional CVA display

is given by

$$(\bar{\mathbf{x}}^j)' \mathbf{A}^{-1} \mathbf{M}_r^* = (\bar{\mathbf{x}}^j)' \mathbf{A}^{-1} \mathbf{A} \mathbf{M}_r = (\bar{\mathbf{x}}^j)' \mathbf{M}_r .$$

Note that $(\bar{\mathbf{x}}^j)' \mathbf{M}_r$ is the point representing the observed j th group mean, $\bar{\mathbf{x}}^j$, in the r -dimensional CVA display constructed from the unstandardised measurements. Similarly, the coordinate vector of the point representing the i th standardised sample of the j th group, $(\mathbf{x}_i^j)' \mathbf{A}^{-1}$, in the r -dimensional CVA display constructed from the standardised measurements, that is $(\mathbf{x}_i^j)' \mathbf{A}^{-1} \mathbf{M}_r^*$, is identical to that representing $(\mathbf{x}_i^j)'$ in the r -dimensional CVA display constructed from the unstandardised measurements:

$$(\mathbf{x}_i^j)' \mathbf{A}^{-1} \mathbf{M}_r^* = (\mathbf{x}_i^j)' \mathbf{A}^{-1} \mathbf{A} \mathbf{M}_r = (\mathbf{x}_i^j)' \mathbf{M}_r .$$

CVA is not only invariant to transformations of \mathbf{X} by diagonal transformation matrices like \mathbf{A}^{-1} but is invariant to all non-singular linear transformations of the form

$$\mathbf{x} \longrightarrow \mathbf{F}' \mathbf{x} \tag{4.2.38}$$

where \mathbf{F} is any $p \times p$ non-singular matrix. This is evident upon substituting an arbitrary $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in all of the expressions above. It is also evident from the fact that the distance metric associated with CVA, namely the Mahalanobis distance metric, is invariant to changes in the scales of the measurements of the form given in equation (4.2.38). This is indicated below by showing that the squared Mahalanobis distance between two points $\mathbf{F}' \mathbf{x}_i^h$ and $\mathbf{F}' \mathbf{x}_k^j$ is identical to the squared Mahalanobis distance between the two points \mathbf{x}_i^h and \mathbf{x}_k^j :

$$\begin{aligned} & \left((\mathbf{x}_i^h)' \mathbf{F} - (\mathbf{x}_k^j)' \mathbf{F} \right) \left(\frac{1}{n} \mathbf{F}' \mathbf{W} \mathbf{F} \right)^{-1} (\mathbf{F}' \mathbf{x}_i^h - \mathbf{F}' \mathbf{x}_k^j) \\ &= n (\mathbf{x}_i^h - \mathbf{x}_k^j)' \mathbf{F} \mathbf{F}^{-1} \mathbf{W}^{-1} (\mathbf{F}')^{-1} \mathbf{F}' (\mathbf{x}_i^h - \mathbf{x}_k^j) \\ &= (\mathbf{x}_i^h - \mathbf{x}_k^j)' \Sigma_{\mathbf{W}}^{-1} (\mathbf{x}_i^h - \mathbf{x}_k^j) . \end{aligned}$$

4.2.5 Important hypotheses to test prior to performing CVA

Prior to performing CVA or constructing a CVA biplot, it is very important to test whether the assumption of identical within-group covariance matrices is appropriate

for the data set at hand. If the observations from the different groups are normally distributed, the null hypothesis of equality of covariance matrices can be tested using Box's M-test (Box, 1949), which is the multivariate analogue of Bartlett's univariate test for the homogeneity of variance across groups (Bartlett, 1938). Unfortunately, Box's M-test is very sensitive to departures from normality (Box, 1953; Johnson and Wichern, 2002). A method of testing the equality of covariance matrices which is less sensitive to departure from normality was proposed by Tiku and Balakrishnan (1985). If the hypothesis test suggest that the assumption of equal covariance matrices is not appropriate, the data should not be analysed by means of CVA and hence should not be graphically represented by means of a CVA biplot. In such a case it may be more appropriate to analyse the data using Analysis of Distance (AOD) and graphically representing the data set by means of an AOD biplot (Gardner *et al.*, 2005).

If the assumption of identical within-group covariance matrices is appropriate for the data set at hand, it is important to check whether in fact the J groups are truly different or whether two or more of the prespecified groups are not in fact part of the same group. The first hypothesis that needs to be tested in an attempt to determine this is the null hypothesis which states that all J the population group means are identical against the alternative hypothesis which states that at least one of the group means differs from the other group means. This hypothesis test can be performed in exactly the same way in which it is performed in the context of a one-way Multivariate Analysis of Variance (MANOVA). CVA and MANOVA have many aspects in common. Both analyses are used to analyze data sets consisting of individual observations (all of which are measured on the same variables) which are structured into a number of prespecified groups, all of which have identical within-group covariance matrices, and both focus on the magnitude of the between-groups variability relative to the magnitude of the within-group variability. In the context of a MANOVA, between-groups variability and within-group variability are usually referred to as between-treatment variability and within-treatment variability respectively. It is the aims of the two analyses which differ - while CVA is aimed at discriminating amongst the groups and classifying observations of unknown origin, MANOVA is aimed at performing hypothesis tests regarding the group means, in particular the null hypothesis specifying that the J group means are identical against the alternative hypothesis which states that at least one of the group means differs from the other group means. One other regard in which MANOVA differs from CVA is the assumptions required to perform the analysis. In order to perform inference regarding the group means, the observations must be independently and normally distributed - note that these are not requirements for the performance of CVA.

Consider the MANOVA model:

$$\mathbf{x}_i^j = \boldsymbol{\mu} + \boldsymbol{\mu}^j + \boldsymbol{\epsilon}_i^j \quad (4.2.39)$$

where x_i^j is the random variable representing the i th observation of the j th group, $\boldsymbol{\mu}$ denotes the overall population mean, $\boldsymbol{\mu}^j$ denotes the population mean of the j th

group and ξ_i^j denotes the random error on the i th observation of the j th group, $i \in [1 : n_j]$, $j \in [1 : J]$. The MANOVA model assumes that each of the stochastic vector variables, ξ_i^j , follows a multivariate normal distribution with mean $\mathbf{0}$ and a common covariance matrix, $i \in [1 : n_j]$, $j \in [1 : J]$. Let the common covariance matrix of the ξ_i^j be denoted by Σ_W . The observations of the stochastic variables, $\{\mathbf{x}_i^j\}$, can be partitioned in a way analogous to the partitioning of \mathbf{x}_i^j in equation (4.2.39). This partitioning of \mathbf{x}_i^j is shown below:

$$\mathbf{x}_i^j = \bar{\mathbf{x}} + (\bar{\mathbf{x}}^j - \bar{\mathbf{x}}) + (\mathbf{x}_i^j - \bar{\mathbf{x}}^j) .$$

The null hypothesis of no difference between the J group means (or equivalently, the null hypothesis of no treatment effect, as it is usually referred to in the context of a MANOVA), that is

$$H_0 : \boldsymbol{\mu}^1 = \boldsymbol{\mu}^2 = \dots = \boldsymbol{\mu}^J , \quad (4.2.40)$$

can be tested against the alternative hypothesis which states that at least one of the J group means differs from the rest, by comparing the magnitude of the between group variability to the magnitude of the within-group variability. The variability within a particular group reflects the variability which can result from chance alone. MANOVA investigates whether the variability of the group means is greater than would be expected as the result of chance alone. If the magnitude of the between-groups variability is very similar to the magnitude of the within-group variability, then the group means are not significantly different and the null hypothesis in equation (4.2.40) will not be rejected. In order to conclude that the J group means are in fact significantly different, that is in order to reject the null hypothesis in equation (4.2.40), the magnitude of the between-groups variability relative to the magnitude of the within-group variability must be large, or equivalently, the magnitude of the total variability relative to the magnitude of the within-group variability must be large or when the magnitude of the within-group variability relative to the total variability is small.

Recall from Section 4.2 that the matrix of sums of squares and cross products corrected for the mean,

$$\mathbf{T} = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}) (\mathbf{x}_i^j - \bar{\mathbf{x}})' ,$$

can be partitioned into the matrix of between-groups sums of squares and cross products, $\mathbf{B} = \bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}}$, and the matrix of within-group sums of squares and cross products, \mathbf{W} , with \mathbf{B} and \mathbf{W} as defined in equation (4.2.5) and equation (4.2.6)

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 187

respectively (see equation (4.2.4)). There are four statistics which can be used to test the null hypothesis in equation (4.2.40), namely Wilk's lambda statistic, the Lawley-Hotelling trace statistic, Pillai's trace statistic and Roy's largest root statistic. All four of these statistics can be expressed as functions of the eigenvalues of the matrix, $\mathbf{W}^{-1}\mathbf{B}$. The four mentioned statistics are approximately equivalent if the total sample size, $n = \sum_{j=1}^J n_j$, is large (Johnson and Wichern, 2002). The statistic which will be focused on here, is Wilk's lambda. One of the reasons for focusing on this particular statistic is that it is the multivariate analogue of the statistic used to test the null hypothesis of no treatment effect in the univariate setting (that is, in the setting of an ANOVA). Furthermore, Wilk's lambda is proportional to the likelihood ratio statistic associated with testing the null hypothesis of no differences between the group means (Johnson and Wichern, 2002). In the case where there are only two prespecified groups, Wilk's lambda is also proportional to the inverse of Hotelling's T^2 -statistic, which can be used to test the null hypothesis of no treatment effect in the case of only two groups. Wilk's lambda is defined to be the following ratio:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|} = \frac{|\mathbf{W}|}{|\mathbf{T}|}.$$

Wilk's lambda can also be expressed as the following function of the eigenvalues of the matrix, $\mathbf{W}^{-1}\mathbf{B}$:

$$\sum_{i=1}^K \frac{1}{1 + \tau_i}$$

where $K = \min(J - 1, p)$ is the rank of \mathbf{B} . If the sample size, n , is large, then a constant multiple of the natural log transform of Wilk's lambda follows approximately a Chi-squared distribution with $p(J - 1)$ degrees of freedom:

$$-\left(n - 1 - \frac{p + J}{2}\right) \ln \Lambda \sim \chi_{p(J-1)}^2.$$

The null hypothesis of no difference between the J group means will therefore be rejected at a significance level of $(\alpha \times 100)\%$ when

$$-\left(n - 1 - \frac{p + J}{2}\right) \ln \Lambda > \chi_{p(J-1)}^2(\alpha)$$

where $X_{p(J-1)}^2(\alpha)$ denotes the upper $(\alpha \times 100)$ th percentile of the Chi-squared distribution with $p(J-1)$ degrees of freedom. Since the natural log is a monotonically increasing function of its argument, it follows that, the smaller the value of Wilk's lambda, the greater the extent to which the J group means differ from each other and the greater the likelihood that the null hypothesis of no difference between the group means will be rejected. The magnitude of Wilk's lambda for the observed J group means can therefore be used, not only to determine whether the null hypothesis of no difference in group means should be rejected or not at a particular significance level, but also to provide a measure of the extent to which the J group means differ. When the sample size, n , is not large, then, for certain combinations of J and p , Wilk's lambda follows an F-distribution (see Johnson and Wichern (2002), p. 303).

If the null hypothesis of no differences between the group means is rejected, then the difference between the means of each pair of groups can be investigated. This can be done by, for each pair of groups, testing the null hypothesis of no difference between the two group means against the alternative hypothesis of different group means. Hotelling's T^2 -statistic can be used to test hypotheses regarding differences between two population means. Specifically, Hotelling's T^2 -statistic can be used to test the null hypothesis of no difference between two groups' population means. In the discussion which follows, it will be shown that Hotelling's T^2 -statistic is proportional to an approximation of the squared Mahalanobis distance between the two group means under consideration and hence also proportional to the Pythagorean distance between the corresponding two canonical means.

Consider two groups, one consisting of n_1 observations from a p -variate population with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$ and the other consisting of n_2 observations from a p -variate population with mean $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}$. Let n denote the total number of observations, that is $n = n_1 + n_2$. Hotelling's T^2 statistic used to test the null hypothesis,

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$$

against the alternative hypothesis

$$H_a : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0$$

is defined as

$$T^2 = [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2 - \boldsymbol{\delta}_0]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \widehat{\boldsymbol{\Sigma}}_{\text{pooled}} \right]^{-1} [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2 - \boldsymbol{\delta}_0] \quad (4.2.41)$$

where

$$\begin{aligned}\widehat{\Sigma}_{\text{pooled}} &= \frac{1}{n-2} \sum_{j=1}^2 \sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j) (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \\ \longrightarrow \widehat{\Sigma}_{\text{pooled}} &= \frac{1}{n-2} \mathbf{W}\end{aligned}$$

where \mathbf{W} denotes the matrix of within-group sums of squares and cross products as before. Consider the situation where the null hypothesis

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0} \quad (4.2.42)$$

has to be tested against the alternative hypothesis

$$H_a : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0}.$$

Hotelling's T^2 -statistic under the null hypothesis in equation (4.2.42) is obtained by setting $\boldsymbol{\delta}_0$ equal to $\mathbf{0}$ in equation (4.2.41). Let T_0^2 denote Hotelling's T^2 -statistic under the null hypothesis given in (4.2.42):

$$T_0^2 = [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \widehat{\Sigma}_{\text{pooled}} \right]^{-1} [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]. \quad (4.2.43)$$

When the observations from the two groups are distributed according to a multivariate normal distribution, the distribution of Hotelling's T^2 -statistic under the null hypothesis given in equation (4.2.42) is proportional to an F-distribution with p (numerator) and $n-p-1$ (denominator) degrees of freedom (Johnson and Wichern, 2002). Specifically, the distribution of T_0^2 is given by

$$T_0^2 \sim \frac{p(n-2)}{n-p-1} F_{(p, n-p-1)}.$$

The null hypothesis in equation (4.2.42) can be tested at an $(\alpha \times 100)\%$ significance

level by comparing the observed value of T_0^2 to

$$\frac{p(n-2)}{n-p-1} F_{(p,n-p-1)}(\alpha)$$

where $F_{(p,n-p-1)}(\alpha)$ denotes the upper $(\alpha \times 100)$ th percentile of the F distribution with p and $n-p-1$ degrees of freedom. The greater the observed value of T_0^2 , the stronger the evidence against the null hypothesis given in equation (4.2.42). If the observed value of T_0^2 is larger than $\frac{p(n-2)}{n-p-1} F_{(p,n-p-1)}(\alpha)$, then the null hypothesis in equation (4.2.42) is rejected at an $(\alpha \times 100)\%$ significance level. It can then be concluded that the means, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, are different and hence that the two populations are different. When the observed value of T^2 is smaller than $\frac{p(n-2)}{n-p-1} F_{(p,n-p-1)}(\alpha)$, there is not sufficient evidence to reject the null hypothesis in equation (4.2.42). It is evident that T_0^2 is a measure of the dissimilarity (or separation) between the two populations.

When the two populations are not multivariate normal and n_1 and n_2 are large, Hotelling's T^2 -statistic given in equation (4.2.41), has a distribution which is approximately Chi-squared with p degrees of freedom. This follows from the Central Limit Theorem, according to which the sample mean of a large number of observations is approximately normally distributed with mean equal to the population mean and covariance matrix equal to the population covariance matrix. Specifically, the approximate distributions of $\bar{\mathbf{x}}^1$ and $\bar{\mathbf{x}}^2$, as defined above, are given by

$$\begin{aligned} \bar{\mathbf{x}}^1 &\sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ \text{and } \bar{\mathbf{x}}^2 &\sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) . \end{aligned}$$

Consequently, the approximate distribution of $\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2$ when n_1 and n_2 are large, is given by

$$\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2 \sim N\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{\Sigma}\right) .$$

Under the null hypothesis given in equation (4.2.42), the large sample distribution of $\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2$ is given by

$$\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2 \sim N\left(\mathbf{0}, \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{\Sigma}\right) .$$

It follows that the distribution of

$$(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$$

under the null hypothesis,

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$$

tends to a Chi-squared distribution with p degrees of freedom as $(n_1 + n_2)$ becomes large. When n is large, the difference between $\widehat{\boldsymbol{\Sigma}}_{\text{pooled}}$ and $\widehat{\boldsymbol{\Sigma}}$ be very small with high probability and consequently, the large sample distribution of Hotelling's T^2 -statistic is approximately equal to the Chi-squared distribution with p degrees of freedom. It follows that if the observed value of Hotelling's T^2 -statistic is greater than $X_p^2(\alpha)$, the upper $(\alpha \times 100)$ th percentile of the Chi-squared distribution with p degrees of freedom, then there is enough evidence to reject the null hypothesis given in equation (4.2.42) and hence conclude that the two groups' population means are not equal. Note that when the sample sizes, n_1 and n_2 are small, the assumption that the two populations are multivariate normal must be appropriate in order to make inferences about $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

It can be illustrated that the (sample) Mahalanobis distance between the two group means, $\bar{\mathbf{x}}^1$ and $\bar{\mathbf{x}}^2$, can be used as a measure of the dissimilarity between the two groups, just like T_0^2 , Hotelling's T^2 -statistic under the null hypothesis,

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}.$$

In order to see why this is true, equation (4.2.43) needs to be rewritten in the following way:

$$\begin{aligned} T^2 &= [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \widehat{\boldsymbol{\Sigma}}_{\text{pooled}} \right]^{-1} [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2] \\ &\rightarrow \left(\frac{1}{n_1} + \frac{1}{n_2} \right) T^2 = [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]' \left[\frac{1}{n-2} \mathbf{W} \right]^{-1} [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2] \\ &\rightarrow \frac{n}{n-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) T^2 = [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]' \left[\frac{1}{n} \mathbf{W} \right]^{-1} [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]. \end{aligned}$$

It is evident that the squared Mahalanobis distance between $\bar{\mathbf{x}}^1$ and $\bar{\mathbf{x}}^2$ is proportional to the observed value of T_0^2 and that, given that n is greater than 2, the Mahalanobis distance is a monotonically increasing function of T_0^2 . This means that

the greater the Mahalanobis distance between the two population means, $\bar{\mathbf{x}}^1$ and $\bar{\mathbf{x}}^2$, the stronger the evidence against the null hypothesis in equation (4.2.42). This implies that the Mahalanobis distance between $\bar{\mathbf{x}}^1$ and $\bar{\mathbf{x}}^2$ can be used as a measure of the dissimilarity between the two groups. Remembering that the Pythagorean distance between the two canonical means, $(\bar{\mathbf{x}}^1)' \mathbf{M}$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}$, is proportional to the Mahalanobis distance between the two group means, $\bar{\mathbf{x}}^1$ and $\bar{\mathbf{x}}^2$, the constant of proportionality being a positive constant, it is evident that the greater the Pythagorean distance between the two canonical means, $(\bar{\mathbf{x}}^1)' \mathbf{M}$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}$, the stronger the evidence against the null hypothesis, $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$. This means that the Pythagorean distance between the two canonical means, $(\bar{\mathbf{x}}^1)' \mathbf{M}$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}$, can be used as a measure of the dissimilarity between the two groups. Also, since the Pythagorean distance between the two points, $(\bar{\mathbf{x}}^1)' \mathbf{M}_r$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}_r$, which represents the two canonical means, $(\bar{\mathbf{x}}^1)' \mathbf{M}$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}$, in the r -dimensional CVA display approximates the Pythagorean distance between the two points, $(\bar{\mathbf{x}}^1)' \mathbf{M}$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}$, the Pythagorean distance between the two points, $(\bar{\mathbf{x}}^1)' \mathbf{M}_r$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}_r$, can be used as a measure of the dissimilarity between the two groups:

$$\begin{aligned} \frac{n}{n-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) T^2 &= [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]' \left[\frac{1}{n} \mathbf{W} \right]^{-1} [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2] \\ \frac{1}{n-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) T^2 &= [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]' [\mathbf{W}]^{-1} [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2] \\ &= [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2]' \mathbf{M} \mathbf{M}' [\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2] \\ &= [(\bar{\mathbf{x}}^1)' \mathbf{M} - (\bar{\mathbf{x}}^2)' \mathbf{M}] [(\bar{\mathbf{x}}^1)' \mathbf{M} - (\bar{\mathbf{x}}^2)' \mathbf{M}]' \\ &\approx [(\bar{\mathbf{x}}^1)' \mathbf{M}_r - (\bar{\mathbf{x}}^2)' \mathbf{M}_r] [(\bar{\mathbf{x}}^1)' \mathbf{M}_r - (\bar{\mathbf{x}}^2)' \mathbf{M}_r]' . \end{aligned}$$

When the two populations are multivariate normal, it is evident that the null hypothesis in equation (4.2.42) will be rejected when the squared Pythagorean distance between $(\bar{\mathbf{x}}^1)' \mathbf{M}$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}$ is larger than the critical value,

$$\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \frac{p}{n-p-1} F_{(p, n-p-1)}(\alpha) .$$

When however the two populations are not multivariate normal and n_1 and n_2 are large, the null hypothesis in equation (4.2.42) will be rejected when the squared Pythagorean distance between $(\bar{\mathbf{x}}^1)' \mathbf{M}$ and $(\bar{\mathbf{x}}^2)' \mathbf{M}$ is larger than the critical value, $\frac{n}{n-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \chi_p^2$. As mentioned earlier, when n_1 and n_2 are small, inferences about $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ can be made using a permutation test.

Consider now the situation where the samples are structured into more than two groups. Let, as before, the number of groups be denoted by J and the size of the i th

4.2. CVA IS EQUIVALENT TO LDA FOR THE MULTI-GROUP CASE 193

group be n_i with $\sum_{i=1}^J n_i = n$. Hotelling's T^2 -statistic used to test the null hypothesis

$$H_0 : \boldsymbol{\mu}_i - \boldsymbol{\mu}_j = \boldsymbol{\delta}_0$$

against the alternative hypothesis

$$H_a : \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \neq \boldsymbol{\delta}_0$$

is given by

$$T^2 = [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j - \boldsymbol{\delta}_0]' \left[\left(\frac{1}{n_i} + \frac{1}{n_j} \right) \widehat{\boldsymbol{\Sigma}}_{\text{pooled}} \right]^{-1} [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j - \boldsymbol{\delta}_0]$$

where

$$\widehat{\boldsymbol{\Sigma}}_{\text{pooled}}^{ij} = \frac{1}{n_i + n_j - 2} \left(\sum_{k=1}^{n_i} (\mathbf{x}_k^i - \bar{\mathbf{x}}^i) (\mathbf{x}_k^i - \bar{\mathbf{x}}^i)' + \sum_{k=1}^{n_j} (\mathbf{x}_k^j - \bar{\mathbf{x}}^j) (\mathbf{x}_k^j - \bar{\mathbf{x}}^j)' \right).$$

Consider the situation where the null hypothesis

$$H_0 : \boldsymbol{\mu}_i - \boldsymbol{\mu}_j = \mathbf{0} \tag{4.2.44}$$

has to be tested against the alternative hypothesis

$$H_a : \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \neq \mathbf{0}.$$

Hotelling's T^2 -statistic under the null hypothesis in equation (4.2.44) is given by:

$$T_0^2 = [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j]' \left[\left(\frac{1}{n_i} + \frac{1}{n_j} \right) \widehat{\boldsymbol{\Sigma}}_{\text{pooled}}^{ij} \right]^{-1} [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j].$$

In this case (where $J > 2$) there is not an exact relationship between the squared

Pythagorean distance between the two canonical means, $(\bar{\mathbf{x}}^i)' \mathbf{M}$ and $(\bar{\mathbf{x}}^j)' \mathbf{M}$, and T_0^2 since $\hat{\Sigma}_{\text{pooled}}$ is not a constant multiple of the matrix of within-group sums of squares and cross products, \mathbf{W} , as in the case where the samples are structured into only two groups. Since however it is assumed that the within-group covariance matrices of the J populations are identical, the matrix,

$$\frac{1}{n_i + n_j} \left(\sum_{k=1}^{n_i} (\mathbf{x}_k^i - \bar{\mathbf{x}}^i) (\mathbf{x}_k^i - \bar{\mathbf{x}}^i)' + \sum_{k=1}^{n_j} (\mathbf{x}_k^j - \bar{\mathbf{x}}^j) (\mathbf{x}_k^j - \bar{\mathbf{x}}^j)' \right),$$

which is the weighted average of the within-group sample covariance matrices of the i th and j th groups, should be a reasonably good approximation of the common within-group covariance matrix,

$$\frac{1}{n} \mathbf{W} = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j) (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' .$$

Given this, it can be illustrated that the squared Mahalanobis distance between the two group means, $\bar{\mathbf{x}}^i$ and $\bar{\mathbf{x}}^j$, or equivalently the squared Pythagorean distance between the i th and j th canonical means, $(\bar{\mathbf{x}}^i)' \mathbf{M}$ and $(\bar{\mathbf{x}}^j)' \mathbf{M}$, can be used as a measure of the dissimilarity between the i th and j th groups.

$$\begin{aligned} T_0^2 &= [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j]' \left[\left(\frac{1}{n_i} + \frac{1}{n_j} \right) \hat{\Sigma}_{\text{pooled}}^{ij} \right]^{-1} [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j] \\ &\rightarrow \left(\frac{1}{n_i} + \frac{1}{n_j} \right) T_0^2 = [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j]' \left[\frac{1}{n_i + n_j - 2} \left(\sum_{k=1}^{n_i} (\mathbf{x}_k^i - \bar{\mathbf{x}}^i) (\mathbf{x}_k^i - \bar{\mathbf{x}}^i)' \right. \right. \\ &\quad \left. \left. + \sum_{k=1}^{n_j} (\mathbf{x}_k^j - \bar{\mathbf{x}}^j) (\mathbf{x}_k^j - \bar{\mathbf{x}}^j)' \right) \right]^{-1} [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j] \\ &\rightarrow \frac{n_i + n_j}{n_i + n_j - 2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) T_0^2 = [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j]' \left[\frac{1}{n_i + n_j} \left(\sum_{k=1}^{n_i} (\mathbf{x}_k^i - \bar{\mathbf{x}}^i) (\mathbf{x}_k^i - \bar{\mathbf{x}}^i)' \right. \right. \\ &\quad \left. \left. + \sum_{k=1}^{n_j} (\mathbf{x}_k^j - \bar{\mathbf{x}}^j) (\mathbf{x}_k^j - \bar{\mathbf{x}}^j)' \right) \right]^{-1} [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j] \\ &\rightarrow \frac{n_i + n_j}{n_i + n_j - 2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) T_0^2 \approx [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j]' \left[\frac{1}{n} \mathbf{W} \right]^{-1} [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j] \quad (4.2.45) \\ &\rightarrow \frac{1}{n} \frac{n_i + n_j}{n_i + n_j - 2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) T_0^2 \approx [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j]' \mathbf{W}^{-1} [\bar{\mathbf{x}}^i - \bar{\mathbf{x}}^j] \\ &\rightarrow \frac{1}{n} \frac{n_i + n_j}{n_i + n_j - 2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) T_0^2 \approx [(\bar{\mathbf{x}}^i)' \mathbf{M} - (\bar{\mathbf{x}}^j)' \mathbf{M}] [(\bar{\mathbf{x}}^i)' \mathbf{M} - (\bar{\mathbf{x}}^j)' \mathbf{M}]' . \end{aligned}$$

(4.2.46)

It is evident from equation (4.2.45) that if $\frac{1}{n}\mathbf{W}$ is a good approximation of the pooled covariance matrix, $\widehat{\Sigma}_{\text{pooled}}^{ij}$, then an increase in the Mahalanobis distance between the i th and j th group means, $\bar{\mathbf{x}}^i$ and $\bar{\mathbf{x}}^j$ should imply an increase in the value of T_0^2 i.e. an increase in the strength of the evidence against the null hypothesis, $H_0 : \boldsymbol{\mu}_i - \boldsymbol{\mu}_j = \boldsymbol{\delta}_0$. Similarly, it is clear from equation (4.2.46) that if $\frac{1}{n}\mathbf{W}$ is a good approximation of the pooled covariance matrix, $\widehat{\Sigma}_{\text{pooled}}^{ij}$, then an increase in the Pythagorean distance between the i th and j th canonical means, $(\bar{\mathbf{x}}^i)' \mathbf{M}$ and $(\bar{\mathbf{x}}^j)' \mathbf{M}$ should imply an increase in the value of T_0^2 . It follows that if $\frac{1}{n}\mathbf{W}$ is a good approximation of the pooled covariance matrix, $\widehat{\Sigma}_{\text{pooled}}^{ij}$, then the Mahalanobis distance between the i th and j th group means as well as the Pythagorean distance between the i th and j th canonical means can be used as a measure of the dissimilarity between the groups. Since the Pythagorean distance between the two points, $(\bar{\mathbf{x}}^i)' \mathbf{M}_r$ and $(\bar{\mathbf{x}}^j)' \mathbf{M}_r$, representing the i th and j th canonical means in the r -dimensional CVA display approximates the Pythagorean distance between $(\bar{\mathbf{x}}^i)' \mathbf{M}$ and $(\bar{\mathbf{x}}^j)' \mathbf{M}$ (the i th and j th canonical means in the p -dimensional CVA display), it follows that if $\frac{1}{n}\mathbf{W}$ is a good approximation of the pooled covariance matrix, $\widehat{\Sigma}_{\text{pooled}}^{ij}$, the Pythagorean distance between the two points, $(\bar{\mathbf{x}}^i)' \mathbf{M}_r$ and $(\bar{\mathbf{x}}^j)' \mathbf{M}_r$, representing the i th and j th canonical means in the r -dimensional CVA display can be used as a measure of the dissimilarity between the i th and j th groups.

After the groups have been redefined, if at all necessary, it is important to check whether the within-group covariance matrices of all the groups are the same. Box's M-test (Box, 1949) can be used for this purpose. If all the within-group covariance matrices are not equal, CVA is not an appropriate analysis to perform on the data - an analysis of distance (AOD) would then be more appropriate.

4.3 Deriving CVA as a special case of Canonical Correlation Analysis (CCA)

4.3.1 Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) is an analysis that investigates the (linear) relationships between two sets of variables of different but associated kinds (Gittins, 1985). The relationships between two sets of variables are of interest in many fields of study. In ecology, for example, the relationships between animal and plant constituents of a biotic community is a typical example of such relationships which are of interest in practice. The relationships investigated by CCA are symmetric in nature. This means that the variables of the two sets have the same roles and therefore it doesn't matter which set is the first set and which is the second. When investigating symmetric relationships, it is the variance common to both sets of variables which is of interest. CCA has a strong connection with multivariate regression which also investigates the linear relationships between two sets of variables of different types. The relationships investigated by multivariate regression however are non-symmetric in nature, meaning that the variables from the different sets have different roles. Specifically, the one set of variables is a set of predictor variables

while the other is a set of response variables. When non-symmetric relationships are investigated, it is the proportion of the total variance associated with the set of response variables which can be explained by the set of predictor variables which is of interest. When one of the two sets of variables consist of one variable only, multiple correlation and multiple regression can be used to investigate symmetric and non-symmetric relationships between the two sets of variables respectively. In practice, multiple regression is often used inadvisably to investigate the relationships between two multivariate sets of variables. This is done by dismembering one of the two sets of variables and then regressing each of the individual variables onto the other set of variables. This series of multiple regressions however ignores the relationships between the set of variables which gets dismembered. When both sets of variables are multivariate, CCA and multivariate regression are the appropriate analyses to use for investigating symmetric and non-symmetric relationships between two sets of variables respectively - CCA being a generalisation of multiple correlation and multivariate regression being a generalisation of multiple regression. A strong connection exists between CCA and multiple regression - this will be discussed shortly to better the understanding of CCA and the interpretation of canonical coefficients and canonical correlations.

Let p and q denote the number of variables of the first and second set of variables respectively. Let \mathbf{x} denote the stochastic vector variable of the first set of variables and \mathbf{y} the stochastic vector variable of the second set.

As mentioned earlier, CCA studies the linear relationships between the two sets of variables of different kinds. All the information on the linear relationships within and between the two sets of variables is contained in the joint covariance matrix (or correlation matrix) of the two vector variables. Let Σ denote the joint covariance matrix associated with the stochastic vector variables, \mathbf{x} and \mathbf{y} , Σ_{11} denote the covariance matrix associated with \mathbf{x} , Σ_{22} denote the covariance matrix associated with \mathbf{y} and Σ_{12} denote the matrix with ij th element equal to the covariance between x_i and y_j . The covariance matrix, Σ , is therefore given by

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

All the information on the linear associations between the two sets of variables is summarised in the matrix $\Sigma_{12} = \Sigma'_{21}$.

When dealing with samples, the joint covariance matrix associated with \mathbf{x} and \mathbf{y} is unknown and an estimate of the matrix needs to be used. All the information on the linear relationships within and between the two sets of variables which is contained in the n observed samples is summarised in the joint sample covariance matrix (or correlation matrix or some comparable inner-products matrix) of the two vector variables, \mathbf{x} and \mathbf{y} . Denote the joint sample covariance matrix of \mathbf{x} and \mathbf{y} by $\hat{\Sigma}$. The observations made in studies involving two sets of variables of different kinds can be represented by partitioned vectors, each partitioned vector consisting of two subvectors corresponding to the two sets of variables. Let n denote the number of

4.3. DERIVING CVA AS A SPECIAL CASE OF CANONICAL
CORRELATION ANALYSIS (CCA)

197

samples that were observed on each of the $p + q$ variables and let the j th observed sample be denoted by $[\mathbf{x}'_j | \mathbf{y}'_j]$. Let \mathbf{X} denote the $n \times p$ matrix, the j th row of which is given by \mathbf{x}'_j and \mathbf{Y} the matrix the j th row of which is given by \mathbf{y}'_j . The joint sample covariance matrix, $\widehat{\Sigma}$, can then be expressed as:

$$\widehat{\Sigma} = \begin{bmatrix} \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{bmatrix}$$

where

$$\begin{aligned} \widehat{\Sigma}_{11} &= c\mathbf{X}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{X} \\ \widehat{\Sigma}_{22} &= c\mathbf{Y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{Y} \\ \widehat{\Sigma}_{12} &= c\mathbf{X}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{Y} \\ \widehat{\Sigma}_{21} &= c\mathbf{Y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{X} = \widehat{\Sigma}'_{12}. \end{aligned}$$

The data can be regarded as generating two sample spaces - a p -dimensional sample space corresponding to the set of x -variables and a q -dimensional sample space corresponding to the set of y -variables. For convenience, the p -dimensional sample space generated by the set of x -variables will from now on be referred to as the x -space while the q -dimensional sample space generated by the set of y -variables will be referred to as the y -space.

All the information on the linear associations between the two sets of variables that is contained in the n observations is summarised in the matrix, $\widehat{\Sigma}_{12}$. Unfortunately the internal structure of $\widehat{\Sigma}_{12}$ is rarely evident upon inspection, especially when pq is large. Hotelling (1936) proposed that the maximum correlation between a linear combination of the x -variables and a linear combination of the y -variables can be used to measure the linear association between the two sets of variables. This one correlation however contains only some of the information on the linear association between the two sets of variables. In order to describe the complete linear association between the measurement domains, linear combinations other than those that yield the maximum correlation must also be considered. Naturally, it is desirable to summarise the association between the two sets of variables by means of as few correlations as possible - the fewer correlations, the simpler the interpretation of the results and the easier the visual representation of the correlation structure of the data will be. For example, in the case where only one large correlation between a linear combination of the x -variables and a linear combination of the y -variables exists, the correlation structure of the data can be visually represented in two dimensions defined by the pair of linear combinations which produced this large correlation. The aim of CCA therefore is to summarise the linear associations

between the two measurement domains by means of as few correlations as possible. For this reason successive pairs of linear combinations are required to maximise the correlation between them conditional on being uncorrelated with all preceding pairs. The successive pairs of linear combinations are required to be uncorrelated with each other in order to significantly reduce the possibility of linear combinations containing the same information.

Note that the correlation between a linear combination of the x -variables, $\mathbf{a}'\underline{\mathbf{x}}$, and a linear combination of the y -variables, $\mathbf{b}'\underline{\mathbf{y}}$, is identical to the correlation between any scalar multiples of the two linear combinations. Letting d and e denote two arbitrary scalar values, then

$$\begin{aligned} \text{corr}(\mathbf{d}\mathbf{a}'\underline{\mathbf{x}}, \mathbf{e}\mathbf{b}'\underline{\mathbf{y}}) &= \frac{\text{cov}(\mathbf{d}\mathbf{a}'\underline{\mathbf{x}}, \mathbf{e}\mathbf{b}'\underline{\mathbf{y}})}{\sqrt{\text{var}(\mathbf{d}\mathbf{a}'\underline{\mathbf{x}})}\sqrt{\text{var}(\mathbf{e}\mathbf{b}'\underline{\mathbf{y}})}} \\ &= \frac{(de)\text{cov}(\mathbf{a}'\underline{\mathbf{x}}, \mathbf{b}'\underline{\mathbf{y}})}{d\sqrt{\text{var}(\mathbf{a}'\underline{\mathbf{x}})}e\sqrt{\text{var}(\mathbf{b}'\underline{\mathbf{y}})}} \\ &= \frac{\text{cov}(\mathbf{a}'\underline{\mathbf{x}}, \mathbf{b}'\underline{\mathbf{y}})}{\sqrt{\text{var}(\mathbf{a}'\underline{\mathbf{x}})}\sqrt{\text{var}(\mathbf{b}'\underline{\mathbf{y}})}} \\ &= \text{corr}(\mathbf{a}'\underline{\mathbf{x}}, \mathbf{b}'\underline{\mathbf{y}}). \end{aligned}$$

This implies that the linear combinations of the x and y -variables that maximise the correlation between them subject to being uncorrelated with all preceding pairs of linear combinations, are not unique. Constraints therefore need to be imposed on the coefficient vectors defining the linear combinations in order to obtain unique solutions. In CCA it is customary to require that the coefficient vector of a linear combination be such that the variate defined by the linear combination has unit sample variance. The variables defined by those linear combinations which satisfies all the above mentioned requirements are called canonical variables. It is clear that the canonical variables are structured in pairs - each pair consisting of a canonical variate constructed from the x -variables and a canonical variate constructed from the y -variables. When \mathbf{X} and \mathbf{Y} are of full column rank, as will be assumed here, then at most

$$s = \min(p, q)$$

canonical variate pairs exist. The correlation between the two canonical variables constituting a canonical variate pair is called a canonical correlation. The pair of linear combinations, $\mathbf{a}'_1\underline{\mathbf{x}}$ and $\mathbf{b}'_1\underline{\mathbf{y}}$ which yields the maximum obtainable correlation between a linear combination of the x -variables and a linear combinations of the y -variables is called the first canonical variate pair where $\mathbf{a}'_1\underline{\mathbf{x}}$ is the first canoni-

4.3. DERIVING CVA AS A SPECIAL CASE OF CANONICAL CORRELATION ANALYSIS (CCA)

199

cal variate constructed from the x -variables and $\mathbf{b}'_1\mathbf{y}$ is the first canonical variate constructed from the y -variables. In general, the i th canonical variate pair consists of the linear combination of the x -variables, $\mathbf{a}'_i\mathbf{x}$, and the linear combination of the y -variables, $\mathbf{b}'_i\mathbf{y}$, which yield the maximum correlation between a linear combination of the x -variables and a linear combination of the y -variables under the constraint that the canonical variate pair is uncorrelated with each of the previous $i - 1$ canonical variate pairs. Let \mathbf{u} denote the vector of s canonical variables constructed from the set of \mathbf{x} -variables and \mathbf{v} denote the vector of s canonical variables constructed from the set of \mathbf{y} -variables. The joint sample covariance of \mathbf{u} and \mathbf{v} is given by

$$\text{cov} \left\{ \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{u}' & \mathbf{v}' \end{bmatrix} \right\} = \begin{bmatrix} \mathbf{A}' & \mathbf{0}' \\ \mathbf{0} & \mathbf{B}' \end{bmatrix} \begin{bmatrix} \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0}' & \mathbf{B} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_p & \mathbf{\Gamma} \\ \mathbf{\Gamma}' & \mathbf{I}_q \end{bmatrix}.$$

The term, canonical, refers to the mathematical concept, canonical form, which is defined to be a mathematical entity which allows properties of interest to be perceived readily (Gittins, 1985). The covariance matrix between the two sets of canonical variables therefore can be viewed as a canonical form of the covariance matrix between the two sets of measured variables since it summarises the linear relationships between the two measurement domains, which are the properties of interest, in terms of only s non-zero quantities.

The s canonical variables constructed from the \mathbf{x} -variables and those constructed from the \mathbf{y} -variables correspond to the axes of the \mathbf{x} -space and \mathbf{y} -space respectively after the coordinate frames of the respective sample spaces have been rotated (simultaneously) to a new position in which the correlation structure of the data is emphasised. This resembles PCA which involves the rotation of the coordinate frame of the sample space to a new position in which the total variance associated with the single set of variables of interest is summarised and represented. It is possible for the canonical variables associated with a set of variables to coincide with the principal components associated with that set of variables, in which case the canonical variables will also optimally summarise the total variance associated with the set of variables. Generally however, this will not be the case.

The classic solution to CCA, as given by Hotelling (1935), Hotelling (1936) and Anderson (1958), makes use of Lagrange multipliers and eigen-analysis. CCA firstly seeks the linear composites, $u = \mathbf{a}'\mathbf{x}$ and $v = \mathbf{b}'\mathbf{y}$, of the original measurement vectors, \mathbf{x} and \mathbf{y} , which maximises the correlation coefficient between u and v amongst all such pairs of linear composites, that is, CCA seeks $u_1 = \mathbf{a}'_1\mathbf{x}$ and $v_1 = \mathbf{b}'_1\mathbf{y}$ where

$$(\mathbf{a}_1, \mathbf{b}_1) = \underset{(\mathbf{a}, \mathbf{b})}{\text{argmax}} \{r(\mathbf{a}, \mathbf{b})\} = \underset{(\mathbf{a}, \mathbf{b})}{\text{argmax}} \left\{ \frac{\widehat{\text{cov}}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y})}{(\mathbf{a}'\widehat{\Sigma}_{11}\mathbf{a})^{1/2} (\mathbf{b}'\widehat{\Sigma}_{22}\mathbf{b})^{1/2}} \right\}.$$

It is clear that $r(\mathbf{a}, \mathbf{b})$ is independent of the scales of the coefficient vectors, \mathbf{a} and

b. In order to obtain unique solutions to the coefficient vectors, constraints need to be placed on these vectors. It is convenient and customary to use the constraints, $\mathbf{a}'\widehat{\Sigma}_{11}\mathbf{a} = 1$ and $\mathbf{b}'\widehat{\Sigma}_{22}\mathbf{b} = 1$. Under these constraints the correlation coefficient, $r(\mathbf{a}, \mathbf{b})$, between $u = \mathbf{a}'\mathbf{x}$ and $v = \mathbf{b}'\mathbf{y}$ is equal to $\mathbf{a}'\widehat{\Sigma}_{12}\mathbf{b} = \lambda$:

$$\begin{aligned} r(\mathbf{a}, \mathbf{b}) &= \frac{\widehat{\text{cov}}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y})}{\sqrt{\widehat{\text{var}}(\mathbf{a}'\mathbf{x})}\sqrt{\widehat{\text{var}}(\mathbf{b}'\mathbf{y})}} \\ &= \frac{\mathbf{a}'\widehat{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\widehat{\Sigma}_{11}\mathbf{a}}\sqrt{\mathbf{b}'\widehat{\Sigma}_{22}\mathbf{b}}} \\ \rightarrow r(\mathbf{a}, \mathbf{b}) &= \mathbf{a}'\widehat{\Sigma}_{12}\mathbf{b} \text{ if } \mathbf{a}'\widehat{\Sigma}_{11}\mathbf{a} = 1 \text{ and } \mathbf{b}'\widehat{\Sigma}_{22}\mathbf{b} = 1. \end{aligned}$$

The coefficient vectors, \mathbf{a} and \mathbf{b} , that maximise the correlation coefficient under the constraints, $\mathbf{a}'\widehat{\Sigma}_{11}\mathbf{a} = 1$ and $\mathbf{b}'\widehat{\Sigma}_{22}\mathbf{b} = 1$ can be found by equating the partial derivatives of

$$\xi = \mathbf{a}'\widehat{\Sigma}_{12}\mathbf{b} - \frac{1}{2}\lambda(\mathbf{a}'\widehat{\Sigma}_{11}\mathbf{a} - 1) - \frac{1}{2}\mu(\mathbf{b}'\widehat{\Sigma}_{22}\mathbf{b} - 1)$$

with respect to \mathbf{a} and \mathbf{b} respectively to the null vector and then solving for \mathbf{a} and \mathbf{b} respectively. The equations resulting from this procedure are provided below:

$$\begin{aligned} \frac{\partial \xi}{\partial \mathbf{a}} &= \mathbf{0} \\ \rightarrow \widehat{\Sigma}_{12}\mathbf{b} - \lambda\widehat{\Sigma}_{11}\mathbf{a} &= \mathbf{0} \end{aligned} \quad (4.3.1)$$

$$\begin{aligned} \frac{\partial \xi}{\partial \mathbf{b}} &= \mathbf{0} \\ \rightarrow \widehat{\Sigma}_{21}\mathbf{a} - \mu\widehat{\Sigma}_{22}\mathbf{b} &= \mathbf{0}. \end{aligned} \quad (4.3.2)$$

The following two equations are obtained by premultiplying equation (4.3.1) and equation (4.3.2) by \mathbf{a}' and \mathbf{b}' respectively:

$$\mathbf{a}'\widehat{\Sigma}_{12}\mathbf{b} - \lambda\mathbf{a}'\widehat{\Sigma}_{11}\mathbf{a} = 0 \quad (4.3.3)$$

$$\mathbf{b}'\widehat{\Sigma}_{21}\mathbf{a} - \mu\mathbf{b}'\widehat{\Sigma}_{22}\mathbf{b} = 0. \quad (4.3.4)$$

4.3. DERIVING CVA AS A SPECIAL CASE OF CANONICAL
CORRELATION ANALYSIS (CCA)

201

Since $\mathbf{a}'\widehat{\Sigma}\mathbf{a} = 1$ and $\mathbf{b}'\widehat{\Sigma}\mathbf{b} = 1$, both equation (4.3.3) and equation (4.3.4) reduce to

$$\mathbf{a}'\widehat{\Sigma}_{12}\mathbf{b} = \lambda = \mu.$$

Hence, the quantity to be maximised, $r(\mathbf{a}, \mathbf{b})$, is λ . After some algebraic manipulation the following equations are obtained (Gittins, 1985):

$$\widehat{\Sigma}_{21}\widehat{\Sigma}_{11}^{-1}\widehat{\Sigma}_{12}\mathbf{b} = \lambda^2\widehat{\Sigma}_{22}\mathbf{b} \quad (4.3.5)$$

$$\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}\mathbf{a} = \lambda^2\widehat{\Sigma}_{11}\mathbf{a}. \quad (4.3.6)$$

Equation (4.3.5) and equation (4.3.6) are two-sided eigenvalue problems having q and p solutions respectively. Let λ_i^2 denote the i th largest eigenvalue of the two-sided eigenvalue problem in equation (4.3.5), $i \in [1 : q]$. Note that λ_j^2 is the j th largest eigenvalue of the two-sided eigenvalue problems in equation (4.3.5) and equation (4.3.6) for $j \in [1 : s]$, where

$$s = \min(p, q)$$

and that $\lambda_k = 0$ for $k > s$. Let \mathbf{a}_i denote the eigenvector of the two-sided eigenvalue problem in equation (4.3.5) associated with the eigenvalue, λ_i^2 for $i \in [1 : q]$ and let \mathbf{b}_j denote the eigenvector of the two-sided eigenvalue problem in equation (4.3.6) associated with the eigenvalue, λ_j^2 for $j \in [1 : p]$. Since $r(\mathbf{a}, \mathbf{b}) = \lambda$, it follows that the maximum correlation coefficient between a linear composite of the \mathbf{x} -variables and a linear composite of the \mathbf{y} -variables is given by λ_1 , the square root of the largest eigenvalue of both the two-sided eigenvalue-problem in equation (4.3.5) and equation (4.3.6). The linear composites of the x and y -variables respectively which produce this maximum correlation coefficient are $\mathbf{a}'_1\mathbf{x}$ and $\mathbf{b}'_1\mathbf{y}$. The pair of variates, $\mathbf{a}'_1\mathbf{x}$ and $\mathbf{b}'_1\mathbf{y}$, are called the first canonical variate pair or the first pair of canonical variables. Assuming that \mathbf{X} and \mathbf{Y} are of full column rank, there exist s such pairs of canonical variates. The s pairs of canonical variates are defined by those eigenvectors of the two-sided eigenvalue-problems that correspond to non-null eigenvalues. The k th canonical variate pair is given by $\mathbf{a}'_k\mathbf{x}$ and $\mathbf{b}'_k\mathbf{y}$ and the correlation coefficient between the pair of variates is given by the k th largest eigenvalue, λ_k^2 , of the two-sided eigenvalue-problems, equation (4.3.5) and equation (4.3.6).

As mentioned earlier, CCA has a strong connection with multiple regression. In fact, CCA can be viewed as a generalisation of multiple regression from strictly one response variable to two or more response variables. CCA's connection to multiple regression can be described as follows: if the k th canonical variate of the second set of variables, $\mathbf{b}'_k\mathbf{y}$, is regressed onto \mathbf{x} then the vector of regression coefficients is proportional to \mathbf{a}_k (the coefficient vector of the other member of the k th canonical

variate pair), $k = 1, 2, \dots, s$. This means that $\mathbf{a}'_k \mathbf{x}$ is the best predictor of $\mathbf{b}'_k \mathbf{y}$ in the least squares sense, that is

$$\mathbf{a}'_k = \operatorname{argmax}_{\mathbf{a}} \left\{ \sum_{i=1}^n (\mathbf{b}'_k \mathbf{y}_i - \mathbf{a}' \mathbf{x}_i)^2 \right\}.$$

Similarly, if $\mathbf{a}'_k \mathbf{x}$, is regressed onto \mathbf{y} , the vector of regression coefficients is proportional to \mathbf{b}_k . It follows that CCA contains within it s pairs of multiple regression relationships. It is well known that multiple regression minimises the sum of the squared residuals formed by the regression. Since the total sums of squares, which is fixed, can be partitioned into the residual sums of squares and the regression sums of squares, multiple regression maximises the regression sums of squares. Multiple regression therefore also maximises the proportion of predictable variance r^2 (the squared multiple correlation coefficient), which is equal to the ratio of the regression sums of squares to the total sums of squares. This means that the k th canonical variate of the first set of variables, $\mathbf{a}'_k \mathbf{x}$, is the best predictor of $\mathbf{b}'_k \mathbf{y}$ in the least squares sense, and vice versa. Also, the multiple correlation between $\mathbf{b}'_k \mathbf{y}$ and \mathbf{x} is equal to the correlation between $\mathbf{b}'_k \mathbf{y}$ and $\mathbf{a}'_k \mathbf{x}$, that is the k th canonical correlation. Similarly, the multiple correlation between $\mathbf{a}'_k \mathbf{x}$ and \mathbf{y} is equal to the correlation between $\mathbf{a}'_k \mathbf{x}$ and $\mathbf{b}'_k \mathbf{y}$. Since every canonical correlation is equal to some multiple correlation, its square is interpretable as a proportion of explained variance. For example, the squared k th canonical correlation is the proportion of the variance of $\mathbf{b}'_k \mathbf{y}$ explained by \mathbf{x} or the proportion of the variance of $\mathbf{a}'_k \mathbf{x}$ explained by \mathbf{y} .

Canonical correlations and canonical variables are invariant to non-singular linear transformations of the variables of either or both sets (Hotelling, 1936; Kshirsagar, 1972; Mardia *et al.*, 1979). The coefficient vectors of the canonical variables on the other hand are not invariant to non-singular linear transformations of the variables, however, those associated with the transformed measurements are closely related to those associated with the original measurements. When the variables are transformed to have zero mean and unit variance, the covariance matrix associated with the two transformed vector variables is equal to the correlation matrix associated with the two original vector variables. Since the transformation via which variables are transformed to have zero mean and unit variance is a non-singular linear transformation, canonical variables and canonical correlations calculated from the joint sample covariance matrix of \mathbf{x} and \mathbf{y} and those calculated from joint sample correlation matrix of \mathbf{x} and \mathbf{y} , are identical. Therefore, if \mathbf{R}_{11} denotes the correlation matrix associated with \mathbf{x} , \mathbf{R}_{22} denotes the correlation matrix associated with \mathbf{y} and \mathbf{R}_{12} denotes the correlation matrix between \mathbf{x} and \mathbf{y} , then substituting \mathbf{R}_{ij} for $\widehat{\Sigma}_{ij}$ for $i, j \in [1 : 2]$ in the derivations above will yield the same canonical variables and canonical correlations as obtained previously.

4.3.2 CVA as a special case of CCA

When CCA investigates the relationships between two sets of variables, where the one set is a set of 0/1 binary valued dummy variables indicating group membership, the analysis is known as Canonical Variate Analysis (CVA) (Gittins, 1985). CVA therefore investigates the affinities amongst the groups constituting the sample at hand with respect to the set of measured (response) variables considered simultaneously. This type of study is called a structural study. When the samples constituting the data set at hand are structured into J distinct groups, only $(J - 1)$ binary valued dummy variables are needed to perfectly describe the group membership of the samples since samples not belonging to any of the preceding $(J - 1)$ groups, obviously belong to the J th group. Let \mathbf{y} denote the $(J - 1)$ -component vector of 0/1 binary dummy variables with k th element defined as:

$$y_k = \begin{cases} 1 & \text{if the observation belongs to group } k \\ 0 & \text{otherwise.} \end{cases} \quad (4.3.7)$$

Let \mathbf{Y} denote the $n \times (J - 1)$ group indicator matrix which is defined in the following way:

$$[\mathbf{Y}]_{ik} = \begin{cases} 1 & \text{if the } i\text{th observation belongs to group } k \\ 0 & \text{otherwise.} \end{cases}$$

This definition of \mathbf{Y} yields the following expressions for $\mathbf{Y}'\mathbf{Y}$, $\mathbf{Y}'\mathbf{1}$, $\mathbf{Y}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{X}\mathbf{1}$:

$$\mathbf{Y}'\mathbf{Y} = \mathbf{N}_{J-1}$$

where \mathbf{N}_{J-1} is a $(J - 1) \times (J - 1)$ diagonal matrix with k th diagonal element equal to n_k ,

$$\mathbf{Y}\mathbf{1} = \begin{bmatrix} \mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} \\ \vdots \\ \mathbf{1}_{n_{J-1}} \\ \mathbf{0}_{n_J} \end{bmatrix}$$

$$\mathbf{Y}'\mathbf{X} = \begin{bmatrix} n_1 (\bar{\mathbf{x}}^1)' \\ n_2 (\bar{\mathbf{x}}^2)' \\ \vdots \\ n_{J-1} (\bar{\mathbf{x}}^{J-1})' \end{bmatrix}$$

$$\begin{aligned}
\text{and } \mathbf{Y}'\mathbf{X}\mathbf{1} &= \sum_{i=1}^{J-1} n_i (\bar{\mathbf{x}}^i)' \\
&= \sum_{i=1}^J n_i (\bar{\mathbf{x}}^i)' - n_J (\bar{\mathbf{x}}^J)' \\
&= \mathbf{0}' - n_J (\bar{\mathbf{x}}^J)' \\
\longrightarrow \mathbf{Y}'\mathbf{X}\mathbf{1} &= -n_J (\bar{\mathbf{x}}^J)' .
\end{aligned}$$

Since \mathbf{y} has only one non-zero component, $\mathbf{b}'\mathbf{y}$ can take on only one of the following J values: 0 (when all the components of \mathbf{y} are zero i.e. when the observation belongs to group J), b_1, b_2, \dots, b_{J-1} . Each of the J groups are therefore associated with a specific score - group 1 being associated with b_1 , group 2 with b_2 etc. and group J with 0. Since the group to which an object belongs to is in a way dependent on the object's observed measurements on the p x -variables, it makes sense to find a scoring system for the groups such that the p x -variables can be used to optimally predict the group-scores, that is to optimally predict the group to which an object of unknown origin belongs to. It also seems natural that the more the groups are separated, that is the more the group-scores differ, the easier it will be to correctly predict the group to which an object belongs to.

Consider the matrix of predictions associated with the regression of \mathbf{x} onto \mathbf{y} :

$$\widehat{\mathbf{X}} = \mathbf{1}\bar{\mathbf{x}}' + (\mathbf{Y} - \mathbf{1}\bar{\mathbf{y}}') \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21}$$

where $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{1}'\mathbf{X}$ and $\bar{\mathbf{y}} = \frac{1}{n}\mathbf{1}'\mathbf{Y}$. It can be shown that if \mathbf{y} is the $(J-1)$ -component vector defined in (4.3.7), then the matrix of regression sums of squares and cross products corresponding to the regression of \mathbf{x} onto \mathbf{y} ,

$$\mathbf{X}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{Y} \left(\mathbf{Y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{Y} \right)^{-1} \mathbf{Y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{X} = \frac{1}{c} \widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21},$$

is equal to the corrected (for the mean) between-groups sums of squares and cross-products matrix,

$$(\bar{\mathbf{X}} - \mathbf{1}\bar{\mathbf{x}}')' \mathbf{N} (\bar{\mathbf{X}} - \mathbf{1}\bar{\mathbf{x}}')$$

(Kshirsagar, 1972). If $\bar{\mathbf{X}}$ is calculated from a centred data matrix, then $\bar{\mathbf{x}} = \mathbf{0}$ and

hence

$$\frac{1}{c} \widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21} = \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}}.$$

Substituting $\overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}}$ for $\frac{1}{c} \widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21}$ and $\mathbf{X}' \mathbf{X}$ for $\frac{1}{c} \widehat{\Sigma}_{11}$ in the equation,

$$\left(\widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21} - \lambda^2 \widehat{\Sigma}_{11} \right) \mathbf{a} = \mathbf{0}$$

which was derived in Section 4.3 yields

$$\begin{aligned} \left(c \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}} - c \lambda^2 \mathbf{X}' \mathbf{X} \right) \mathbf{a} &= \mathbf{0} \\ \therefore \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}} \mathbf{a} &= \lambda^2 \mathbf{X}' \mathbf{X} \mathbf{a}. \end{aligned} \quad (4.3.8)$$

By the same argument as in Section 4.2, the eigenvector of the two-sided eigenvalue problem in (4.3.8), that corresponds to the largest eigenvalue of (4.3.8), maximises the sums of squares ratio,

$$\frac{\mathbf{a}' \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}} \mathbf{a}}{\mathbf{a}' \mathbf{X}' \mathbf{X} \mathbf{a}}. \quad (4.3.9)$$

It follows that the vector \mathbf{a}_1 that defines the linear combination of the x -variables which optimally predicts the group scores, $\mathbf{b}'\mathbf{y}$, also defines the linear combination of the x -variables which maximally separates the J groups in that it maximises the ratio in (4.3.9). The eigenvector of the two-sided eigenvalue problem in (4.3.8) that corresponds to the i th largest eigenvalue of (4.3.8), \mathbf{a}_i , maximises the ratio in (4.3.9) conditional on being orthogonal to \mathbf{a}_j in the metric $\widehat{\Sigma}_{11}$ for all $j < i$, $i \in [1 : s]$ where $s = \min(p, J - 1)$. Note however that since

$$\mathbf{X}' \mathbf{X} = \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}} + \mathbf{W}$$

a vector \mathbf{a} that maximises the ratio in (4.3.9) also maximises the ratio

$$\frac{\mathbf{a}' \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}} \mathbf{a}}{\mathbf{a}' \mathbf{W} \mathbf{a}}. \quad (4.3.10)$$

It follows that the vector \mathbf{a}_1 that defines the linear combination of the x -variables which optimally predicts the group scores, $\mathbf{b}'\mathbf{y}$, also defines the linear combination of the x -variables which maximally separates the J groups in that it maximises the ratio in (4.3.10). In general, the vector \mathbf{a}_i defines the linear combination of the x -variables that optimally predicts the groups scores conditional on being orthogonal to \mathbf{a}_j in the metric $\widehat{\Sigma}_{11}$ for all $j < i$ and maximises the ratio in (4.3.10) conditional on it being orthogonal to \mathbf{a}_j in the metric $\widehat{\Sigma}_{11}$ for all $j < i$. Recall that the r -dimensional CVA display space described in Section 4.2 is defined by the set of r linear combinations $\left\{\mathbf{m}'_{(i)}\mathbf{x}\right\}_{i=1}^r$ where $\mathbf{m}_{(1)}$ maximises the ratio in (4.3.10) under the constraint $\mathbf{m}'_{(1)}\mathbf{W}\mathbf{m}_{(1)} = 1$ while $\mathbf{m}_{(i)}$ maximises the ratio in (4.3.10) under the constraint $\mathbf{m}'_{(i)}\mathbf{W}\mathbf{m}_{(i)} = 1$ and conditional on being orthogonal to $\mathbf{m}_{(j)}$ in the metric \mathbf{W} , $j < i$. It follows that the r -linear combinations $\left\{\mathbf{a}'_i\mathbf{x}\right\}_{i=1}^r$, like the r linear combinations $\left\{\mathbf{m}'_{(i)}\mathbf{x}\right\}_{i=1}^r$, optimally separates the J groups in r -dimensional space and only differ from the latter with respect to their scaling and the metric with respect to which they are orthogonal to each other.

4.4 CVA as a two-step procedure

From Section 4.2 and Section 4.3 it may seem as if obtaining the r -dimensional CVA solution is comprised of one step only, namely finding the first r eigenvectors of the two-sided eigenvalue problem,

$$\mathbf{B}\mathbf{m} = \lambda\mathbf{W}\mathbf{m}$$

where $\mathbf{B} = \overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}}$ yields the weighted CVA solution and $\mathbf{B} = \overline{\mathbf{X}}'\overline{\mathbf{X}}$ yields the unweighted CVA solution, $r \in [1 : K]$. Although this is true mathematically, the process of obtaining the r -dimensional CVA solution can be best explained as consisting of two steps, the first of which is a transformation from the p -dimensional measurement space to a new p -dimensional space and the second of which is a least squares approximation performed in the transformed space. This approach to CVA has not received much attention in the past but is discussed in detail in Gower *et al.* (2011). The rest of this section is devoted to this approach.

Recall from Section 4.2 that the r linear combinations that define the r -dimensional weighted CVA solution, that is the r linear combinations, $\left\{\mathbf{x}'\mathbf{f}_{(i)}\right\}_{i=1}^r$, which maximally separate the J groups in r -dimensional space in that they yield the largest possible value of the trace,

$$\text{tr} \left\{ \left(\mathbf{F}'_r \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}} \mathbf{F}_r \right) \left(\mathbf{F}'_r \mathbf{W} \mathbf{F}_r \right)^{-1} \right\} \quad (4.4.1)$$

across all $p \times r$ matrices of full column rank, \mathbf{F}_r , are given by $\left\{\mathbf{x}'\mathbf{m}_{(i)}\right\}_{i=1}^r$, where the

matrix \mathbf{M} satisfies the two-sided eigenvalue problem,

$$\bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \mathbf{M} = \mathbf{W} \mathbf{M} \mathbf{\Lambda}$$

$r \in [1 : K]$. That is, $\mathbf{F}_r = \mathbf{M}_r$ yields the largest possible value of the trace in (4.4.1) across all $p \times r$ matrices of full column rank. Recall that the Pythagorean distances in the p -dimensional canonical space (that is the row space of $\mathbf{X}\mathbf{M}$) are proportional to the corresponding Mahalanobis distances in the p -dimensional measurement space (see Section 4.2).

It can be shown that given a $p \times p$ non-singular matrix \mathbf{L} which is such that

$$\mathbf{L}\mathbf{L}' = \mathbf{W}^{-1}$$

the matrix

$$\mathbf{M} = \mathbf{L}\mathbf{V}^{\mathbf{N}}$$

where $\mathbf{V}^{\mathbf{N}}$ is the $p \times p$ orthogonal matrix with column vectors equal to the right singular vectors of the matrix $\mathbf{N}^{1/2} \bar{\mathbf{X}} \mathbf{L}$, satisfies the two-sided eigenvalue problem,

$$\bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \mathbf{M} = \mathbf{W} \mathbf{M} \mathbf{\Lambda} .$$

Letting the svd of the matrix $\mathbf{N}^{1/2} \bar{\mathbf{X}} \mathbf{L}$ be given by

$$\mathbf{N}^{1/2} \bar{\mathbf{X}} \mathbf{L} = \mathbf{U}^{\mathbf{N}} \mathbf{D}^{\mathbf{N}} (\mathbf{V}^{\mathbf{N}})'$$

this is shown below:

$$\begin{aligned} \mathbf{N}^{1/2} \bar{\mathbf{X}} \mathbf{L} &= \mathbf{U}^{\mathbf{N}} \mathbf{D}^{\mathbf{N}} (\mathbf{V}^{\mathbf{N}})' \\ \longrightarrow \mathbf{L}' \bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \mathbf{L} &= \mathbf{V}^{\mathbf{N}} (\mathbf{D}_p^{\mathbf{N}})^2 (\mathbf{V}^{\mathbf{N}})' \\ \longrightarrow \bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \mathbf{L} \mathbf{V} &= (\mathbf{L}^{-1})' \mathbf{V}^{\mathbf{N}} (\mathbf{D}_p^{\mathbf{N}})^2 \\ \longrightarrow \bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \mathbf{L} \mathbf{V} &= (\mathbf{L}^{-1})' \mathbf{L}^{-1} \mathbf{L} \mathbf{V}^{\mathbf{N}} (\mathbf{D}_p^{\mathbf{N}})^2 \\ &\longrightarrow \mathbf{B} \mathbf{M} = \mathbf{W} \mathbf{M} \mathbf{\Lambda} \end{aligned}$$

where $\mathbf{B} = \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}}$, $\mathbf{M} = \mathbf{L} \mathbf{V}^{\mathbf{N}}$ and $\mathbf{\Lambda} = (\mathbf{D}_p^{\mathbf{N}})^2$. It follows that the coordinate vectors of the points representing the i th sample and the j th group mean in the p -dimensional canonical space are given by the i th row vector of

$$\mathbf{X} \mathbf{M} = \mathbf{X} \mathbf{L} \mathbf{V}^{\mathbf{N}}$$

and the j th row vector of

$$\overline{\mathbf{X}} \mathbf{M} = \overline{\mathbf{X}} \mathbf{L} \mathbf{V}^{\mathbf{N}}$$

respectively, $i \in [1 : n]$, $j \in [1 : J]$. Note that $\mathbf{x}_i' \mathbf{L} \mathbf{V}^{\mathbf{N}}$ and $(\overline{\mathbf{x}}^j)' \mathbf{L} \mathbf{V}^{\mathbf{N}}$ are the coordinate vectors of the points $\mathbf{x}_i' \mathbf{L}$ and $(\overline{\mathbf{x}}^j)' \mathbf{L}$ with respect to the column vectors of $\mathbf{V}^{\mathbf{N}}$, $i \in [1 : n]$, $j \in [1 : J]$. It follows that the p -dimensional canonical space, that is the row space of $\mathbf{X} \mathbf{M}$, is identical to the row space of $\mathbf{X} \mathbf{L}$. Recall that the Pythagorean distances in the p -dimensional canonical space are proportional to the corresponding Mahalanobis distances in the p -dimensional measurement space (see Section 4.2). Since each of the row vectors of $\overline{\mathbf{X}} \mathbf{L}$ is a linear combination of row vectors of $\mathbf{X} \mathbf{L}$ and $\overline{\mathbf{X}} \mathbf{L}$ is of rank K , the row space of $\overline{\mathbf{X}} \mathbf{L}$ is a K -dimensional subspace of the row space of $\mathbf{X} \mathbf{L}$. Note that since the row space of $\mathbf{N}^{1/2} \overline{\mathbf{X}} \mathbf{L}$ is identical to the row space of $\overline{\mathbf{X}} \mathbf{L}$, the column vectors of $\mathbf{V}_K^{\mathbf{N}}$ form an orthogonal basis for the row space of $\overline{\mathbf{X}} \mathbf{L}$, that is $\mathcal{V}((\overline{\mathbf{X}} \mathbf{L})') = \mathcal{V}(\mathbf{V}_K^{\mathbf{N}})$:

$$\begin{aligned} \mathcal{V}((\overline{\mathbf{X}} \mathbf{L})') &= \mathcal{V}(\mathbf{L}' \overline{\mathbf{X}}') \\ &= \mathcal{V}(\mathbf{L}' \overline{\mathbf{X}}' \mathbf{N}^{1/2}) \\ &= \mathcal{V}(\mathbf{L}' \overline{\mathbf{X}}' \mathbf{N}^{1/2} \mathbf{N}^{1/2} \overline{\mathbf{X}} \mathbf{L}) \\ &= \mathcal{V}(\mathbf{V}^{\mathbf{N}} \mathbf{D}^{\mathbf{N}} (\mathbf{U}^{\mathbf{N}})' \mathbf{U}^{\mathbf{N}} \mathbf{D}^{\mathbf{N}} (\mathbf{V}^{\mathbf{N}})') \\ &= \mathcal{V}(\mathbf{V}_K^{\mathbf{N}} (\mathbf{D}_K^{\mathbf{N}})^2 (\mathbf{V}_K^{\mathbf{N}})') \\ &= \mathcal{V}(\mathbf{V}_K^{\mathbf{N}} \mathbf{D}_K^{\mathbf{N}}) \\ \rightarrow \mathcal{V}((\overline{\mathbf{X}} \mathbf{L})') &= \mathcal{V}(\mathbf{V}_K^{\mathbf{N}}) . \end{aligned}$$

Since the last $p - K$ column vectors of the matrix $\mathbf{V}^{\mathbf{N}}$ lie orthogonal to $\mathcal{V}(\mathbf{V}_K^{\mathbf{N}})$, it follows that they form an orthogonal basis for the $(p - K)$ -dimensional orthogonal complement of $\mathcal{V}(\mathbf{V}_K^{\mathbf{N}})$. This means that the p column vectors of $\mathbf{V}^{\mathbf{N}}$ form an orthogonal basis for the row space of $\mathbf{X} \mathbf{L}$, that is the p column vectors of $\mathbf{V}^{\mathbf{N}}$ form an orthogonal basis for the p -dimensional canonical space.

The coordinate vectors of the points representing the i th sample and the j th

group mean in the r -dimensional weighted CVA display space are given by the i th row vector of

$$\mathbf{X}\mathbf{M}_r = \mathbf{X}\mathbf{L}\mathbf{V}_r^{\mathbf{N}}$$

and the j th row vector of the

$$\overline{\mathbf{X}}\mathbf{M}_r = \overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{N}}$$

respectively. Note that $\mathbf{x}'_i\mathbf{L}\mathbf{V}_r^{\mathbf{N}}$ and $(\overline{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_r^{\mathbf{N}}$ are the coordinate vectors of the points $\mathbf{x}'_i\mathbf{L}\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'$ and $(\overline{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'$ with respect to the column vectors of $\mathbf{V}_r^{\mathbf{N}}$ respectively, $i \in [1 : n]$, $j \in [1 : J]$. It is evident that the r -dimensional weighted CVA display space is equal to the row space of $\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'$ or equivalently, the column space of $\mathbf{V}_r^{\mathbf{N}}$:

$$\begin{aligned} \mathcal{V}\left(\left(\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\right)'\right) &= \mathcal{V}\left(\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\mathbf{L}'\overline{\mathbf{X}}'\right) \\ &= \mathcal{V}\left(\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\mathbf{L}'\overline{\mathbf{X}}'\mathbf{N}^{1/2}\right) \\ &= \mathcal{V}\left(\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\mathbf{L}'\overline{\mathbf{X}}'\mathbf{N}^{1/2}\mathbf{N}^{1/2}\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\right) \\ &= \mathcal{V}\left(\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\mathbf{V}_K^{\mathbf{N}}(\mathbf{D}_K^{\mathbf{N}})^2(\mathbf{V}_K^{\mathbf{N}})'\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\right) \\ &= \mathcal{V}\left(\mathbf{V}_r^{\mathbf{N}}\mathbf{D}_r^{\mathbf{N}}\mathbf{D}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\right) \\ &= \mathcal{V}\left(\mathbf{V}_r^{\mathbf{N}}\mathbf{D}_r^{\mathbf{N}}\right) \\ \longrightarrow \mathcal{V}\left(\left(\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'\right)'\right) &= \mathcal{V}\left(\mathbf{V}_r^{\mathbf{N}}\right). \end{aligned}$$

Consider the trace,

$$\text{tr}\left\{\left(\mathbf{N}^{1/2}\overline{\mathbf{X}}\mathbf{L} - \mathbf{F}\right)\left(\mathbf{N}^{1/2}\overline{\mathbf{X}}\mathbf{L} - \mathbf{F}\right)'\right\} \quad (4.4.2)$$

where \mathbf{F} denotes an arbitrary $J \times p$ matrix of rank r . According to the Eckart-Young theorem the trace in (4.4.2) is minimised across all $J \times p$ matrices of rank r , \mathbf{F} , when

$$\mathbf{F} = \mathbf{N}^{1/2}\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{N}}(\mathbf{V}_r^{\mathbf{N}})'. \quad .$$

Since any $J \times p$ matrix of rank r can be expressed as $\mathbf{Y}\mathbf{Q}_r\mathbf{Q}'_r$ where \mathbf{Y} is a $J \times p$ matrix and \mathbf{Q}_r is a $p \times r$ matrix with orthogonal column vectors, the trace in (4.4.2) can be expressed as

$$\text{tr} \left\{ (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{P}\mathbf{Q}_r\mathbf{Q}'_r) (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{P}\mathbf{Q}_r\mathbf{Q}'_r)' \right\} \quad (4.4.3)$$

$$\begin{aligned} &= \text{tr} \left\{ (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r) (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r)' \right\} \\ &+ \text{tr} \left\{ (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r - \mathbf{P}\mathbf{Q}_r\mathbf{Q}'_r) (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r - \mathbf{P}\mathbf{Q}_r\mathbf{Q}'_r)' \right\}. \end{aligned} \quad (4.4.4)$$

Hence, according to the Eckart-Young theorem the trace in (4.4.3) is minimised across all $J \times p$ matrices of rank r , $\mathbf{P}\mathbf{Q}_r\mathbf{Q}'_r$, when

$$\mathbf{P} = \mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} \text{ and } \mathbf{Q}_r = \mathbf{V}_r^{\mathbf{N}}.$$

It is evident that $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$ is the r -dimensional subspace of the p -dimensional canonical space that is closest to the set of J points $\left\{ n_j^{1/2} (\bar{\mathbf{x}}^j)' \mathbf{L} \right\}$ in that $\mathbf{Q}_r = \mathbf{V}_r^{\mathbf{N}}$ yields the smallest possible value of the trace,

$$\begin{aligned} &\text{tr} \left\{ (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r) (\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r)' \right\} \\ &= \sum_{j=1}^J \left(n_j^{1/2} (\bar{\mathbf{x}}^j)' \mathbf{L} - n_j^{1/2} (\bar{\mathbf{x}}^j)' \mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r \right) \left(n_j^{1/2} (\bar{\mathbf{x}}^j)' \mathbf{L} - n_j^{1/2} (\bar{\mathbf{x}}^j)' \mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r \right)'. \end{aligned} \quad (4.4.5)$$

That is, the r -dimensional weighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$, is that r -dimensional subspace of the p -dimensional canonical space that yields the smallest possible value of the sum of the squared Pythagorean distances between the points $\left\{ n_j^{1/2} (\bar{\mathbf{x}}^j)' \mathbf{L} \right\}$ and their orthogonal projections onto the r -dimensional subspace. It follows that the points representing the group means in the r -dimensional weighted CVA display, $(\bar{\mathbf{x}}^j)' \mathbf{L}\mathbf{V}_r^{\mathbf{N}} (\mathbf{V}_r^{\mathbf{N}})'$, are obtained by firstly transforming the group means to the p -dimensional canonical space, that is the p -dimensional space in which the Pythagorean distances are proportional to the corresponding Mahalanobis distances in the p -dimensional measurement space, and then projecting the transformed group means onto the r -dimensional subspace which is closest to the set of J points, $\left\{ n_j^{1/2} (\bar{\mathbf{x}}^j)' \mathbf{L} \right\}$, in terms of least squares. Note that the minimisation criteria in (4.4.5) can be expressed as the weighted sum of squared Pythagorean distances between the points $\left\{ (\bar{\mathbf{x}}^j)' \mathbf{L} \right\}$ and their orthogonal projections onto an r -dimensional subspace of the

p -dimensional canonical space:

$$\begin{aligned} & \text{tr} \left\{ (\mathbf{N}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{N}^{1/2} \bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r) (\mathbf{N}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{N}^{1/2} \bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r)' \right\} \\ &= \sum_{j=1}^J n_j \left((\bar{\mathbf{x}}^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r \right) \left((\bar{\mathbf{x}}^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r \right)' . \end{aligned} \quad (4.4.6)$$

This means that the r -dimensional weighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$, is that r -dimensional subspace of the p -dimensional canonical space that yields the smallest possible weighted sum of the squared Pythagorean distances between the points that represent the group means in the p -dimensional canonical space, $\{(\bar{\mathbf{x}}^j)' \mathbf{L}\}$, and their orthogonal projections onto the r -dimensional subspace, the weights being equal to the corresponding group sizes. The fact that the larger groups receive greater weights in the minimization criteria in (4.4.6), implies that the r -dimensional weighted CVA display space will be drawn towards the points representing the group means of the larger groups in the p -dimensional canonical space.

It is shown below that the right singular vectors of the matrix $\mathbf{X}\mathbf{L}$ are identical to the right singular vectors of the matrix $\mathbf{N}^{1/2} \bar{\mathbf{X}}\mathbf{L}$:

$$\begin{aligned} \mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L} &= \mathbf{L}' \left(\bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} + \mathbf{W} \right) \mathbf{L} \\ &= \mathbf{L}' \bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}} \mathbf{L} + \mathbf{L}' \mathbf{W} \mathbf{L} \\ &= \mathbf{V}^{\mathbf{N}} \boldsymbol{\Lambda} (\mathbf{V}^{\mathbf{N}})' + \mathbf{I} \\ \longrightarrow \mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L} &= \mathbf{V}^{\mathbf{N}} (\boldsymbol{\Lambda} + \mathbf{I}) (\mathbf{V}^{\mathbf{N}})' . \end{aligned}$$

It is evident that the p right singular vectors of $\mathbf{N}^{1/2} \bar{\mathbf{X}}\mathbf{L}$ form an orthogonal basis for the row space of $\mathbf{X}\mathbf{L}$, that is an orthogonal basis for the p -dimensional canonical space. Since the r -dimensional weighted CVA display space is spanned by the first r right singular vectors of $\mathbf{X}\mathbf{L}$, the r -dimensional weighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$, is the r -dimensional linear subspace of the p -dimensional canonical space that is closest to the set of n points, $\{\mathbf{x}'_i \mathbf{L}\}$, in that

$$\text{tr} \left\{ (\mathbf{X}\mathbf{L} - \mathbf{X}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r) (\mathbf{X}\mathbf{L} - \mathbf{X}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r)' \right\}$$

is minimised over all $p \times r$ matrices of full column rank, \mathbf{Q}_r , when $\mathbf{Q}_r = \mathbf{V}_r^{\mathbf{N}}$, $r \in [1 : K]$. Since $\mathbf{X}\mathbf{L}$ is centred such that

$$\mathbf{1}'\mathbf{X}\mathbf{L} = \mathbf{0}'$$

the Pythagorean distances between the points representing the n individual samples in the p -dimensional canonical space are optimally approximated by the corresponding Pythagorean distances in the r -dimensional weighted CVA display space (see Section 1.6.11). Since the Pythagorean distances in the p -dimensional canonical space are proportional to the corresponding Mahalanobis distances in the p -dimensional measurement space, it follows that the r -dimensional weighted CVA display space is the r -dimensional subspace of the p -dimensional canonical space which yields the most accurate representation of the true intersample relationships, as measured by the Mahalanobis distance metric.

The $p \times p$ non-singular transformation matrix \mathbf{L} which is such that $\mathbf{L}\mathbf{L}' = \mathbf{W}^{-1}$ can be obtained from the spectral decomposition of the matrix \mathbf{W} or its inverse, \mathbf{W}^{-1} (Gower *et al.*, 2011). Consider the spectral decomposition of \mathbf{W} :

$$\mathbf{W} = \mathbf{E}\mathbf{\Xi}\mathbf{E}'$$

where \mathbf{E} is a $p \times p$ orthogonal matrix and $\mathbf{\Xi}$ is a $p \times p$ diagonal matrix. Since \mathbf{W} is a positive definite matrix, all p its eigenvalues are positive values and hence each of the diagonal matrices $\mathbf{\Xi}$, $\mathbf{\Xi}^{-1}$, $\mathbf{\Xi}^{1/2}$ and $\mathbf{\Xi}^{-1/2}$ is a $p \times p$ non-singular matrix. The spectral decomposition of \mathbf{W}^{-1} follows as:

$$\mathbf{W}^{-1} = \mathbf{E}\mathbf{\Xi}^{-1}\mathbf{E}' \quad (4.4.7)$$

and hence the matrix \mathbf{W}^{-1} can be expressed as

$$\begin{aligned} \mathbf{W}^{-1} &= (\mathbf{E}\mathbf{\Xi}^{-1/2})(\mathbf{E}\mathbf{\Xi}^{-1/2})' \\ \longrightarrow \mathbf{W}^{-1} &= \mathbf{L}\mathbf{L}' \text{ where } \mathbf{L} = \mathbf{E}\mathbf{\Xi}^{-1/2}. \end{aligned} \quad (4.4.8)$$

Since both \mathbf{E} and $\mathbf{\Xi}^{-1/2}$ are $p \times p$ non-singular matrices, the matrix $\mathbf{L} = \mathbf{E}\mathbf{\Xi}^{-1/2}$ is a $p \times p$ non-singular matrix. Since the column vectors of the matrix \mathbf{E} are eigenvectors of \mathbf{W} as well as of \mathbf{W}^{-1} and the column vectors of

$$\mathbf{E}\mathbf{\Xi}^{-1/2}$$

are scalar multiples of the column vectors of \mathbf{E} , it follows that the column vectors of $\mathbf{L} = \mathbf{E}\mathbf{\Xi}^{-1/2}$ are eigenvectors of \mathbf{W} as well as of \mathbf{W}^{-1} . The column vectors of the $p \times p$ non-singular matrix, \mathbf{L} which is such that $\mathbf{L}\mathbf{L}' = \mathbf{W}^{-1}$ are therefore eigenvectors

of the matrices \mathbf{W} and \mathbf{W}^{-1} , normalised such that

$$\mathbf{L}'\mathbf{W}\mathbf{L} = \mathbf{I}.$$

In the context of this two-step approach to CVA and the CVA display, the i th canonical variable is defined as being the linear combination, $y_i = \underline{\mathbf{x}}'\boldsymbol{\ell}_{(i)}$ where $\boldsymbol{\ell}_{(i)}$ denotes the i th column vector of the matrix \mathbf{L} defined in (4.4.8) (Gower *et al.*, 2011). The p -dimensional canonical space, that is the row space of $\mathbf{X}\mathbf{L}$, contains all possible observations of the vector of p canonical variables, $\underline{\mathbf{y}} = \mathbf{L}'\underline{\mathbf{x}}$. Let \mathbf{Y} denote the $n \times p$ matrix with i th row vector equal to the i th canonical sample, $\mathbf{y}'_i = \mathbf{x}'_i\mathbf{L}$, that is:

$$\mathbf{Y} = \mathbf{X}\mathbf{L}.$$

The j th canonical mean is defined to be the transformed j th group mean, that is $(\bar{\mathbf{y}}^j)' = (\bar{\mathbf{x}}^j)'\mathbf{L}$. The $J \times p$ matrix with j th row vector equal to the j th canonical mean is given by:

$$\begin{aligned} \bar{\mathbf{Y}} &= \bar{\mathbf{X}}\mathbf{L} \\ \longrightarrow \bar{\mathbf{Y}} &= \mathbf{N}^{-1}\mathbf{G}'\mathbf{X}\mathbf{L} \\ \longrightarrow \bar{\mathbf{Y}} &= \mathbf{N}^{-1}\mathbf{G}'\mathbf{Y}. \end{aligned}$$

Recall that the row space of $\bar{\mathbf{X}}\mathbf{L}$ is a K -dimensional subspace of the row space of $\mathbf{X}\mathbf{L}$. This means that the J canonical means are contained in a K -dimensional subspace of the p -dimensional canonical space. The common within-group sample covariance matrix associated with the vector of canonical variables, $\underline{\mathbf{y}}$, is given by:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}}^{\mathbf{y}} &= \mathbf{L}'\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}}\mathbf{L} \\ \longrightarrow \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}}^{\mathbf{y}} &= \frac{1}{n}\mathbf{L}'\mathbf{W}\mathbf{L} \\ \longrightarrow \widehat{\boldsymbol{\Sigma}}_{\mathbf{W}}^{\mathbf{y}} &= \frac{1}{n}\mathbf{I}. \end{aligned}$$

It is evident that within each group, the p canonical variables are uncorrelated and have identical sample variances.

Recall that the r linear combinations that define the r -dimensional unweighted CVA solution, that is the r linear combinations, $\{\mathbf{x}'\mathbf{f}_{(i)}\}_{i=1}^r$, which maximally separate the J groups in r -dimensional space in that they yield the largest possible value

of the trace,

$$\text{tr} \left\{ \left(\mathbf{F}'_r \bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{F}_r \right) \left(\mathbf{F}'_r \mathbf{W} \mathbf{F}_r \right)^{-1} \right\}$$

across all $p \times r$ matrices of full column rank, \mathbf{F}_r , are given by $\{\mathbf{x}'\mathbf{m}_{(i)}\}_{i=1}^r$, where the matrix \mathbf{M} satisfies

$$\bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{M} = \mathbf{W} \mathbf{M} \mathbf{\Lambda}$$

$r \in [1 : K]$. It can be shown that the matrix

$$\mathbf{M}^{\mathbf{I}} = \mathbf{L} \mathbf{V}^{\mathbf{I}}$$

where \mathbf{L} is as defined in (4.4.8) and $\mathbf{V}^{\mathbf{I}}$ is the $p \times p$ orthogonal matrix with column vectors equal to the right singular vectors of the matrix $\bar{\mathbf{X}} \mathbf{L}$, satisfies the two-sided eigenvalue problem,

$$\bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{M} = \mathbf{W} \mathbf{M} \mathbf{\Lambda} . \quad (4.4.9)$$

Letting the svd of the matrix $\bar{\mathbf{X}} \mathbf{L}$ be given by

$$\bar{\mathbf{X}} \mathbf{L} = \mathbf{U}^{\mathbf{I}} \mathbf{D}^{\mathbf{I}} (\mathbf{V}^{\mathbf{I}})'$$

this is shown below:

$$\begin{aligned} \rightarrow \mathbf{L}' \bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{L} &= \mathbf{V}^{\mathbf{I}} (\mathbf{D}_p^{\mathbf{I}})^2 (\mathbf{V}^{\mathbf{I}})' \\ \rightarrow \bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{L} \mathbf{V} &= (\mathbf{L}^{-1})' \mathbf{V}^{\mathbf{I}} (\mathbf{D}_p^{\mathbf{I}})^2 \\ \rightarrow \bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{L} \mathbf{V} &= (\mathbf{L}^{-1})' \mathbf{L}^{-1} \mathbf{L} \mathbf{V}^{\mathbf{I}} (\mathbf{D}_p^{\mathbf{I}})^2 \\ \rightarrow \bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{M}^{\mathbf{I}} &= \mathbf{W} \mathbf{M}^{\mathbf{I}} \mathbf{\Lambda}^{\mathbf{I}} \end{aligned}$$

where $\mathbf{M}^{\mathbf{I}} = \mathbf{L} \mathbf{V}^{\mathbf{I}}$ and $\mathbf{\Lambda}^{\mathbf{I}} = (\mathbf{D}_p^{\mathbf{I}})^2$. Hence, the coordinate vectors of the points representing the i th sample and j th group mean in the r -dimensional unweighted

CVA display space are given by the i th row vector of

$$\mathbf{X}\mathbf{M}_r^{\mathbf{I}} = \mathbf{X}\mathbf{L}\mathbf{V}_r^{\mathbf{I}}$$

and the j th row vector of

$$\overline{\mathbf{X}}\mathbf{M}_r^{\mathbf{I}} = \overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{I}}$$

respectively, $r \in [1 : K]$. Recall that the r -dimensional CVA display space where $r > K$ is not uniquely defined as each of the last $p - K$ eigenvalues of the two-sided eigenvalue problem in (4.4.9) is equal to zero.

Note that the column vectors of $\mathbf{V}_K^{\mathbf{I}}$, like the column vectors of $\mathbf{V}_K^{\mathbf{N}}$, form an orthogonal basis for the row space of $\overline{\mathbf{X}}\mathbf{L}$:

$$\begin{aligned} \nu((\overline{\mathbf{X}}\mathbf{L})') &= \nu(\mathbf{L}'\overline{\mathbf{X}}') \\ &= \nu(\mathbf{L}'\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{L}) \\ &= \nu(\mathbf{V}^{\mathbf{I}}\mathbf{D}^{\mathbf{I}}(\mathbf{U}^{\mathbf{I}})'\mathbf{U}^{\mathbf{I}}\mathbf{D}^{\mathbf{I}}(\mathbf{V}^{\mathbf{I}})') \\ &= \nu(\mathbf{V}_K^{\mathbf{I}}(\mathbf{D}_K^{\mathbf{I}})^2(\mathbf{V}_K^{\mathbf{I}})') \\ &= \nu(\mathbf{V}_K^{\mathbf{I}}\mathbf{D}_K^{\mathbf{I}}) \\ \longrightarrow \nu((\overline{\mathbf{X}}\mathbf{L})') &= \nu(\mathbf{V}_K^{\mathbf{I}}) . \end{aligned}$$

It follows that the K -dimensional unweighted and K -dimensional weighted CVA display spaces are identical. Due to the fact that projection matrices are unique, it follows that

$$\mathbf{V}_K^{\mathbf{I}}(\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\mathbf{N}}(\mathbf{V}_K^{\mathbf{N}})' . \quad (4.4.10)$$

Since $(\overline{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_K^{\mathbf{I}}$ and $(\overline{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_K^{\mathbf{N}}$ are the coordinate vectors of the points $(\overline{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_K^{\mathbf{I}}(\mathbf{V}_K^{\mathbf{I}})'$ and $(\overline{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_K^{\mathbf{N}}(\mathbf{V}_K^{\mathbf{N}})'$ with respect to the column vectors of $\mathbf{V}_K^{\mathbf{I}}$ and $\mathbf{V}_K^{\mathbf{N}}$ respectively, it follows that the points representing the group means in the K -dimensional unweighted and K -dimensional weighted CVA displays are identical. This implies that the classification regions of the K -dimensional unweighted and K -dimensional weighted CVA displays are identical. As a result of (4.4.10) the points representing an arbitrary sample in the K -dimensional unweighted and K -dimensional weighted CVA displays are also identical. This implies that all classifications made using the

classification regions of the K -dimensional weighted and K -dimensional unweighted CVA displays are identical. This is in agreement with Section 4.2.

Given that the K column vectors of $\mathbf{V}_K^{\mathbf{I}}$ form an orthogonal basis for a K -dimensional subspace of the p -dimensional canonical space and the last $p-K$ column vectors of $\mathbf{V}^{\mathbf{I}}$ form an orthogonal basis for the orthogonal complement of $\mathcal{V}(\mathbf{V}_K^{\mathbf{I}})$, the p column vectors of $\mathbf{V}^{\mathbf{I}}$ (like the p column vectors of $\mathbf{V}^{\mathbf{N}}$) form an orthogonal basis for the p -dimensional canonical space.

Consider the trace,

$$\text{tr} \left\{ (\bar{\mathbf{X}}\mathbf{L} - \mathbf{F})(\bar{\mathbf{X}}\mathbf{L} - \mathbf{F})' \right\} \quad (4.4.11)$$

where \mathbf{F} denotes an arbitrary $J \times p$ matrix of rank r . According to the Eckart-Young theorem the trace in (4.4.2) is minimised across all $J \times p$ matrices of rank r , \mathbf{F} , when

$$\mathbf{F} = \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r^{\mathbf{I}}(\mathbf{V}_r^{\mathbf{I}})'$$

Since any $J \times p$ matrix of rank r can be expressed as $\mathbf{Y}\mathbf{Q}_r\mathbf{Q}_r'$ where \mathbf{Y} is a $J \times p$ matrix and \mathbf{Q}_r is a $p \times r$ matrix with orthogonal column vectors, the trace in (4.4.11) can be expressed as

$$\begin{aligned} & \text{tr} \left\{ (\bar{\mathbf{X}}\mathbf{L} - \mathbf{P}\mathbf{Q}_r\mathbf{Q}_r')(\bar{\mathbf{X}}\mathbf{L} - \mathbf{P}\mathbf{Q}_r\mathbf{Q}_r')' \right\} \quad (4.4.12) \\ &= \text{tr} \left\{ (\bar{\mathbf{X}}\mathbf{L} - \bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}_r')(\bar{\mathbf{X}}\mathbf{L} - \bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}_r')' \right\} \\ &+ \text{tr} \left\{ (\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}_r' - \mathbf{P}\mathbf{Q}_r\mathbf{Q}_r')(\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}_r' - \mathbf{P}\mathbf{Q}_r\mathbf{Q}_r')' \right\}. \end{aligned}$$

Hence, according to the Eckart-Young theorem the trace in (4.4.12) is minimised across all $J \times p$ matrices of rank r , $\mathbf{P}\mathbf{Q}_r\mathbf{Q}_r'$, when

$$\mathbf{P} = \bar{\mathbf{X}}\mathbf{L} \text{ and } \mathbf{Q}_r = \mathbf{V}_r^{\mathbf{I}}.$$

It is evident that $\mathcal{V}(\mathbf{V}_r^{\mathbf{I}})$ is the r -dimensional subspace of the p -dimensional canonical space that is closest to the set of J points $\{(\bar{\mathbf{x}}^j)' \mathbf{L}\}$ in that $\mathbf{Q}_r = \mathbf{V}_r^{\mathbf{I}}$ yields the smallest possible value of the trace,

$$\text{tr} \left\{ (\bar{\mathbf{X}}\mathbf{L} - \bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}_r')(\bar{\mathbf{X}}\mathbf{L} - \bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}_r')' \right\}$$

$$= \sum_{j=1}^J \left((\bar{\mathbf{x}}^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{Q}_r \mathbf{Q}_r' \right) \left((\bar{\mathbf{x}}^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{Q}_r \mathbf{Q}_r' \right)' . \quad (4.4.13)$$

That is, the r -dimensional unweighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{I}})$, is that r -dimensional subspace of the p -dimensional canonical space that yields the smallest possible value of the sum of the squared Pythagorean distances between the points representing the group means in the p -dimensional canonical space, that is $\left\{ (\bar{\mathbf{x}}^j)' \mathbf{L} \right\}$, and their orthogonal projections onto the r -dimensional subspace. It follows that the points representing the group means in the r -dimensional unweighted CVA display, that is $\left\{ (\bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{V}_r^{\mathbf{I}} (\mathbf{V}_r^{\mathbf{I}})' \right\}$, are obtained by firstly transforming the group means to the p -dimensional canonical space, that is the p -dimensional space in which the Pythagorean distances are proportional to the corresponding Mahalanobis distances in the p -dimensional measurement space, and then projecting the transformed group means onto the r -dimensional subspace which is closest to the set of J points, $\left\{ (\bar{\mathbf{x}}^j)' \mathbf{L} \right\}$, in terms of least squares. Recall from Section 4.2 that the weighted CVA display is identical to the unweighted CVA display when the J groups are of identical sizes and that the more \mathbf{N} differs from $\frac{n}{J} \mathbf{I}$, the more the CVA display corresponding to $\mathbf{C} = \mathbf{N}$ will differ from the CVA display corresponding to $\mathbf{C} = \mathbf{I}$.

Note that since the centroid of the set of points, $\left\{ (\bar{\mathbf{x}}^j)' \mathbf{L} \right\}$, is not given by the null vector, the Pythagorean distances between the points representing the group means in the p -dimensional canonical space are not optimally approximated by the corresponding Pythagorean distances between the points representing the group means in the r -dimensional subspace, $\mathcal{V}(\mathbf{V}_r)$, that is, the r -dimensional unweighted CVA display space (Section 1.6.11). Note that since $\bar{\mathbf{X}}' \bar{\mathbf{X}}$ and \mathbf{W} do not add up to $\mathbf{X}' \mathbf{X}$, the right singular vectors of $\mathbf{X} \mathbf{L}$ and $\bar{\mathbf{X}} \mathbf{L}$ are not the same. Hence the Pythagorean distance between the points in the p -dimensional canonical space are not optimally approximated by the corresponding Pythagorean distances in the r -dimensional unweighted CVA display space.

Note that the J groups carry identical weights in the minimisation criteria that determines the r -dimensional unweighted CVA display space, that is

$$\sum_{j=1}^J \left((\bar{\mathbf{x}}^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{Q}_r \mathbf{Q}_r' \right) \left((\bar{\mathbf{x}}^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{Q}_r \mathbf{Q}_r' \right)' .$$

Recall that the larger groups carry greater weights than the smaller groups in the minimisation criteria that determines the r -dimensional weighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$. This implies that the r -dimensional weighted CVA display space tends to be more drawn towards the points representing the group means of the larger groups in the p -dimensional canonical space than the unweighted CVA display space. That is, the points representing the group means of larger groups in the p -dimensional canonical space tend to lie closer to their orthogonal projections onto the r -dimensional weighted CVA display space than to their orthogonal projec-

tions onto the r -dimensional unweighted CVA display space, $r \in [1 : K]$. Similarly, the points representing the group means of the smaller groups in the p -dimensional canonical space will tend to lie further from their orthogonal projections onto the r -dimensional weighted CVA display space than from their orthogonal projections onto the r -dimensional unweighted CVA display space, $r \in [1 : K]$. This is however not necessarily the case - if the point representing the group mean of a very small group in the p -dimensional canonical space lies close to the point representing the group mean of the largest group, then the point representing the group mean of the small group will likely lie closer to its projection onto the r -dimensional weighted CVA display space than to its projection onto the r -dimensional unweighted CVA display space. Remember that the K -dimensional weighted and unweighted CVA display spaces are identical. Hence, the Pythagorean distance between any point in the p -dimensional canonical space and its orthogonal projection onto the K -dimensional weighted CVA display space are identical to the Pythagorean distance between that point and its orthogonal projection onto the K -dimensional unweighted CVA display space. Furthermore, since the canonical mean of a group is a measure of the central locality of the canonical samples belonging to that group, it follows that the points representing the individual samples belonging to the larger groups in the p -dimensional canonical space, on average tend to lie closer to their orthogonal projections onto the r -dimensional weighted CVA display than to their orthogonal projections onto the r -dimensional unweighted CVA display space, $r \in [1 : K]$. Similarly, the points representing the individual samples belonging to the smaller groups in the p -dimensional canonical space, on average tend to lie further from their orthogonal projections onto the r -dimensional weighted CVA display than from their orthogonal projections onto the r -dimensional unweighted CVA display space, $r \in [1 : K]$.

Note that the Pythagorean distances between the points $\{\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\}$ are not optimally approximated by the Pythagorean distances between the points representing the group means in either the r -dimensional weighted CVA display or the r -dimensional unweighted CVA display. Note also that the Pythagorean distance between the two points $\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}$ is identical to the Pythagorean distance between the two points $\mathbf{e}'_j (\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_k (\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}$, for all $j, k \in [1 : J]$. Since the matrix $(\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}$ is centred such that $\mathbf{1}' (\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L} = \mathbf{0}'$, the Pythagorean distances between the points,

$$\left\{ \mathbf{e}'_j \left(\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{L} \right\}$$

are optimally approximated by the corresponding Pythagorean distances between the orthogonal projections of these points onto the r -dimensional subspace of $\mathcal{V} \left(\left(\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{L} \right)'$ that is spanned by the first r right singular vectors of $(\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}$ (see Section 1.6.11). Letting the svd of the matrix $(\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}$,

where \mathbf{L} is as defined in (4.4.8), be given by

$$\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L} = \mathbf{U}^{\text{Cent}}\mathbf{D}^{\text{Cent}}\left(\mathbf{V}^{\text{Cent}}\right)'$$

this means that the Pythagorean distances between the points $\{\mathbf{e}'_j\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L}\mathbf{V}^{\text{Cent}}\left(\mathbf{V}^{\text{Cent}}\right)'\}$ will optimally approximate the corresponding Pythagorean distances between the points $\{\mathbf{e}'_j\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L}\}$. Since the Pythagorean distance between the two points $\mathbf{e}'_j\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$ and $\mathbf{e}'_k\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{V}_r\mathbf{V}'_r$ is identical to the Pythagorean distance between the two points $\mathbf{e}'_j\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$ and $\mathbf{e}'_k\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$ for all $j, k \in [1 : J]$, it follows that the Pythagorean distances between the points $\{\mathbf{e}'_j\overline{\mathbf{X}}\mathbf{L}\}$ are optimally approximated by the corresponding Pythagorean distances between the orthogonal projections of the points $\{\mathbf{e}'_j\overline{\mathbf{X}}\mathbf{L}\}$ onto the r -dimensional subspace $\mathcal{V}(\mathbf{V}_r^{\text{Cent}})$, $r \in [1 : K]$. That means that the relative magnitudes of the Pythagorean distances between the points $\{\mathbf{e}'_j\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L}\mathbf{V}^{\text{Cent}}\left(\mathbf{V}^{\text{Cent}}\right)'\}$ will provide the optimal representation of the relative magnitudes of the corresponding Mahalanobis distances between the points $\{\mathbf{e}'_j\overline{\mathbf{X}}\}$.

Note that the r -dimensional subspace $\mathcal{V}(\mathbf{V}_r^{\text{Cent}})$ is the r -dimensional subspace of the row space of $\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L}$ that is closest to the set of points $\{\mathbf{e}'_j\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L}\}$ in terms of least squares, that is the matrix $\mathbf{V}_r^{\text{Cent}}$ minimises the trace

$$\text{tr}\left\{\left(\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L} - \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{Q}_r\mathbf{Q}'_r\right)\left(\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L} - \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\overline{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r\right)'\right\} \quad (4.4.14)$$

across all $J \times r$ orthonormal matrices, \mathbf{Q}_r . It follows that if the r -dimensional CVA display space is taken to be the subspace $\mathcal{V}(\mathbf{V}_r^{\text{Cent}})$ and consequently the j th group mean and i th sample are represented by points with coordinate vectors given by the j th row vector of

$$\overline{\mathbf{X}}\mathbf{M}_r^{\text{Cent}} \text{ where } \mathbf{M}_r^{\text{Cent}} = \mathbf{L}\mathbf{V}_r^{\text{Cent}}$$

and the i th row vector of

$$\mathbf{X}\mathbf{M}_r^{\text{Cent}} = \mathbf{X}\mathbf{L}\mathbf{V}_r^{\text{Cent}}$$

respectively, then the Pythagorean distances between the points representing the

group means in the p -dimensional canonical space will be optimally approximated by the corresponding Pythagorean distances between the points representing the group means in the r -dimensional CVA display space. Since the J groups receive identical weights in the minimisation criteria in (4.4.14), the CVA display constructed from the svd of $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ is referred to as an unweighted CVA display (like the CVA display constructed from the svd of $\bar{\mathbf{X}}\mathbf{L}$).

Recall that the larger groups carry greater weights than the smaller groups in the minimisation criteria that determines the r -dimensional weighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$. It follows that the points representing the group means of larger groups in the p -dimensional canonical space tend to lie closer to their orthogonal projections onto the r -dimensional weighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$, than to their orthogonal projections onto the r -dimensional unweighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{Cent}})$, $r \in [1 : K]$. Similarly, the points representing the group means of the smaller groups in the p -dimensional canonical space will tend to lie further from their orthogonal projections onto the r -dimensional weighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{N}})$, than from their orthogonal projections onto the r -dimensional unweighted CVA display space, $\mathcal{V}(\mathbf{V}_r^{\mathbf{Cent}})$, $r \in [1 : K]$.

The matrix

$$\mathbf{M}^{\mathbf{Cent}} = \mathbf{L}\mathbf{V}^{\mathbf{Cent}}$$

satisfies the two-sided eigenvalue problem

$$\bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{M}^{\mathbf{Cent}} = \mathbf{W}\mathbf{M}^{\mathbf{Cent}}\boldsymbol{\Lambda}_r^{\mathbf{Cent}}. \quad (4.4.15)$$

This is shown below:

$$\begin{aligned} & \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{L} = \mathbf{U}^{\mathbf{Cent}}\mathbf{D}^{\mathbf{Cent}}(\mathbf{V}^{\mathbf{Cent}})' \\ \rightarrow & \mathbf{L}'\bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{L} = \mathbf{V}^{\mathbf{Cent}}(\mathbf{D}_p^{\mathbf{Cent}})^2(\mathbf{V}^{\mathbf{Cent}})' \\ \rightarrow & \bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{L}\mathbf{V} = (\mathbf{L}^{-1})'\mathbf{V}^{\mathbf{Cent}}(\mathbf{D}_p^{\mathbf{Cent}})^2 \\ \rightarrow & \bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{L}\mathbf{V} = (\mathbf{L}^{-1})'\mathbf{L}^{-1}\mathbf{L}\mathbf{V}^{\mathbf{Cent}}(\mathbf{D}_p^{\mathbf{Cent}})^2 \\ \rightarrow & \bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{M}^{\mathbf{Cent}} = \mathbf{W}\mathbf{M}^{\mathbf{Cent}}\boldsymbol{\Lambda}^{\mathbf{Cent}} \end{aligned}$$

where $\mathbf{M}^{\mathbf{Cent}} = \mathbf{L}\mathbf{V}^{\mathbf{Cent}}$ and $\boldsymbol{\Lambda} = (\mathbf{D}_p^{\mathbf{Cent}})^2$. The two-sided eigenvalue problem in (4.4.15) has $p - K$ zero eigenvalues since the matrix $\mathbf{W}^{-1/2}\bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}}\mathbf{W}^{-1/2}$ is

of rank $K = \min(J - 1, p)$:

$$\begin{aligned}
 \text{rank}\left(\mathbf{W}^{-1/2}\bar{\mathbf{X}}'\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\mathbf{W}^{-1/2}\right) &= \text{rank}\left(\bar{\mathbf{X}}'\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\right) \\
 &= \text{rank}\left(\bar{\mathbf{X}}'\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\right) \\
 &= \text{rank}\left(\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\right) \\
 &= \min(J - 1, p) .
 \end{aligned}$$

The last step follows since the matrix $\bar{\mathbf{X}}$ is of rank $K = \min(J - 1, p)$ and because pre-multiplication by $\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)$ simply changes the one linear dependence between the row vectors of $\bar{\mathbf{X}}$, that is

$$\sum_{j=1}^J n_j (\bar{\mathbf{x}}^j)' = \mathbf{0}'$$

to the following linear dependence between the row vectors of $\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}$:

$$\sum_{j=1}^J \mathbf{e}_j' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}} = \mathbf{0}' .$$

It can be shown that the last $p - K$ coordinates of the J points $\left\{(\bar{\mathbf{x}}^j)'\mathbf{M}\right\}$ are identical. A derivation of this result is provided in the appendix at the end of this chapter. Consequently the classifications based on the K -dimensional CVA display constructed from the svd of $\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\mathbf{L}$ are identical to the classifications based on the p -dimensional canonical space. This means that the classifications based on the K -dimensional CVA displays constructed from the svd of $\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\mathbf{L}$, $\bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$, are identical.

It will now be shown that the K -dimensional CVA display space, $\mathcal{V}(\mathbf{V}_K^{\text{Cent}})$ is identical to the row space of $\bar{\mathbf{X}}\mathbf{L}$, just like the K -dimensional CVA display space $\mathcal{V}(\mathbf{V}_K^{\mathbf{I}})$ and the K -dimensional CVA display space $\mathcal{V}(\mathbf{V}_K^{\mathbf{N}})$. Consider the $J \times p$ matrix $\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\mathbf{L}$. When $p < J - 1$, this matrix is of rank p while when $p > J - 1$ the matrix is of rank $J - 1$. Let the svd's of $\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\mathbf{L}$, $\bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ be given by:

$$\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\mathbf{L} = \mathbf{U}^{\text{Cent}}\mathbf{D}^{\text{Cent}}(\mathbf{V}^{\text{Cent}})'$$

$$\begin{aligned}\bar{\mathbf{X}}\mathbf{L} &= \mathbf{U}^{\mathbf{I}}\mathbf{D}^{\mathbf{I}}(\mathbf{V}^{\mathbf{I}})' \\ \mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L} &= \mathbf{U}^{\mathbf{N}}\mathbf{D}^{\mathbf{N}}(\mathbf{V}^{\mathbf{N}})'. \end{aligned}$$

When $p < J - 1$ and hence the rank of each of the matrices, $\bar{\mathbf{X}}\mathbf{L}$, $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ is equal to $K = p$, it follows that

$$\mathbf{V}_K^{\mathbf{I}}(\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\text{Cent}}(\mathbf{V}_K^{\text{Cent}})' = \mathbf{V}_K^{\mathbf{N}}(\mathbf{V}_K^{\mathbf{N}})' = \mathbf{I}_K.$$

Since the three projection matrices above are identical, it follows that the column space of $\mathbf{V}^{\mathbf{I}}$, \mathbf{V}^{Cent} and $\mathbf{V}^{\mathbf{N}}$ are identical. It is known that

$$\begin{aligned}\nu(\mathbf{V}^{\mathbf{I}}) &= \nu((\bar{\mathbf{X}}\mathbf{L})') \\ \nu(\mathbf{V}^{\text{Cent}}) &= \nu\left(\left(\left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\bar{\mathbf{X}}\mathbf{L}\right)'\right) \\ \nu(\mathbf{V}^{\mathbf{N}}) &= \nu\left(\left(\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}\right)'\right). \end{aligned}$$

It follows that the row space of $\bar{\mathbf{X}}\mathbf{L}$, $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ are identical and hence that the p column vectors of each of the matrices, $\mathbf{V}^{\mathbf{I}}$, \mathbf{V}^{Cent} and $\mathbf{V}^{\mathbf{N}}$, form an orthogonal basis for the row space of $\bar{\mathbf{X}}\mathbf{L}$.

Consider now the case where $p > J - 1$ and hence where the rank of each of the matrices, $\bar{\mathbf{X}}\mathbf{L}$, $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ is equal to $K = J - 1$. Since elementary row operations do not change the row space of a matrix, the row space of $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ is identical to the row space of the matrix obtained by subtracting from each of the first $J - 1$ rows of $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ the J th row of $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$, that is the matrix

$$\begin{bmatrix} (\bar{\mathbf{x}}^1)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} - \left((\bar{\mathbf{x}}^J)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} \right) \\ (\bar{\mathbf{x}}^2)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} - \left((\bar{\mathbf{x}}^J)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} \right) \\ \vdots \\ (\bar{\mathbf{x}}^{J-1})' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} - \left((\bar{\mathbf{x}}^J)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} \right) \\ (\bar{\mathbf{x}}^J)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} \end{bmatrix} = \begin{bmatrix} (\bar{\mathbf{x}}^1)' \mathbf{L} - (\bar{\mathbf{x}}^J)' \mathbf{L} \\ (\bar{\mathbf{x}}^2)' \mathbf{L} - (\bar{\mathbf{x}}^J)' \mathbf{L} \\ \vdots \\ (\bar{\mathbf{x}}^{J-1})' \mathbf{L} - (\bar{\mathbf{x}}^J)' \mathbf{L} \\ (\bar{\mathbf{x}}^J)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} \end{bmatrix}. \quad (4.4.16)$$

Note that the J th row vector of the above matrix can be expressed as a linear combination of the other $J - 1$ row vectors

$$(\bar{\mathbf{x}}^J)' \mathbf{L} - \frac{1}{J} \mathbf{1}' \bar{\mathbf{X}}\mathbf{L} = -\frac{1}{J} \left((\bar{\mathbf{x}}^1)' \mathbf{L} - (\bar{\mathbf{x}}^J)' \mathbf{L} + (\bar{\mathbf{x}}^2)' \mathbf{L} - (\bar{\mathbf{x}}^J)' \mathbf{L} + \dots + (\bar{\mathbf{x}}^{J-1})' \mathbf{L} - (\bar{\mathbf{x}}^J)' \mathbf{L} \right).$$

Note that since the matrix $\mathbf{M}^{\text{Cent}} = \mathbf{L}\mathbf{V}^{\text{Cent}}$ satisfies the two-sided eigenvalue problem in (4.4.15), the matrix $\mathbf{M}_r^{\text{Cent}} = \mathbf{L}\mathbf{V}_r^{\text{Cent}}$ maximises the trace

$$\text{tr} \left\{ \left(\mathbf{F}'_r \bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}' \right) \bar{\mathbf{X}} \mathbf{F}_r \right) (\mathbf{F}'_r \mathbf{W} \mathbf{F}_r)^{-1} \right\} \quad (4.4.17)$$

across all $p \times r$ matrices of full column rank, \mathbf{F}_r , $r \in [1 : K]$. It is evident that the points representing the group means in the r -dimensional space in which the groups are optimally separated in that the trace in (4.4.17) is maximised across all $p \times r$ matrices of full column rank, \mathbf{F}_r , can be obtained by first transforming the group means and individual observations to the p -dimensional canonical space and then projecting the points orthogonally onto the best fitting r -dimensional subspace to the set of points $\{\mathbf{e}'_j (\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}') \bar{\mathbf{X}} \mathbf{L}\}$ in terms of least squares. It follows that if the r -dimensional CVA display space is taken to be the subspace $\mathcal{V}(\mathbf{V}_r^{\text{Cent}})$ then the groups will be maximally separated in that the trace in (4.4.17) is maximised and the Pythagorean distances between the points representing the group means in the p -dimensional canonical space will be optimally approximated by the corresponding Pythagorean distances between the points representing the group means in the r -dimensional CVA display space.

Let \mathbf{C} denote a $J \times J$ matrix which is either positive definite and symmetric or idempotent and let $\mathbf{C}^{1/2}$ denote the square root matrix of \mathbf{C} if \mathbf{C} is positive definite and symmetric (see Chapter 1) and let $\mathbf{C}^{1/2}$ denote \mathbf{C} when \mathbf{C} is idempotent. It is evident that the r -dimensional CVA display space is the r -dimensional column space of the matrix that yields the smallest possible value of the trace,

$$\text{tr} \left\{ (\mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} - \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{Q}_r \mathbf{Q}'_r) (\mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} - \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{Q}_r \mathbf{Q}'_r)' \right\}$$

where \mathbf{L} is a $p \times p$ non-singular matrix which is such that $\mathbf{L}\mathbf{L}' = \mathbf{W}^{-1}$, across all $p \times r$ orthonormal matrices, \mathbf{Q}_r . It follows from the Eckart-Young theorem that the r -dimensional CVA display space is the subspace $\mathcal{V}(\mathbf{V}_r)$, where \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$. In the r -dimensional CVA display, the j th group mean and i th individual sample are represented by the points,

$$(\bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{V}_r = (\bar{\mathbf{x}}^j)' \mathbf{M}_r$$

and

$$\mathbf{x}'_i \mathbf{L} \mathbf{V}_r = \mathbf{x}'_i \mathbf{M}_r .$$

The matrix $\mathbf{M}_r = \mathbf{L}\mathbf{V}_r$ satisfies

$$\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \mathbf{M}_r = \mathbf{W} \mathbf{M}_r \mathbf{\Lambda}_r .$$

and is the $p \times r$ matrix of full column rank that yields the largest possible value of the ratio

$$\text{tr} \left\{ \mathbf{F}_r' \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \mathbf{F}_r (\mathbf{F}_r' \mathbf{W} \mathbf{F}_r)^{-1} \right\} \quad (4.4.18)$$

across all $p \times r$ matrices of full column rank, \mathbf{F}_r . That is, the r -dimensional CVA display space is that subspace in which the groups are maximally separated in that it yields the maximum value of the trace in (4.4.18).

As an example of an unweighted CVA display constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}')$, consider the two-dimensional CVA display of the simulated data set introduced in Section 4.2.1.1 provided in Figure 4.3. Note that this unweighted CVA display is almost indistinguishable from the unweighted CVA display of this data set constructed with $\mathbf{C} = \mathbf{I}$ in Figure 4.2.

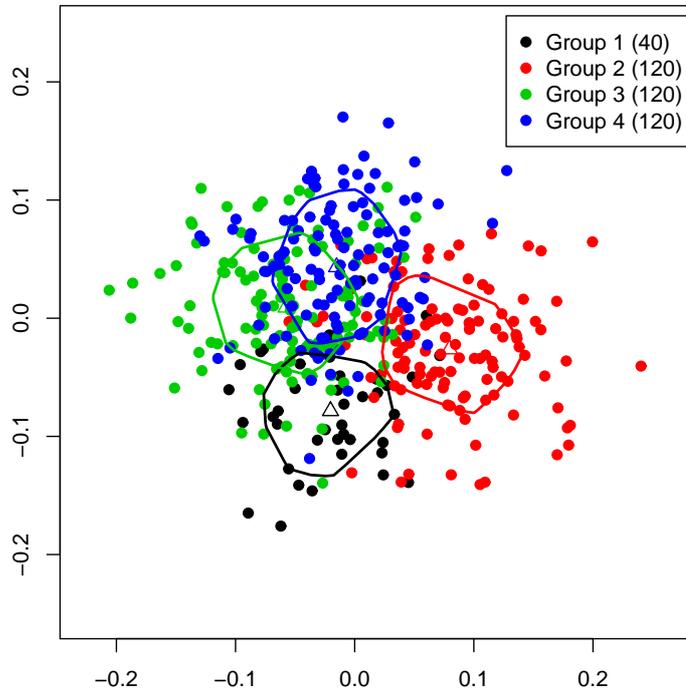


Figure 4.3: *The two-dimensional unweighted CVA display of the simulated data set constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}')$.*

Which of $\mathbf{C} = \mathbf{N}$, $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$ should be used in the construction of the CVA display depends for what purpose the investigator of the CVA display wants to use it. If the investigator merely wants to visualise the most accurate representation of the relative magnitudes of the Mahalanobis distances between the group centroids, then constructing the CVA display from the svd of $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ would be best. If however the investigator wants to visualise the most accurate representation of the relative magnitudes of the Mahalanobis distances between the individual samples, then the CVA display should be constructed from the svd of $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$. If the investigator wants to use the classification regions of the CVA display to classify a set of new samples and he/she expects the relative group sizes associated with the set of new samples to be very similar to those associated with the set of samples from which $\bar{\mathbf{X}}$ is calculated, that is the set of samples upon which the construction of the classification regions were based, then constructing the CVA display from the svd of $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ would probably be best. If on the other hand the investigator believes that the relative group sizes associated with the set of new samples differ substantially from those associated with the set of samples from which $\bar{\mathbf{X}}$ is calculated, then constructing the CVA display from the svd of $\bar{\mathbf{X}}\mathbf{L}$ or $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$ would probably be best.

4.5 The CVA biplot

The CVA displays that have been discussed up to now were ordinary MDS displays, containing no information on the measured variables. The CVA biplot is obtained by adding information on the measured variables to the CVA display in the form of calibrated linear axes called biplot axes.

The construction of the CVA biplot is easiest to understand in the light of the two-step approach discussed in Section 4.4. The CVA biplot can be constructed from the svd of $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$, $\bar{\mathbf{X}}\mathbf{L}$ or $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$, just as the CVA displays discussed in the foregoing sections. The CVA biplot constructed from the svd of $\mathbf{N}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ is called the weighted CVA biplot, while both the CVA biplot constructed from the svd of $\bar{\mathbf{X}}\mathbf{L}$ and the CVA biplot constructed from the svd of $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$, are referred to as unweighted CVA biplots.

Gabriel (1972) introduced the (weighted) CVA biplot, which he referred to as the MANOVA biplot, as the simultaneous graphical representation of the group means and the original measured variables in a space in which the Pythagorean distances between the pairs of transformed group means approximate a constant multiple of the Mahalanobis distances between the corresponding pairs of group means. Gabriel (1972) proposed that the group means and measured variables be represented by sets of uncalibrated vectors emanating from the origin. He furthermore suggested that these vectors be such that the inner product between the vector representing the j th group mean and the vector representing the i th measured variable equals the approximation of the i th element of the j th group mean, \bar{x}_i^j , the approximation

being given by the ij th element of the matrix,

$$\overline{\mathbf{X}}\mathbf{M}_r\mathbf{M}_r^r$$

where r denotes the dimension of the biplot and \mathbf{M}_r satisfies

$$\overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}}\mathbf{M}_r = \mathbf{W}\mathbf{M}_r\mathbf{\Lambda}$$

$i, j \in [1 : J]$, $r \in [1 : p]$. Gower and Hand (1996) adjusted the CVA biplot proposed by Gabriel in the same way in which they adjusted the traditional PCA biplot. They proposed that the group means be represented by the endpoints of the vectors representing them in the biplot proposed by Gabriel and that the measured variables be represented by calibrated linear axes lying collinear with the vectors representing the measured variables in the biplot proposed by Gabriel. Gower and Hand (1996) proposed that these axes, referred to as biplot axes, be such that the approximation to \overline{x}_i^j , that is $(\overline{\mathbf{x}}^j)'\mathbf{M}_r\mathbf{M}_r^r\mathbf{e}_i$, is read off from the biplot axis representing the i th measured variable at the orthogonal projection of the point representing the j th group mean onto the biplot axis representing the i th measured variable, $i \in [1 : p]$, $j \in [1 : J]$. Note that this CVA biplot proposed by Gower and Hand (1996) is called a predictive CVA biplot. Imposing such calibrated linear axes onto the unweighted CVA displays discussed in Sections 4.2 and 4.4 yields the corresponding unweighted predictive CVA biplots. In general, the i th predictive biplot axis of the CVA biplot constructed from the svd of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ is calibrated with the value $(\overline{\mathbf{x}}^j)'\mathbf{M}_r\mathbf{M}_r^r$, where \mathbf{M} satisfies

$$\overline{\mathbf{X}}'\mathbf{C}\overline{\mathbf{X}}\mathbf{M}_r = \mathbf{W}\mathbf{M}_r\mathbf{\Lambda}$$

at the orthogonal projection of the point $(\overline{\mathbf{x}}^j)'\mathbf{M}$ onto that biplot axis. Gower and Hand (1996) also proposed another CVA biplot, called the interpolative CVA biplot in which the measured variables are represented by calibrated linear axes which are used to position the group means and individual samples in the CVA biplot space. The CVA biplot space will henceforth be denoted by \mathcal{L} . The interpolative and predictive CVA biplots will be discussed in more detail in Sections 4.5.1 and 4.5.2.

Like the CVA displays discussed earlier, the main strength of the CVA biplot is its ability to exhibit patterns in the data which might have been very difficult to detect using traditional analytic methods. The CVA biplot allows visual appraisal of the approximate nature and degree of overlap amongst the various groups. Just like the K -dimensional CVA display, the K -dimensional CVA biplot contains all the information useful for discrimination and classification purposes. Classification

regions can be constructed on the CVA biplots proposed by both Gabriel (1971) and Gower and Hand (1996) in exactly the same way as on the CVA displays discussed in Sections 4.2 and 4.4. Imposing classification regions onto the CVA biplot allows visualisation of the approximate CVA classification rules if the dimension of the CVA biplot, r , is less than K and the exact CVA classification rules if $r = K$.

When it comes to the visualisation of the (approximate) CVA classification process via the constructed classification regions, the biplot proposed by Gower and Hand (1996) has the additional advantage that it provides information about the contributions of the different measured variables to the classification of an observation of unknown origin via the linear biplot axes (specifically, the predictive biplot axes discussed in Section 4.5.2). Conclusions about a measured variable can however only be made from the visual inspection of the CVA biplot if the corresponding biplot axis accurately predicts the measurements on that variable. For example, when two observations of unknown origin are classified to two different groups, projection of the two points onto each of the p biplot axes and comparing the values read off from the biplot axes at the projections of the two points, will indicate with respect to which measured variables the two observations differ more than others, given that the biplot axes accurately predicts the measurements of the corresponding variables. If the values read off from the biplot axes are very similar for two observations, except for the values read off from the k th biplot axis and all the biplot axes accurately predicts the measurements of the corresponding variables, then it can be concluded that it is the two observations' large difference with respect to the k th measured variable that resulted in the them being classified to different groups. Projecting the points representing the J groups centroids onto the p biplot axes will also indicate with respect to which variables the groups differ substantially and with respect to which they are quite similar, provided that the biplot axes accurately predicts the corresponding measurements. The quality of the individual CVA biplot axes with respect to their predictive abilities will be discussed in Sections 5.4 and 5.13.1.

If in addition to the group centroids, the individual samples are represented in the CVA biplot, the biplot allows for a more detailed representation of the approximate degree and nature of the overlap and/or separation amongst the groups as well as visualisation of the within-group dispersion. Imposing α -bags onto the biplot further facilitates the visualisation of the extent of the overlap between the groups. The smallest value of α for which there is any overlap between the groups can be used to quantify the amount of overlap. The smaller this α value, the greater the extent of the overlap between the groups. Recall from Section 4.2.5 that the value of Wilk's lambda calculated under the null hypothesis of no difference between the J group means can also be used as a measure of the amount of overlap amongst the groups. The greater the magnitude of Wilk's lambda under the null hypothesis of no difference between the group means, the more evidence there is against the null hypothesis of no difference, that is, the smaller the amount of overlap between the groups.

Like the CVA display, the CVA biplot should only be used to represent the group structure of a data set if the assumption of identical within-group covariance matrices is appropriate for that data set. If this assumption is not appropriate, the group structure underlying the data set can be represented in an AOD biplot.

The definition of the terms interpolation and prediction are the same as for the PCA biplot - interpolation is the process of finding the position of a sample in the biplot space, \mathcal{L} , given the sample's measurements on the original measured variables while prediction is the inverse process, that is, the process of inferring the measurements of a sample on the original measured variables given the sample's position in the biplot space. As for the PCA biplot, each of these two processes are performed by relating the given values to a set of biplot axes. The biplot axes used for interpolation are called interpolative biplot axes and the CVA biplot containing these axes is called the interpolative CVA biplot. On the other hand, the biplot axes used to perform prediction are called predictive biplot axes and the CVA biplot containing these axes is called the predictive CVA biplot. The interpolation and prediction processes corresponding to the CVA biplot are discussed in Sections 4.5.1 and 4.5.2 respectively

4.5.1 Interpolation

The coordinate vector of the point that represents an observation \mathbf{x} in the p -dimensional canonical space is given by:

$$\begin{aligned} \mathbf{y}' &= \mathbf{x}'\mathbf{M} \\ \longrightarrow \mathbf{y}' &= \sum_{k=1}^p x_k \mathbf{e}'_k \mathbf{M} \\ \longrightarrow \mathbf{y}' &= p \left(\frac{1}{p} \sum_{k=1}^p x_k \mathbf{e}'_k \mathbf{M} \right). \end{aligned} \quad (4.5.1)$$

Note that $\mathbf{e}'_k \mathbf{M}$ is the interpolant of one unit of the k th variable, x_k , in the p -dimensional canonical space. Hence, in the p -dimensional interpolative CVA biplot, the point $\mathbf{e}'_k \mathbf{M}$ will be calibrated with one unit of the k th variable. Similarly, the point $x_k \mathbf{e}'_k \mathbf{M}$ will be calibrated x_k units of x_k . It is evident that the biplot axis representing the x_k in the p -dimensional interpolative CVA biplot is linear and lies in the direction of the vector, $\mathbf{e}'_k \mathbf{M}$, $k \in [1 : p]$. It is evident that the interpolant of \mathbf{x} in the full canonical space is the weighted vector-sum of the interpolants of the unit points of the p x -variables where the weights are given by the corresponding elements of \mathbf{x} . The expression of the interpolant of \mathbf{x} in (4.5.1) shows that if the vector emanating from the origin to the centroid of the set of points, $\{x_k \mathbf{e}'_k \mathbf{M}\}$, is extended p times, then the endpoint of that vector gives the position of the interpolant of \mathbf{x} .

The interpolant of \mathbf{x} in the r -dimensional subspace of the canonical space defined by the first r canonical variables is given by

$$\begin{aligned} \mathbf{z}' &= \mathbf{x}'\mathbf{M}_r \\ &= \sum_{k=1}^p x_k \mathbf{e}'_k \mathbf{M}_r. \end{aligned}$$

By the same argument as before, the interpolative axis representing x_k in the r -dimensional interpolative CVA biplot is linear and lies in the direction of the vector $\mathbf{e}'_k \mathbf{M}_r$ with the point $x_k \mathbf{e}'_k \mathbf{M}_r$ being calibrated with x_k units of x_k . It follows that the interpolant of \mathbf{x} in the r -dimensional CVA biplot space is positioned at the endpoint of the vector obtained by extending the vector emanating from the origin to the centroid of the set of points, $\{x_k \mathbf{e}'_k \mathbf{M}_r\}$, p times.

In order to obtain the position of a new sample (that is a sample that was not part of the set of samples from which $\bar{\mathbf{X}}$ was calculated) it first needs to be scaled such that its measurements are in the same scales as those of the centred matrix \mathbf{X} . That is, the overall (unconditional) observed means of the measured variables must be subtracted from the corresponding measurements of the new sample. After rescaling the measurements of the new sample, the position of the sample in the existing biplot can be obtained in exactly the same way as that of an original sample. That is, if \mathbf{x}^* is the rescaled new sample, then its position in the existing biplot (constructed from $\bar{\mathbf{X}}$) is given by $(\mathbf{z}^*)' = (\mathbf{x}^*)' \mathbf{M}_r$.

4.5.2 Prediction

Since prediction is the inverse process of interpolation, the predicted values corresponding to a point \mathbf{y} in the p -dimensional canonical space is given by

$$\hat{\mathbf{x}}' = \mathbf{y}' \mathbf{M}^{-1}.$$

When the point, \mathbf{y} which lies in the p -dimensional canonical space, also lies in the r -dimensional biplot space, only its first r elements can be non-zero values. Hence \mathbf{y} can be expressed in the following form:

$$\mathbf{y}' = [\mathbf{z}' \mathbf{0}']$$

where the i th element of \mathbf{z} is identical to the i th element of \mathbf{y} , $i \in [1 : r]$. It follows that the predicted values corresponding to a point \mathbf{z} in \mathcal{L} are given by:

$$\begin{aligned} \hat{\mathbf{x}}' &= \mathbf{y}' \mathbf{M}^{-1} \\ &= [\mathbf{z}' \mathbf{0}'] \mathbf{M}^{-1} \\ &= \mathbf{z}' \mathbf{M}^r. \end{aligned}$$

A point that predicts the value μ for the k th variable, x_k , therefore satisfies

$$\mu = \hat{\mathbf{x}}' \mathbf{e}_k$$

$$\longrightarrow \mu = \mathbf{z}'\mathbf{M}^r\mathbf{e}_k. \quad (4.5.2)$$

Hence, all the points in \mathcal{L} that predict the value μ for x_k lie on a $(r-1)$ -dimensional hyperplane in \mathcal{L} which lies orthogonal to the vector $\mathbf{M}^r\mathbf{e}_k$. The hyperplanes corresponding to different values of μ lie parallel to each other and are all orthogonal to the vector $\mathbf{M}^r\mathbf{e}_k$. The value of x_k predicted by a point \mathbf{z} in \mathcal{L} , is therefore the value of μ which is such that \mathbf{z} lies on the hyperplane,

$$\mu = \mathbf{z}'\mathbf{M}^r\mathbf{e}_k.$$

If a line passing through the origin and orthogonal to these hyperplanes is constructed and calibrated with the value μ at the point where it intersects with the hyperplane $\mu = \mathbf{z}'\mathbf{M}^r\mathbf{e}_k$, then the predicted value for x_k corresponding to a point \mathbf{z}^* in \mathcal{L} , can be obtained by projecting the point \mathbf{z}^* orthogonally onto the line and reading off the calibration at the point onto which \mathbf{z}^* projects. Every point \mathbf{z} on the line passing through the origin and orthogonal to the hyperplanes of the form, $\mu = \mathbf{z}'\mathbf{M}^r\mathbf{e}_k$, can be expressed in the following way:

$$\mathbf{z}' = \sigma\mathbf{e}'_k(\mathbf{M}^r)'$$

If $\sigma\mathbf{e}'_k(\mathbf{M}^r)'$ is substituted for \mathbf{z}' in equation (4.5.2), then the following expression for the predicted value μ for x_k is obtained:

$$\mu = \sigma\mathbf{e}'_k(\mathbf{M}^r)'\mathbf{M}^r\mathbf{e}_k.$$

The value of σ corresponding to the value of μ , is therefore given by

$$\sigma = \frac{\mu}{\mathbf{e}'_k(\mathbf{M}^r)'\mathbf{M}^r\mathbf{e}_k}$$

and hence

$$\begin{aligned} \mathbf{z}' &= \sigma\mathbf{e}'_k(\mathbf{M}^r)' \\ \longrightarrow \mathbf{z}' &= \frac{\mu}{\mathbf{e}'_k(\mathbf{M}^r)'\mathbf{M}^r\mathbf{e}_k}\mathbf{e}'_k(\mathbf{M}^r)'. \end{aligned} \quad (4.5.3)$$

The line defined by points satisfying equation (4.5.3) and which is calibrated as explained above is called the k th predictive biplot axis of the CVA biplot. It is evident from equation (4.5.3) that the predictive biplot axis representing the k th measured variable (or the k th predictive biplot axis for short) lies in the direction of the vector, $\mathbf{e}'_k (\mathbf{M}^r)'$, $k \in [1 : p]$. A CVA biplot with predictive biplot axes is called a predictive CVA biplot. It is important to note that the k th predictive biplot axis of the CVA biplot does not lie in the same direction as the k th interpolative biplot axis of the CVA biplot - the CVA biplot differs from the PCA biplot in this respect. In the rest of this thesis, it will be assumed that if the type of CVA biplot is not specified, it is the predictive CVA biplot being referred to.

Note that since

$$(\mathbf{M}\mathbf{M}')^{-1} = \mathbf{W}$$

the cosine of the angle between the i th and j th predictive biplot axes of the p -dimensional CVA biplot, θ_{ij} , is equal to the correlation coefficient between the i th and j th measured variables:

$$\begin{aligned} \cos(\theta_{ij}) &= \frac{\mathbf{e}'_i (\mathbf{M}^{-1})' \mathbf{M}^{-1} \mathbf{e}_j}{\sqrt{\mathbf{e}'_i (\mathbf{M}^{-1})' \mathbf{M}^{-1} \mathbf{e}_i} \sqrt{\mathbf{e}'_j (\mathbf{M}^{-1})' \mathbf{M}^{-1} \mathbf{e}_j}} \\ &= \frac{[\mathbf{W}]_{ij}}{\sqrt{[\mathbf{W}]_{ii} [\mathbf{W}]_{jj}}} \\ &= \frac{[\widehat{\Sigma}_{\mathbf{W}}]_{ij}}{\sqrt{[\widehat{\Sigma}_{\mathbf{W}}]_{ii} [\widehat{\Sigma}_{\mathbf{W}}]_{jj}}} \\ \rightarrow \cos(\theta_{ij}) &= r_{ij}. \end{aligned}$$

It follows that the correlation coefficient between two measured variables is a decreasing function of the size of the angle between the two predictive biplot axes representing those two variables in the p -dimensional CVA biplot space. The cosine of the angle between the i th and j th predictive biplot axes of the r -dimensional CVA biplot,

$$\frac{\mathbf{e}'_i (\mathbf{M}^r)' \mathbf{M}^r \mathbf{e}_j}{\sqrt{\mathbf{e}'_i (\mathbf{M}^r)' \mathbf{M}^r \mathbf{e}_i} \sqrt{\mathbf{e}'_j (\mathbf{M}^r)' \mathbf{M}^r \mathbf{e}_j}},$$

approximates the correlation coefficient since the elements of the matrix $(\mathbf{M}^r)' \mathbf{M}^r$ approximate the elements of the matrix \mathbf{W} (although it is not clear what the mini-

mization criteria is).

Consider again the expression of the predicted values corresponding to a point \mathbf{z} in \mathcal{L} :

$$\hat{\mathbf{x}}' = \mathbf{z}'\mathbf{M}^r .$$

From Section 4.5.1, it is known that

$$\mathbf{z}' = \mathbf{x}'\mathbf{M}_r .$$

The expression for $\hat{\mathbf{x}}$ can therefore be rewritten as

$$\hat{\mathbf{x}}' = \mathbf{x}'\mathbf{M}_r\mathbf{M}^r .$$

The approximations to the matrix of group means, $\overline{\mathbf{X}}$, and the matrix of individual observations, \mathbf{X} , which are produced by the predictive CVA biplot are therefore given by

$$\begin{aligned} \widehat{\overline{\mathbf{X}}} &= \overline{\mathbf{X}}\mathbf{M}_r\mathbf{M}^r \\ \text{and } \widehat{\mathbf{X}} &= \mathbf{X}\mathbf{M}_r\mathbf{M}^r \end{aligned}$$

respectively. It is evident that each of the J group means as well as each of the n individual samples will be perfectly predicted in the p -dimensional CVA biplot irrespective of the \mathbf{C} -matrix used in the construction of the biplot. Note that since each of the last $p - K$ singular values of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ is equal to zero, the vector defining the i th dimension of the CVA biplot, where $i > K$, is not unique. The $(K + j)$ th dimension of the CVA biplot space can be defined by any one of the right singular vectors of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ which lies orthogonal to the $(K + j - 1)$ -dimensional CVA biplot space, $j \in [1 : p - K]$. An algorithm describing how the second dimension of a two-dimensional CVA biplot of a data set consisting of two groups (i.e. $K = 1$) can be defined, can be found in Gardner and Le Roux (2004). This algorithm can be generalised for the case where $K > 1$. This is however not in the scope of this thesis.

The approximation to a new sample that was interpolated into the biplot can be obtained in exactly the same way as that of an original sample - that is, after it has been rescaled as explained in Section 4.5.1. That is, if \mathbf{x}^* is a new sample, rescaled such that its measurements are in the same scales as the measurements in \mathbf{X} , then

the r -dimensional CVA biplot will yield the following approximation to \mathbf{x}^* :

$$(\hat{\mathbf{x}}^*)' = \mathbf{x}'\mathbf{M}_r\mathbf{M}'_r.$$

It is evident that any new sample will be perfectly predicted by the p -dimensional CVA biplot, irrespective of which \mathbf{C} -matrix is used in the construction of the biplot.

Recall that the r -dimensional CVA biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$ where \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$, is the column space of that matrix for which

$$\text{tr} \left\{ \left(\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r \right) \left(\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r \right)' \right\}$$

is minimised across all $p \times r$ orthonormal matrices, \mathbf{Q}_r . Letting $\widehat{\mathbf{X}}^\# = \bar{\mathbf{X}}\mathbf{L}\mathbf{Q}_r\mathbf{Q}'_r\mathbf{L}^{-1}$, it is evident that the r -dimensional CVA biplot space, \mathcal{L} , is that r -dimensional subspace for which

$$\text{tr} \left\{ \left(\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2}\widehat{\mathbf{X}}^\#\mathbf{L} \right) \left(\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2}\widehat{\mathbf{X}}^\#\mathbf{L} \right)' \right\} \quad (4.5.4)$$

is minimised across all $p \times r$ orthonormal matrices, \mathbf{Q}_r . The best rank r approximation to the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$, that is $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$, where \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$, can be expressed as $\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}$ where

$$\widehat{\mathbf{X}} = \bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}'_r.$$

This is shown below:

$$\begin{aligned} \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r &= \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r\mathbf{L}^{-1}\mathbf{L} \\ \longrightarrow \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r &= \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}'_r\mathbf{L} \\ \longrightarrow \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r &= \mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L} \text{ where } \widehat{\mathbf{X}} = \bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}'_r. \end{aligned}$$

The approximations to the group means which are read off from the predictive biplot axes of the r -dimensional predictive CVA biplot, that is the rows of $\widehat{\mathbf{X}} = \bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}'_r$, therefore yields the smallest possible value of the trace in (4.5.4).

Since the rank of the matrix $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ is equal to $K = \min(J-1, p)$, it follows that

$$\begin{aligned}\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L} &= \mathbf{U}_K\mathbf{D}_K\mathbf{V}'_K \\ &= \mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}\mathbf{V}_K\mathbf{V}'_K \\ &= \mathbf{C}^{1/2}\widehat{\mathbf{X}}_K\mathbf{L} \\ \text{where } \widehat{\mathbf{X}}_K &= \overline{\mathbf{X}}\mathbf{M}_K\mathbf{M}^K.\end{aligned}$$

When $\mathbf{C} = \mathbf{I}$ and $\mathbf{C} = \mathbf{N}$, $\mathbf{C}^{1/2}$ is a non-singular matrix and hence

$$\overline{\mathbf{X}} = \widehat{\mathbf{X}}_K = \overline{\mathbf{X}}\mathbf{M}_K\mathbf{M}^K.$$

Recall that for each of $\mathbf{C} = \mathbf{I}$, $\mathbf{C} = \mathbf{N}$ and $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$,

$$\overline{\mathbf{X}}\mathbf{m}_{(j)} = \mathbf{0} \quad \forall j \in [K+1 : p]. \quad (4.5.5)$$

Consider the following expression of the matrix $\overline{\mathbf{X}}$:

$$\begin{aligned}\overline{\mathbf{X}} &= \overline{\mathbf{X}}\mathbf{M}\mathbf{M}^{-1} \\ \longrightarrow \overline{\mathbf{X}} &= \overline{\mathbf{X}}\mathbf{M}_K\mathbf{M}^K + \overline{\mathbf{X}}\mathbf{M}_{(p-K)}\mathbf{M}^{(p-K)}\end{aligned}$$

where $\mathbf{M}_{(p-K)}$ denotes the submatrix of \mathbf{M} consisting of the last $p-K$ columns of \mathbf{M} and $\mathbf{M}^{(p-K)}$ denotes the submatrix of \mathbf{M}^{-1} consisting of the last $p-K$ rows of \mathbf{M}^{-1} . Given equation (4.5.5) it follows that

$$\overline{\mathbf{X}} = \overline{\mathbf{X}}\mathbf{M}_K\mathbf{M}^K.$$

It follows that the K -dimensional CVA biplot perfectly predicts each of the J group means irrespective of which one of the \mathbf{C} -matrices was used in the construction of the biplot. It is important to note that the first K right singular vectors of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ do not span the row space of $\mathbf{X}\mathbf{L}$. Hence the K -dimensional CVA biplot will not perfectly predict all n the individual samples.

Below, the expression of $\widehat{\mathbf{X}}$ is rewritten in such a way that the geometric inter-

pretation of $\widehat{\mathbf{X}}$ is highlighted:

$$\begin{aligned}\widehat{\mathbf{X}} &= \overline{\mathbf{X}}\mathbf{M}^r\mathbf{M}^r \\ &= \overline{\mathbf{X}}\mathbf{W}^{-1}(\mathbf{M}^r)'\mathbf{M}^r \\ &= \overline{\mathbf{X}}\mathbf{W}^{-1}(\mathbf{M}^r)'(\mathbf{M}^r\mathbf{W}^{-1}(\mathbf{M}^r)')^{-1}\mathbf{M}^r \\ \rightarrow \widehat{\mathbf{X}}' &= (\mathbf{M}^r)'(\mathbf{M}^r\mathbf{W}^{-1}(\mathbf{M}^r)')^{-1}\mathbf{M}^r\mathbf{W}^{-1}\overline{\mathbf{X}}'.\end{aligned}$$

The above expression of $\widehat{\mathbf{X}}'$ indicates that the point, $(\widehat{\mathbf{x}}^j)'$, is the orthogonal projection of the point, $(\overline{\mathbf{x}}^j)'$, onto the row space of \mathbf{M}^r in the metric \mathbf{W}^{-1} , $j \in [1 : J]$ (see Section 1.6.7).

The three different types of predictive CVA biplots will now be illustrated at the hand of the simulated data set introduced in Section 4.2.1.1. Figures 4.4(a), 4.4(b) and 4.5 provide the two-dimensional predictive unweighted CVA biplot constructed with $\mathbf{C} = \mathbf{I}$, the two-dimensional predictive unweighted CVA biplot constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$ and the two-dimensional predictive weighted CVA biplot respectively.

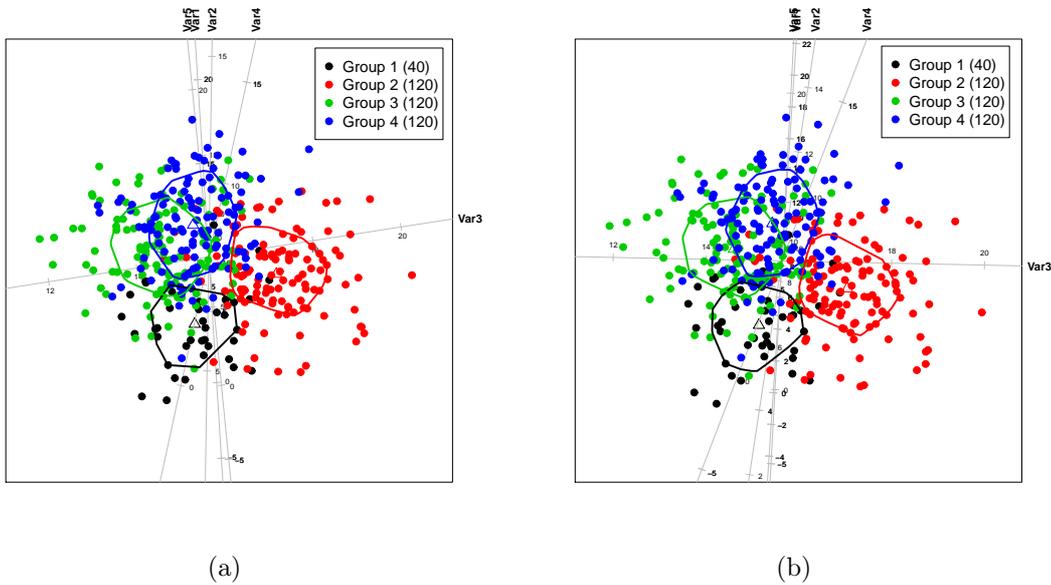


Figure 4.4: (a) The two-dimensional predictive unweighted CVA biplot of the simulated data set constructed with $\mathbf{C} = \mathbf{I}$; (b) The two-dimensional predictive unweighted CVA biplot of the simulated data set constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$.

The unweighted CVA biplots of the simulated data set constructed with $\mathbf{C} = \mathbf{I}$ and $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$ respectively appear to be almost identical. The large differences between the weighted CVA biplot and the two unweighted CVA biplots is attributable to the large difference between the size of Group 1 and the sizes of Groups 2, 3 and 4. Upon comparison of the CVA biplots in Figures 4.4(a), 4.4(b)

and 4.5 to the CVA displays in Figures 4.2, 4.3 and 4.1 respectively, it is clear that the positions of the group centroids and individual samples in the CVA biplots are identical to those in the corresponding CVA displays as expected. The advantage of the CVA biplots to the corresponding CVA displays is the information on the original measured variables contained in the biplots. The weighted CVA biplot in Figure 4.5 suggests that Group 2 differs substantially from Group 3 with respect to the variable $Var3$ while and differs substantially from Group 4 with respect to the variables $Var1$, $Var2$, $Var3$ and $Var5$ while Groups 3 and 4 differ substantially with respect to the variable $Var4$. This biplot also suggests that Group 1 is very similar to Groups 2 and 3 with respect to the variables $Var1$, $Var2$, $Var4$ and $Var5$. Remember that the dissimilarities and similarities between the groups with respect to specific variables that are suggested by the CVA biplot can only be trusted to be true in reality if the biplot axes accurately predict the measurements on the variables under consideration.

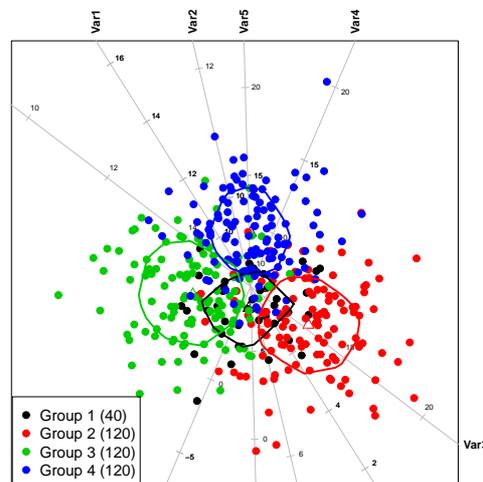


Figure 4.5: *The two-dimensional predictive unweighted CVA biplot of the simulated data set constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$.*

4.6 The scale invariance of the CVA biplot

From Section 4.4 it is known that in general, the point representing the j th group mean in the r -dimensional CVA biplot is given by $(\bar{\mathbf{x}}^j)' \mathbf{M}_r$, where \mathbf{M} satisfies the two-sided eigenvalue problem,

$$\mathbf{B}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{A}$$

where $\mathbf{B} = \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}$, and can be expressed as

$$\mathbf{M} = \mathbf{L}\mathbf{V}$$

where \mathbf{L} is a $p \times p$ non-singular matrix which is such that $\mathbf{L}\mathbf{L}' = \mathbf{W}^{-1}$ and \mathbf{V} is the matrix of right singular vectors of the matrix,

$$\mathbf{C}^{1/2} \overline{\mathbf{X}} \mathbf{L}.$$

Let \mathbf{A} , as in Section 4.2.4, denote the $p \times p$ diagonal matrix with i th diagonal element equal to the sample standard deviation of the i th measured variable, x_i . It was shown in Section 4.6 that the matrix of group means and the within-group matrix of sums of squares and cross products associated with the matrix of standardised measurements, $\mathbf{X}^* = \mathbf{X}\mathbf{A}^{-1}$, are given by

$$\begin{aligned} \overline{\mathbf{X}}^* &= \overline{\mathbf{X}}\mathbf{A}^{-1} \\ \text{and } \mathbf{W}^* &= \mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1} \end{aligned} \quad (4.6.1)$$

respectively. It follows from (4.6.1) and the spectral decomposition of \mathbf{W}^{-1} in (4.4.7) that the $p \times p$ non-singular matrix \mathbf{L}^* which is such that

$$\mathbf{L}^* (\mathbf{L}^*)' = (\mathbf{W}^*)^{-1}$$

is given by $\mathbf{L}^* = \mathbf{A}\mathbf{L}$:

$$\begin{aligned} (\mathbf{W}^*)^{-1} &= \mathbf{A}\mathbf{W}^{-1}\mathbf{A} \\ &= \mathbf{A}\mathbf{E}\mathbf{E}^{-1}\mathbf{E}'\mathbf{A} \\ \longrightarrow (\mathbf{W}^*)^{-1} &= \mathbf{L}^* (\mathbf{L}^*)' \\ \text{where } \mathbf{L}^* &= \mathbf{A}\mathbf{E}\mathbf{E}^{-1/2} \\ \therefore \mathbf{L}^* &= \mathbf{A}\mathbf{L}. \end{aligned}$$

In Section 4.2.4 it was also shown that for $\mathbf{C} = \mathbf{N}$ and $\mathbf{C} = \mathbf{I}$, the matrix $\mathbf{B} = \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}$ associated with the standardised measurements is given by $\mathbf{B}^* = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. It is

shown below that this is also true for $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$:

$$\begin{aligned}\mathbf{B}^* &= (\overline{\mathbf{X}}^*)' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \overline{\mathbf{X}}^* \\ &= \mathbf{A}^{-1} \overline{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}' \right) \overline{\mathbf{X}} \mathbf{A}^{-1} \\ \longrightarrow \mathbf{B}^* &= \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} .\end{aligned}$$

Furthermore, it was shown in Section 4.2.4 that the matrix $\mathbf{M}^* = \mathbf{A}\mathbf{M}$, with k th column vector denoted by $\mathbf{m}_{(k)}^*$, is the matrix with first column vector equal to the vector \mathbf{m}^* that maximises the ratio,

$$\frac{(\mathbf{m}^*)' \mathbf{B}^* \mathbf{m}^*}{(\mathbf{m}^*)' \mathbf{W}^* \mathbf{m}^*} \quad (4.6.2)$$

under the constraint $(\mathbf{m}^*)' \mathbf{W}^* \mathbf{m}^* = 1$ and with i th column vector given by the vector \mathbf{m}^* that maximises the ratio in (4.6.2) under the constraint, $(\mathbf{m}^*)' \mathbf{W}^* \mathbf{m}^* = 1$ and conditional on $(\mathbf{m}^*)' \mathbf{W}^* \mathbf{m}_{(k)}^* = 0$ for all $k < i$, $i \in [2 : p]$. This result holds for all three choices of \mathbf{C} . It follows that for all three choices of \mathbf{C} , the coordinates of the points representing the J group means in the r -dimensional CVA biplot constructed from the standardised data set are therefore given by the rows of the matrix

$$\overline{\mathbf{X}}^* \mathbf{M}_r^* = \overline{\mathbf{X}} \mathbf{A}^{-1} \mathbf{A} \mathbf{M}_r = \overline{\mathbf{X}} \mathbf{M}_r$$

while the coordinates of the points representing the n individual observations in the r -dimensional CVA biplot constructed from the standardised measurements are given by the rows of the matrix,

$$\mathbf{X}^* \mathbf{M}_r^* = \mathbf{X} \mathbf{A}^{-1} \mathbf{A} \mathbf{M}_r = \mathbf{X} \mathbf{M}_r .$$

It is evident that as far as the representation of the group means and the individual observations are concerned, the r -dimensional CVA biplot constructed from the standardised measurements is identical to the r -dimensional CVA biplot constructed from the unstandardised measurements for all three choices of \mathbf{C} .

It follows that in order for the interpolative and predictive CVA biplots to be completely unaffected by standardization, the interpolative and predictive biplot axes need to be unaffected by standardization. From the definition of interpolative and predictive biplot axes in the Sections 4.5.1 and 4.5.2, it is evident that the

interpolative biplot axis representing the k th original variable in the r -dimensional interpolative CVA biplot constructed from the matrix $\mathbf{X}^* = \mathbf{X}\mathbf{A}^{-1}$ is in the direction of the vector,

$$\mathbf{e}'_k \mathbf{M}_r^* = \mathbf{e}'_k \mathbf{A} \mathbf{M}_r = [\mathbf{A}]_{kk} \mathbf{e}'_k \mathbf{M}_r$$

while the predictive biplot axis representing the k th original variable in the r -dimensional predictive CVA biplot constructed from the matrix, $\mathbf{X}^* = \mathbf{X}\mathbf{A}^{-1}$, is in the direction of the vector,

$$\mathbf{e}'_k (\mathbf{M}^{*r})' = \mathbf{e}'_k \mathbf{A}^{-1} (\mathbf{M}^r)' = [\mathbf{A}]_{kk} \mathbf{e}'_k (\mathbf{M}^r)'$$

where $k \in [1 : p]$. Since the vector, $\mathbf{e}'_k \mathbf{M}_r^*$, is simply a scalar multiple of the vector, $\mathbf{e}'_k \mathbf{M}_r$, the k th interpolative biplot axis of the r -dimensional interpolative CVA biplot constructed from the standardised measurements lie in exactly the same direction as that of the r -dimensional interpolative CVA biplot constructed from the unstandardised measurements. Similarly, since the vector, $\mathbf{e}'_k (\mathbf{M}^{*r})'$, is simply a scalar multiple of the vector, $\mathbf{e}'_k (\mathbf{M}^r)'$, the k th predictive biplot axis of the r -dimensional predictive CVA biplot constructed from the standardised measurements lie in exactly in the same direction as that of the r -dimensional predictive CVA biplot constructed from the unstandardised measurements. Note that the axes of the CVA biplot can be calibrated in the scales of either the unstandardised or standardised measurements. If the biplot axes of the r -dimensional interpolative CVA biplot are calibrated in terms of the scales of the standardised measurements (that is, calibrated in standard deviation units), then the point

$$[\mathbf{A}]_{kk} \mathbf{e}'_k \mathbf{M}_r \tag{4.6.3}$$

will be calibrated with the value one while if the interpolative biplot axes are calibrated in terms of the scales of the unstandardised measurements, then the point in (4.6.3) will be calibrated with the value, $\bar{\mathbf{x}}^k + [\mathbf{A}]_{kk}$. When the axes of the r -dimensional predictive CVA biplot are calibrated in the same scales as the unstandardised measurements, then the approximations read off from the predictive biplot axes by orthogonally projecting the point representing the j th group mean in the biplot onto the respective predictive biplot axes, are approximations to the elements of the observed j th group mean, $\bar{\mathbf{x}}^j$, that is

$$(\hat{\bar{\mathbf{x}}}^j)' = (\bar{\mathbf{x}}^j)' \mathbf{M}_r \mathbf{M}^r$$

4.7. A COMPARISON BETWEEN A CVA BILOT AND A PCA BILOT 241

$r \in [1 : p]$, $j \in 1 : J$. Similarly, the approximations read off from the predictive biplot axes by orthogonally projecting the point representing the i th sample of the j th group in the biplot onto the respective predictive biplot axes, are approximations to the elements of the observed i th sample of the j th group, \mathbf{x}_i^j , that is

$$(\hat{\mathbf{x}}_i^j)' = (\mathbf{x}_i^j)' \mathbf{M}_r \mathbf{M}^r$$

$r \in [1 : p]$, $j \in 1 : J$. When however the axes of the r -dimensional predictive CVA biplot are calibrated in the scales of the standardised measurements, then the approximations read off from the predictive biplot axes by orthogonally projecting the point representing the j th group mean and the point representing the i th sample of the j th group onto the respective predictive biplot axes, are approximations to the elements of the j th group mean of the standardised measurements, $(\hat{\bar{\mathbf{x}}}^{j*})' = (\bar{\mathbf{x}}^j)' \mathbf{A}^{-1}$, and the standardised i th sample of the j th group, $(\hat{\mathbf{x}}_i^{j*})' = (\mathbf{x}_i^j)' \mathbf{A}^{-1}$, respectively. The expressions for these approximations are given by

$$\begin{aligned} (\hat{\bar{\mathbf{x}}}^{j*})' &= (\bar{\mathbf{x}}^j)' \mathbf{M}_r^* \mathbf{M}^{r*} \\ &= (\bar{\mathbf{x}}^j)' \mathbf{A}^{-1} \mathbf{A} \mathbf{M}_r \mathbf{M}^r \mathbf{A}^{-1} \\ &= (\bar{\mathbf{x}}^j)' \mathbf{A}^{-1} \\ \text{and } (\hat{\mathbf{x}}_i^{j*})' &= (\mathbf{x}_i^j)' \mathbf{M}_r^* \mathbf{M}^{r*} \\ &= (\mathbf{x}_i^j)' \mathbf{A}^{-1} \mathbf{A} \mathbf{M}_r \mathbf{M}^r \mathbf{A}^{-1} \\ &= (\mathbf{x}_i^j)' \mathbf{A}^{-1} \end{aligned}$$

respectively.

Upon substituting any $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in each of the expressions above, it is evident that CVA and the CVA biplot are invariant to all non-singular linear transformations of the form

$$\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x}.$$

4.7 A comparison between a CVA biplot and a PCA biplot

Unlike the r -dimensional PCA biplot, which is aimed at representing the matrix of individual samples as accurately as possible in r -dimensional space (in terms of least squares), the CVA biplot is aimed representing the group structure underlying the data as well as possible in r -dimensional space and achieves this aim by separating

the groups as much as possible. Consequently, the CVA biplot differs fundamentally from the PCA biplot with respect to the role of the group membership of the individual samples in the construction of the biplot as well as the distance metric central to the biplot and the degree to which the groups are separated in the biplot. Whereas the construction of the PCA biplot is based on the matrix of individual samples and completely unaffected by the group membership of the individual samples and the group means, the construction of the CVA biplot is based on the matrix of group means and hence group membership plays an integral role in the construction. The group means do not influence the construction of the PCA biplot but can be interpolated onto the PCA biplot constructed from the individual samples. In the case of the CVA biplot, the individual samples do not influence the construction of the CVA biplot but can be interpolated onto the CVA biplot constructed from the group means. The J group means are always perfectly predicted in the K -dimensional CVA biplot. In the PCA biplot however, all the group means will only necessarily be perfectly predicted when the biplot is p -dimensional. Recall that the Pythagorean distances in the PCA biplot optimally approximates the corresponding Pythagorean distances in the p -dimensional measurement space. The Pythagorean distances in the CVA biplot on the other hand, approximates the corresponding Mahalanobis distances in the p -dimensional measurement space up to a constant multiple. Recall that for each of the three types of CVA biplots, the groups are maximally separated according to some trace criterion. Consequently the groups will be more separated in the CVA biplot than in the PCA biplot of the same dimension. Another difference between the CVA and PCA biplots is that the CVA biplot is scale invariant while the PCA biplot is scale dependent. As for the PCA biplot, both interpolative and predictive CVA biplot axes are linear. However, unlike in the case of the PCA biplot, the interpolative and predictive CVA biplot axes corresponding to a particular variable do not lie in the same direction.

Recall that in Chapter 2 the *Ocotea* data set, which consist of 37 samples structured into three groups (species), was represented in a two-dimensional PCA biplot (constructed from the standardised measurements of the data set) in which the group structure underlying the data was graphically represented by using different plotting characters and colours to represent samples (and group centroids) belonging to different groups. To illustrate the CVA biplot's superior ability to discriminate between the groups of a data set compared to the PCA biplot, the two dimensional predictive unweighted CVA biplot of the *Ocotea* data set (with $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$) is provided in Figure 4.6(b) alongside the two-dimensional predictive PCA biplot constructed from the standardised measurements of the data set in Figure 4.6(a). Note that since the CVA biplot is scale invariant, the CVA biplot need not be constructed from the standardised measurements. In both of these biplots 95% bags were superimposed onto the biplot for the *O. bullata* and *O. porosa* species while a convex hull was superimposed for the *O. kenyensis* specie (due to the small number of observations belonging to this specie). As in Chapter 2, solid black squares were used to represent the 20 samples belonging to *O. bullata*, solid red triangles were used to represent the seven samples belonging to *O. kenyensis* and solid green circles were used to represent the ten samples belonging to *O. porosa*.

4.7. A COMPARISON BETWEEN A CVA BIPLLOT AND A PCA BIPLLOT 243

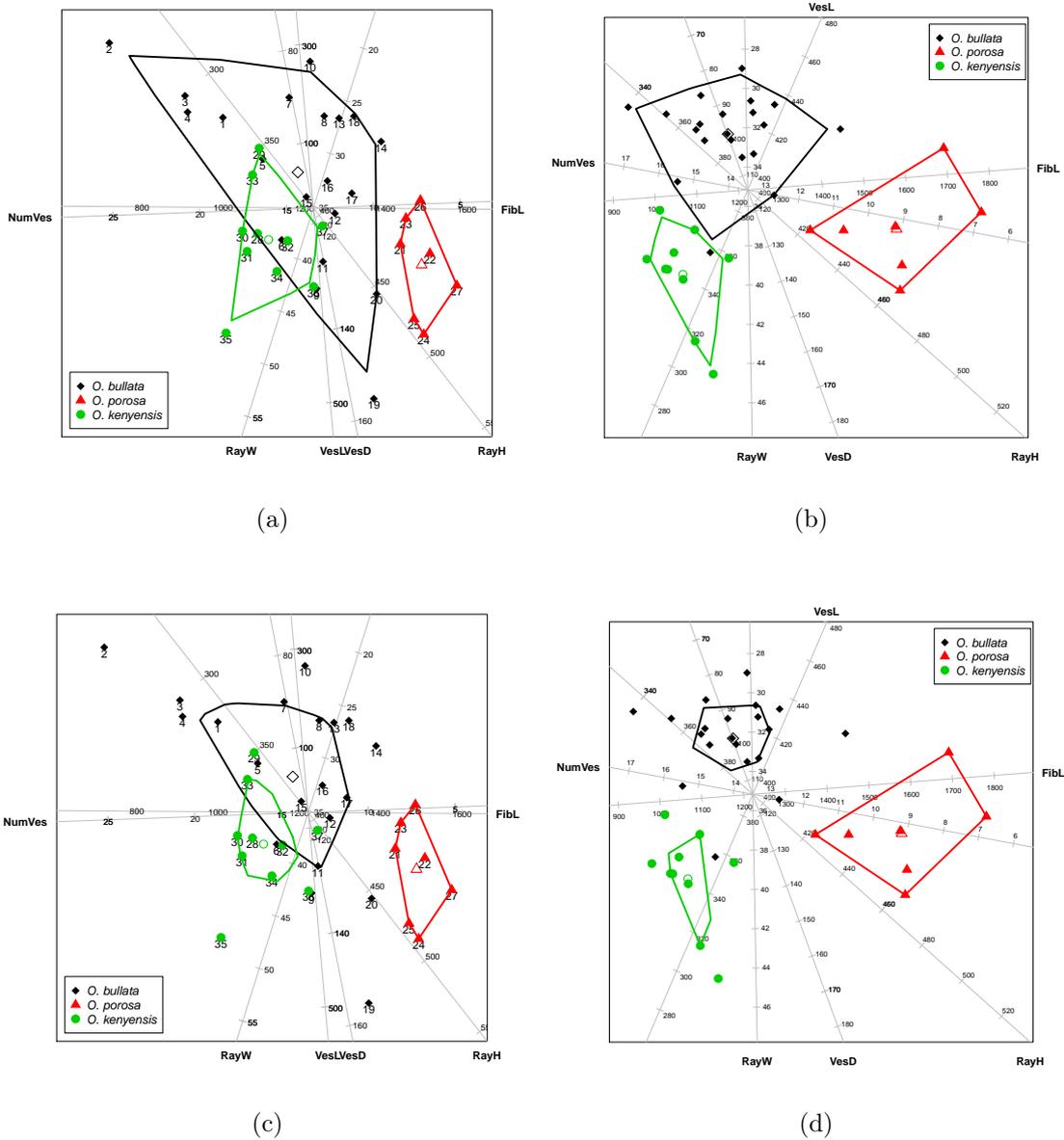


Figure 4.6: (a) and (c) The two-dimensional predictive PCA biplot constructed from the standardised measurements of the *Ocotea* data set; (b) and (d) The two-dimensional predictive CVA biplot of the *Ocotea* data set. In (a) and (b) 95% bags are superimposed for the species *O. bullata* and *O. porosa* while a convex hull is constructed for the specie *O. kenyensis*. In (c) and (d) 50% bags are superimposed for the species *O. bullata* and *O. porosa* while a convex hull is constructed for the specie *O. kenyensis*.

Figure 4.6(c) and Figure 4.6(d) contain the two-dimensional predictive PCA biplot and the two-dimensional predictive unweighted CVA biplot (with $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$) with 50% bags superimposed for the *O. bullata* and *O. porosa* species and a convex hull superimposed for the *O. kenyensis* specie. The groups appear to be substantially more separated in the CVA biplot than in the PCA biplot - there is no overlap between either the 50% bags and convex hull in the CVA biplot in Figure 4.6(d) or

the 95% bags and convex hull in the CVA biplot in Figure 4.6(b), whereas even the 50% bags corresponding to the *O. bullata* and *O. porosa* species overlap in the PCA biplot. Apart from the extent of the separation and/or overlap amongst the groups, the two-dimensional unweighted CVA biplot differs substantially from the PCA biplot of the *Ocotea* data set. This is however to be expected given the fundamental differences in the construction of these two types of biplots.

4.8 The effect of accounting for the group sizes in the CVA biplot

Recall that when the sizes of all the groups in a data set are identical, the weighted CVA biplot and the unweighted CVA biplot with $\mathbf{C} = \mathbf{I}$ are identical. It seems reasonable to expect that, the greater the differences in relative group sizes, the more the r -dimensional weighted and unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplots will differ, $r < K$.

In this section, seven simulated data sets will be used to investigate the effect of taking group sizes into account in the construction of a CVA biplot. Each simulated data set consists of 400 randomly selected observations, each of which is measured on the same five variables and belongs to one of four groups. For each of the data sets the observations of all four the groups are multivariate normally distributed. The four groups have different population means but identical population covariance matrices. The seven data sets were all drawn from the same four multivariate normal distributions but differ with respect to the sizes of the four groups. Tables 4.1 and 4.2 provide the population means of the four groups and the common population correlation matrix of the four groups respectively.

For each of the data sets, the two-dimensional (predictive) weighted CVA biplot will be compared to the two-dimensional (predictive) unweighted CVA biplot with $\mathbf{C} = \mathbf{I}$. The weighted CVA biplots could also have been compared to the corresponding unweighted CVA biplots with $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$ but since the unweighted CVA biplot with $\mathbf{C} = \mathbf{I}$ is more comparable to the weighted CVA biplot (in that it is identical to the weighted CVA biplot when the groups are of equal sizes), the weighted CVA biplots will only be compared to the latter unweighted CVA biplot here.

For each of the simulated data sets the sample sizes of the four groups are large enough so that the sample group means should be accurate estimates of the population group means. Consequently, the sample group means should be similar across the seven simulated data sets. Since the construction of an unweighted CVA biplot is based on the matrix of group means alone, the differences between the seven unweighted CVA biplots with respect to the relative positions of the group means and the biplot axes will be the result of sampling variation alone and are therefore expected to be minimal. For the same reason the positions and shapes of the α -bags, which are based on the positions of the interpolated samples, should not differ much across data sets. On the other hand, since the relative sizes of the four groups differ substantially across the seven simulated data sets, the differences between the seven weighted CVA biplots are expected to be substantial. Furthermore, the more

4.8. THE EFFECT OF ACCOUNTING FOR THE GROUP SIZES IN THE CVA BIPLLOT

245

the relative sizes of the four groups differ, the greater the differences between the unweighted and weighted CVA biplots are expected to be. To ease visualisation of the degree of overlap amongst the four groups, a 50% bag has been superimposed onto each of the biplots for each of the four groups. Also, the four groups are represented by different colours - the observations (solid circles) and mean (asterisk) of Group 1 are black, those of Group 2 are red, those of Group 3 are green while those of Group 4 are blue.

The first data set has equal sized groups, that is 100 observations in each of the four groups. The weighted and unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplots of this data set will therefore be identical as explained in Section 4.2. This is illustrated by Figures 4.7(a) and 4.7(b) which show the unweighted and weighted two-dimensional CVA biplot respectively. The weighted and unweighted CVA biplots of this first data set serve as a benchmark to which the biplots of the other simulated data sets (that do not have equal sized groups) can be compared in order to obtain a better understanding of the effect of accounting for the group sizes in the construction of the CVA biplot when the groups have different sizes.

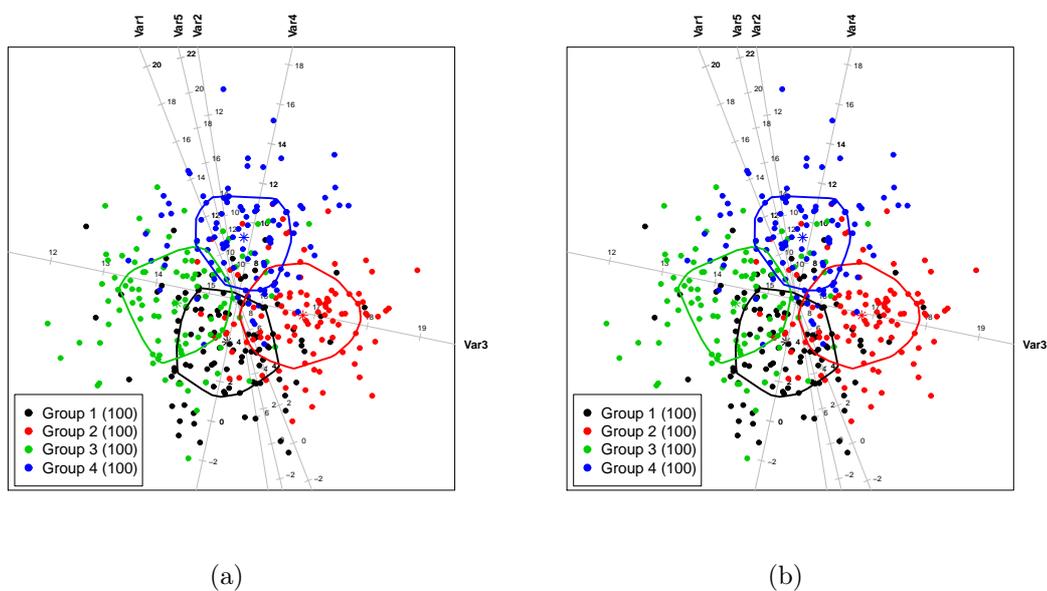


Figure 4.7: (a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the first simulated data set; (b) The two-dimensional weighted CVA biplot of the first simulated data set.

The sizes of the four groups in the second simulated data set differ slightly - Group 1 has 100 observations, Group 2 has 105 observations, Group 3 has 90 observations and Group 4 has 105 observations. Due to the fact that the differences in the group sizes are so small, the differences between the unweighted and weighted CVA biplot of this data set are barely noticeable. The only difference that is relatively easy to spot, is that the angle between the biplot axes representing *Var2* and *Var5* is smaller in the weighted CVA biplot than in the unweighted CVA biplot.

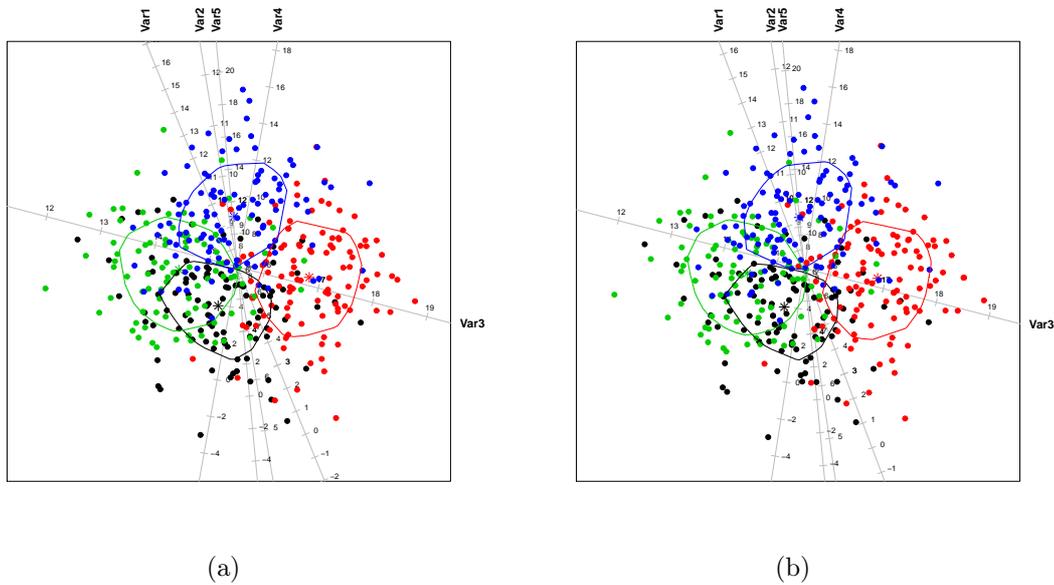


Figure 4.8: (a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the second simulated data set; (b) The two-dimensional weighted CVA biplot of the second simulated data set.

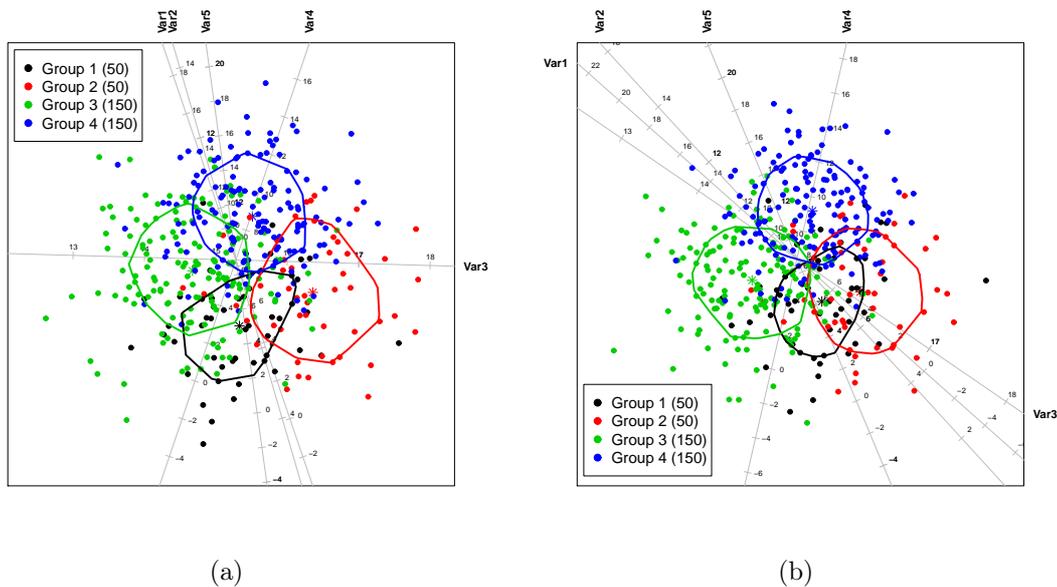


Figure 4.9: (a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the third simulated data set; (b) The two-dimensional weighted CVA biplot of the third simulated data set.

For the third simulated data set, the sizes of Groups 1, 2, 3 and 4 are 50, 50, 150 and 150 respectively. Due to the large differences in the group sizes, the weighted CVA biplot is expected to differ quite substantially from the unweighted CVA biplot. This expectation is confirmed upon comparison of Figures 4.9(a) and 4.9(b) which represents the unweighted and weighted CVA biplots of this data set respectively. The relative positions of the biplot axes in the weighted CVA biplot differ greatly

4.8. THE EFFECT OF ACCOUNTING FOR THE GROUP SIZES IN THE CVA BILOT

247

from that in the unweighted CVA biplot. Also, looking at the Pythagorean distances between the points representing the four group centroids, it is evident that in the weighted CVA biplot the distance between the group centroids of Group 1 and Group 2 is much smaller relative to the other distances compared to in the unweighted CVA biplot. In the unweighted CVA biplot the degree of overlap between the various pairs of groups seems to be comparable, except for Groups 1 and 4 whose 50% bags do not overlap at all. In the weighted CVA biplot however, the degree to which Groups 1 and 2 (the two smaller groups) overlap is much greater than the degree to which Groups 3 and 4 (the two larger groups) overlap with each other as well as with Groups 1 and 2. Also, the extent to which Groups 3 and 4 overlap with each other as well as with Groups 1 and 2 in the weighted CVA biplot is less than in the unweighted CVA biplot while the extent to which Group 1 and Group 2 overlap with each other in the weighted CVA biplot is much greater than in the unweighted CVA biplot.

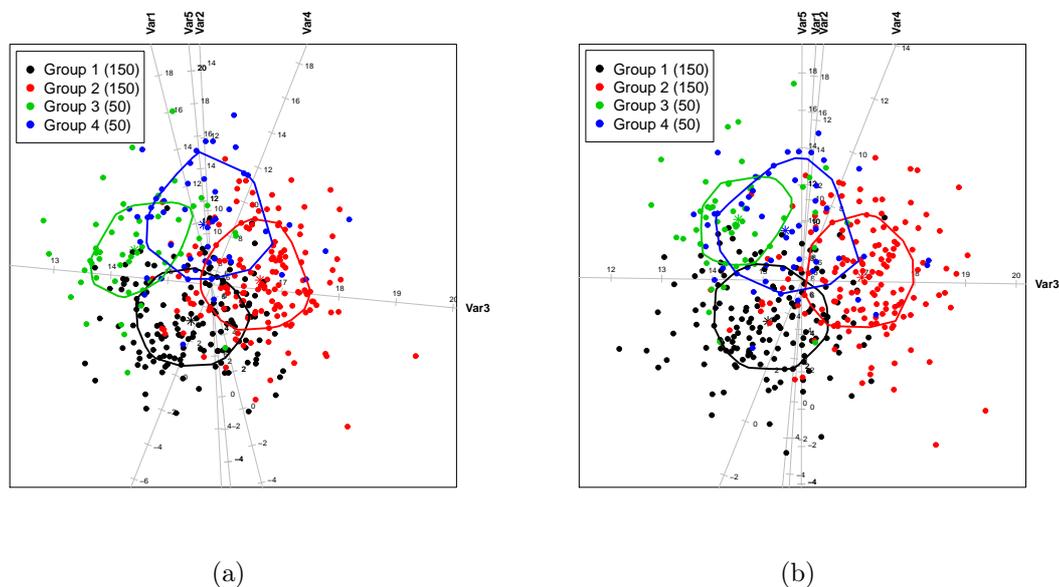


Figure 4.10: (a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the fourth simulated data set; (b) The two-dimensional weighted CVA biplot of the fourth simulated data set.

The fourth simulated data set contains 150 observations belonging to each of Groups 1 and 2 and 50 observations belonging to each of Groups 3 and 4. Due to the large differences in the sizes of the four groups, the differences between the weighted and unweighted CVA biplots with respect to the relative positions of the biplot axes and group centroids, as well as the degree of overlap amongst the various pairs of groups, are clearly visible. The degree of overlap between the various pairs of groups seem to be comparable in the unweighted CVA biplot, which is provided in Figure 4.10(a). However, in the weighted CVA biplot, which is provided in Figure 4.10(b), the degree to which Group 1 and Group 2 overlap with each other and the other groups is much less than the degree to which Group 3 and Group 4 overlap

with each other. Note that this is the opposite of what was the case for the third simulated data set.

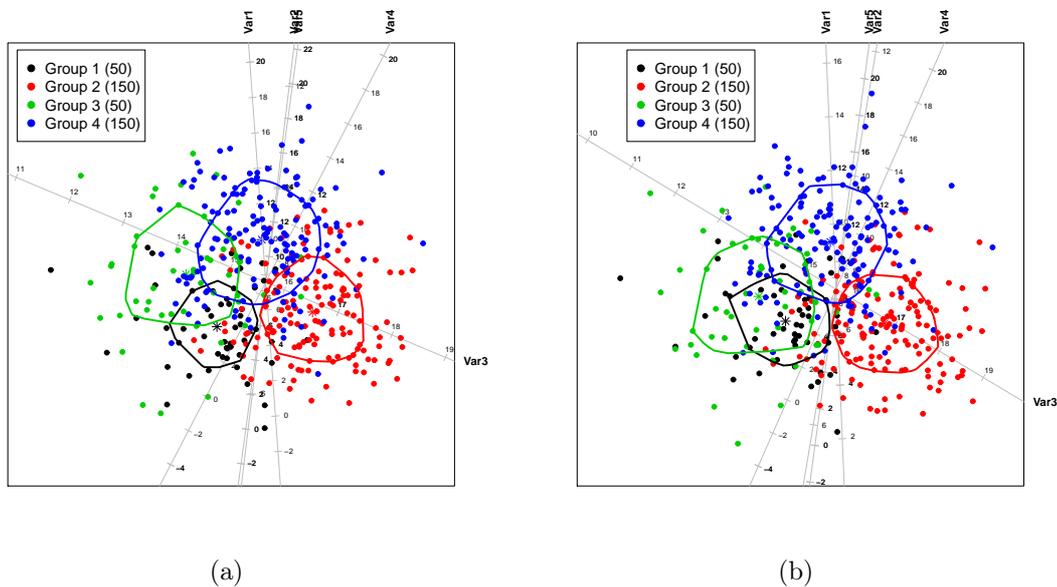


Figure 4.11: (a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the fifth simulated data set; (b) The two-dimensional weighted CVA biplot of the fifth simulated data set.

The fifth simulated data set contains 50 observations belonging to each of Groups 1 and 3 and 150 observations belonging to each of Groups 2 and 4. The relative positions of the biplot axes in the weighted CVA biplot do not seem to differ much from those in the unweighted CVA biplot. The only difference which is clearly visible is the relative positions of the biplot axes that represent *Var2* and *Var5* - it seems as if the two biplot axes almost swapped positions in the weighted CVA biplot. However, given the very small angle between these two biplot axes, the difference is not large. The distance between the points representing the group centroids of Group 1 and Group 3 (i.e. the two smaller groups) in the weighted CVA biplot relative to the distances between the other pairs of group centroids is much smaller than in the unweighted CVA biplot. Accordingly, the degree of overlap between Group 1 and Group 3 in the weighted CVA biplot is much larger than in the unweighted CVA biplot. On the other hand, the degree of overlap between Group 2 and Group 4 (i.e. the two larger groups) in the weighted CVA biplot is slightly less than in the unweighted CVA biplot.

The sixth simulated data set is the data set that was introduced in Section 4.2.1.1. Recall that this data set contains 120 observations belonging to each of Groups 2, 3 and 4 and only 40 observations belonging to Group 1. The very small size of Group 1 relative to the sizes of the other groups results in big differences between the weighted and unweighted CVA biplots of the data set. The relative positions of the biplot axes as well as the points representing the group centroids in the weighted CVA biplot in Figure 4.12(b) differ drastically from that in the unweighted CVA

4.8. THE EFFECT OF ACCOUNTING FOR THE GROUP SIZES IN THE CVA BILOT 249

biplot shown in Figure 4.12(a). In the weighted CVA biplot, the three larger groups are well separated while in the unweighted CVA biplot there is substantial overlap between Groups 3 and 4. On the other hand, Group 1, which is well separated from the other three groups in the unweighted CVA biplot, overlaps substantially with Group 2 and Group 3 in the weighted CVA biplot.

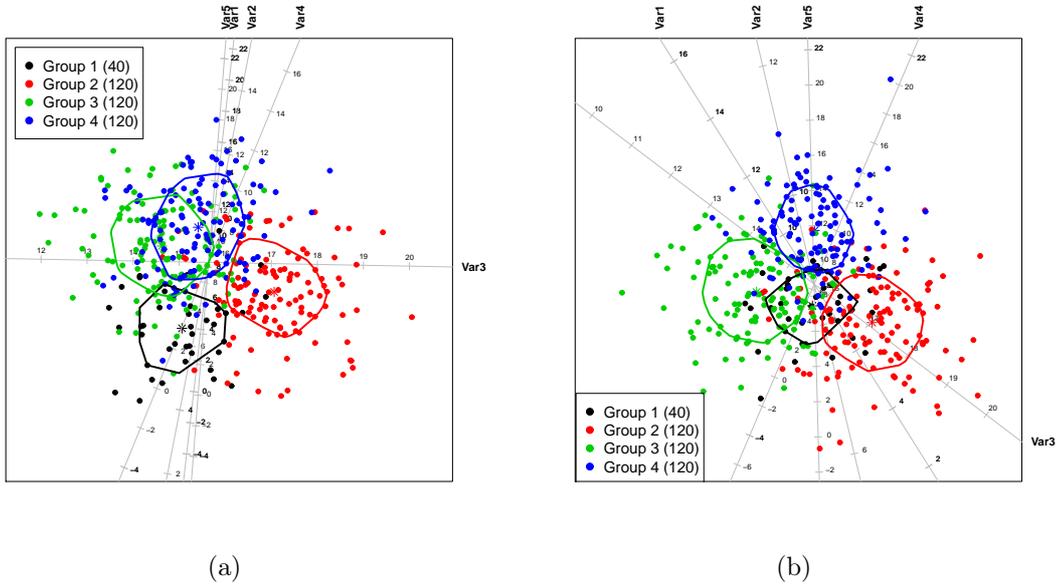


Figure 4.12: (a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the sixth simulated data set; (b) The two-dimensional weighted CVA biplot of the sixth simulated data set.

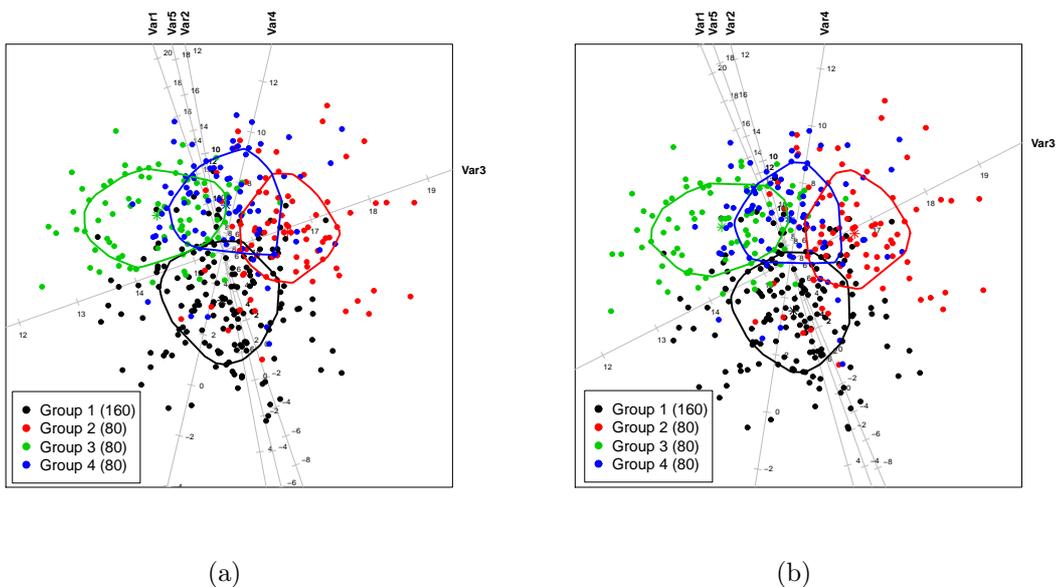


Figure 4.13: (a) The two-dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot of the seventh simulated data set; (b) The two-dimensional weighted CVA biplot of the seventh simulated data set.

The seventh simulated data set contains 160 observations belonging to Group 1 and 80 observations belonging to each of Group 2, Group 3 and Group 4. Contrary to what may have been expected, the weighted and unweighted CVA biplots of the data sets are very similar - so similar in fact that it is difficult to spot the differences. Groups 1 and 3 do however seem to be separately to a slightly greater extent in the weighted CVA biplot than in the unweighted CVA biplot.

In conclusion, it is evident that the more the group sizes differ, the greater the differences between the weighted CVA biplot and the unweighted CVA biplot ($\mathbf{C} = \mathbf{I}$). Since all the data sets originate from the same population and only differ with respect to the relative group sizes, the unweighted CVA biplots of the different data sets (which do not take the group sizes into account) are quite similar. The weighted CVA biplots of the different data sets on the other hand differ quite drastically. Upon comparison of the weighted CVA biplots of the simulated data sets to the corresponding unweighted CVA biplots, it is evident that the larger groups usually overlap to a lesser extent with each other as well as with the other groups in the weighted CVA biplot than in the corresponding unweighted CVA biplot while the smaller groups usually overlap to a greater extent with each other and the other groups in the weighted CVA biplot than in the corresponding unweighted CVA biplot. Compare for example the weighted CVA biplots in Figures 4.9(b), 4.10(b) and 4.12(b) with the corresponding unweighted CVA biplots in Figures 4.9(a), 4.10(a) and 4.12(a) respectively. It can therefore be said that the weighted CVA biplot ‘focuses’ more on the separation of the larger groups and less on the separation of the smaller groups than does the unweighted CVA biplot.

It also seems that the weighted CVA biplot ‘focuses’ more on the separation of the larger groups than on the separation of the smaller groups. Consider for example the weighted CVA biplot in Figure 4.9(b), where Group 3 and Group 4 are equally large and much larger than Group 1 and Group 2 and the weighted CVA biplot in Figure 4.10(b), where Group 1 and Group 2 are equally large and much larger than Group 3 and Group 4. In both these biplots, the degree to which the two larger groups overlap with each other as well as with the two smaller groups is much less than the degree to which the two smaller groups overlap with each other.

The fact that the above-mentioned patterns are observed for a number of simulated data sets drawn from the same multivariate normal population, comprising groups of sizes between 40 and 160 samples, suggests that it is very unlikely that these patterns occurred merely as a result of sampling variation. The effect of taking the (possibly different) group sizes into account in the construction of the CVA biplot, on the quality measures of the CVA biplot will be discussed in Chapter 5.

4.9 Summary

CVA is a linear dimension reduction technique which is designed to discriminate between predefined groups in a data set and to classify observations of unknown origin. CVA can be defined in a number of different ways, including (1) as equivalent to LDA for the multi-group scenario; (2) as a special case of CCA and (3) as a two-step process. CVA is also closely related to MANOVA. A very convenient property of CVA, in contrast with PCA, is the fact that it is scale invariant. CVA can be

weighted or unweighted, depending on whether or not the (possibly) different group sizes are taken into account in the analysis.

The CVA biplot is designed to graphically represent the group structure underlying a data set comprising samples that are structured into a number of predefined groups. The r -dimensional CVA biplot is a graphical representation of the r -dimensional CVA solution with axes superimposed to provide information on the original measured variables, $r \in [1 : 3]$. The CVA biplot allows visualisation of the exact CVA classification regions when the dimension of the biplot is equal to $K = \min(J - 1, p)$ and visualisation of approximate CVA classification regions when the dimension of the biplot is less than K . Unlike the PCA biplot, the CVA biplot is constructed from the matrix of group centroids and not the matrix of individual samples. Interpolating the individual samples onto the CVA biplot allows a more detailed representation of the approximate degree and nature of the overlap and/or separation amongst the groups as well as visualisation of the within-group dispersion. Furthermore, the CVA biplot differs from the PCA biplot in that it is scale invariant and in that the interpolative and predictive CVA biplot axes corresponding to a particular variable are not collinear as in the PCA biplot.

The CVA biplot, like CVA, can be weighted or unweighted, depending on whether or not the (possibly) different group sizes are taken into account in the construction of the biplot. In general, the CVA biplot is constructed from the svd of the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$. The matrix \mathbf{L} transforms the group centroids to the canonical space while the matrix $\mathbf{C}^{1/2}$ determines whether the resulting CVA biplot will be weighted or not. Setting the matrix $\mathbf{C}^{1/2}$ equal to $\mathbf{N}^{1/2}$ yields the weighted CVA biplot while both $\mathbf{C}^{1/2} = \mathbf{I}$ and $\mathbf{C}^{1/2} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$ yield unweighted CVA biplots. The weighted CVA biplot takes the (possibly) different sizes of the groups into account in the construction of the CVA biplot while the two unweighted CVA biplots implicitly assume that the groups are equally sized. When the J groups have identical sizes, the weighted CVA biplot is identical to the unweighted CVA biplot constructed from the svd of $\bar{\mathbf{X}}\mathbf{L}$. The more the group sizes differ, the more the weighted CVA biplot will differ from the two unweighted CVA biplots. The K -dimensional weighted and unweighted CVA biplots are identical.

The relative magnitudes of the Pythagorean distances in the CVA biplot approximate the relative magnitudes of the corresponding Mahalanobis distances in the measurement space. The relative magnitudes of the Mahalanobis distances between the individual samples in the measurement space are optimally approximated in the weighted CVA biplot while the relative magnitudes of the Mahalanobis distances between the group centroids in the measurement space are optimally approximated in the unweighted CVA biplot corresponding to $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$.

The larger groups receive greater weights than the smaller groups in the minimisation criteria which determines the weighted CVA biplot space while all J the groups receive identical weights in the minimisation criteria which determines both the unweighted CVA biplot space corresponding to $\mathbf{C} = \mathbf{I}$ and that corresponding to $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$. Consequently the points representing the group means of the larger groups in the p -dimensional canonical space tend to lie closer to their orthogonal projections onto the r -dimensional weighted CVA biplot space than to their orthogonal projections onto the two r -dimensional unweighted CVA biplot spaces,

$r \in [1 : K - 1]$. The larger groups also tend to be more separated from the other groups in the r -dimensional weighted CVA biplot compared to in the r -dimensional unweighted CVA biplot, $r \in [1 : K]$.

It is important to remember that it is only appropriate to represent the group structure underlying a data set in a CVA biplot if the assumption of identical within-group covariance matrices is appropriate for that data set. If this assumption is not appropriate, an AOD biplot should be used to graphically represent the group structure underlying the data set (Gardner *et al.*, 2005).

4.10 Appendix

4.10.1 The derivation of the result in Section 4.2

It will now be shown that the last $p - K$ elements of the J canonical means, $\{(\bar{\mathbf{x}}^j)'\mathbf{M}\}$, where \mathbf{M} is the $p \times p$ matrix with i th column vector equal to the i th eigenvector of the two-sided eigenvalue problem

$$\mathbf{X}'\mathbf{X}\mathbf{m} = \lambda\mathbf{W}\mathbf{m},$$

are identical and consequently that the J canonical means are contained in a $K = \min(J - 1, p)$ dimensional subspace of the canonical space. The fact that each of the last $p - K$ eigenvalues of the two-sided eigenvalue problem is equal to zero, implies that the diagonal matrix $\mathbf{\Lambda}$ in

$$\mathbf{B}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda}$$

is of the following form:

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

From the form of the matrix $\mathbf{\Lambda}$ it can be derived that the last $p - K$ elements of the J canonical means are identical i.e.

$$(\bar{\mathbf{x}}^1)'\mathbf{m}_{(j)} = (\bar{\mathbf{x}}^2)'\mathbf{m}_{(j)}, \dots, (\bar{\mathbf{x}}^J)'\mathbf{m}_{(j)} \quad \forall j \in [K + 1 : p].$$

A derivation of this result, very similar to a derivation in Gardner (2001), is provided below:

$$\mathbf{B}\mathbf{M} = \mathbf{W}\mathbf{M}\mathbf{\Lambda}$$

$$\longrightarrow \mathbf{M}'\overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}}\mathbf{M} = \mathbf{\Lambda} \quad (4.10.1)$$

$$\longrightarrow \overline{\mathbf{Y}}'\mathbf{N}\overline{\mathbf{Y}} = \mathbf{\Lambda}$$

$$\longrightarrow \sum_{j=1}^J n_j \overline{\mathbf{y}}^j (\overline{\mathbf{y}}^j)' = \mathbf{\Lambda} \quad (4.10.2)$$

$$\longrightarrow \sum_{j=1}^J n_j \begin{bmatrix} \overline{\mathbf{y}}^{j(1)} (\overline{\mathbf{y}}^{j(1)})' & \overline{\mathbf{y}}^{j(1)} (\overline{\mathbf{y}}^{j(2)})' \\ \overline{\mathbf{y}}^{j(2)} (\overline{\mathbf{y}}^{j(1)})' & \overline{\mathbf{y}}^{j(2)} (\overline{\mathbf{y}}^{j(2)})' \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (4.10.3)$$

where $\overline{\mathbf{Y}} = \overline{\mathbf{X}}\mathbf{M}$ and $\overline{\mathbf{y}}^{j(1)}$ is the $K \times 1$ vector with i th element equal to the i th element of the vector $\overline{\mathbf{y}}^j$ for $i \in [1 : K]$ and $\overline{\mathbf{y}}^{j(2)}$ is the $(p - K) \times 1$ vector with i th element equal to the $(K + i)$ th element of the vector $\overline{\mathbf{y}}^j$ for $i \in [1 : p - K]$. It is evident that

$$\sum_{j=1}^J n_j \overline{\mathbf{y}}^{j(2)} (\overline{\mathbf{y}}^{j(2)})' = \mathbf{0}. \quad (4.10.4)$$

Let \mathbf{a}_j be defined as:

$$\mathbf{a}_j = \sqrt{n_j} \overline{\mathbf{y}}^{j(2)}. \quad (4.10.5)$$

Given (4.10.5), equation (4.10.4) can now be expressed as

$$\begin{aligned} & \sum_{j=1}^J \mathbf{a}_j \mathbf{a}_j' = \mathbf{0} \\ \longrightarrow & \sum_{j=1}^J \begin{bmatrix} a_{j1}^2 & a_{j1}a_{j2} & \cdots & a_{j1}a_{j(p-K)} \\ a_{j2}a_{j1} & a_{j2}^2 & \cdots & a_{j2}a_{j(p-K)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{jp}a_{j1} & a_{jp}a_{j2} & \cdots & a_{jp}^2 \end{bmatrix} = \mathbf{0}. \end{aligned}$$

is of the following form:

$$\Lambda = \begin{bmatrix} \Lambda_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Let \mathbf{z}'_i denote the point, $\mathbf{x}'_i \mathbf{M}$, $(\bar{\mathbf{z}}^j)'$ denote the point, $(\bar{\mathbf{x}}^j)' \mathbf{M}$, and let \mathbf{d}' denote the vector, $\frac{1}{J} \mathbf{1}' \bar{\mathbf{X}} \mathbf{M}$, $i \in 1 : n_j$, $j \in 1 : J$. The equation

$$\begin{aligned} \mathbf{M}' \bar{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J} \mathbf{1} \mathbf{1}' \right) \bar{\mathbf{X}} \mathbf{M} &= \Lambda \\ \mathbf{M}' \left(\bar{\mathbf{X}} - \frac{1}{J} \mathbf{1} \mathbf{1}' \bar{\mathbf{X}} \right)' \left(\bar{\mathbf{X}} - \frac{1}{J} \mathbf{1} \mathbf{1}' \bar{\mathbf{X}} \right) \mathbf{M} &= \Lambda \end{aligned}$$

can now be rewritten as

$$\begin{aligned} \sum_{i=1}^J (\bar{\mathbf{z}}^i - \mathbf{d}) (\bar{\mathbf{z}}^i - \mathbf{d})' &= \Lambda \\ \sum_{i=1}^J \begin{bmatrix} (\bar{\mathbf{z}}^{i(1)} - \mathbf{d}^{(1)}) (\bar{\mathbf{z}}^{i(1)} - \mathbf{d}^{(1)})' & (\bar{\mathbf{z}}^{i(1)} - \mathbf{d}^{(1)}) (\bar{\mathbf{z}}^{i(2)} - \mathbf{d}^{(2)})' \\ (\bar{\mathbf{z}}^{i(2)} - \mathbf{d}^{(2)}) (\bar{\mathbf{z}}^{i(1)} - \mathbf{d}^{(1)})' & (\bar{\mathbf{z}}^{i(2)} - \mathbf{d}^{(2)}) (\bar{\mathbf{z}}^{i(2)} - \mathbf{d}^{(2)})' \end{bmatrix} &= \begin{bmatrix} \Lambda_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \end{aligned}$$

It is evident that

$$\sum_{i=1}^J (\bar{\mathbf{z}}^{i(2)} - \mathbf{d}^{(2)}) (\bar{\mathbf{z}}^{i(2)} - \mathbf{d}^{(2)})' = \mathbf{0}. \quad (4.10.7)$$

Let \mathbf{a}_i be defined as:

$$\mathbf{a}_i = \bar{\mathbf{z}}^{i(2)} - \mathbf{d}^{(2)}.$$

Equation (4.10.7) can then be rewritten as

$$\sum_{i=1}^J \mathbf{a}_i \mathbf{a}_i' = \mathbf{0}$$

$$\longrightarrow \sum_{i=1}^J \begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \dots & a_{i1}a_{i(p-K)} \\ a_{i2}a_{i1} & a_{i2}^2 & \dots & a_{i2}a_{i(p-K)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{ip}a_{i1} & a_{ip}a_{i2} & \dots & a_{ip}^2 \end{bmatrix} = \mathbf{0}.$$

It follows that

$$\sum_{i=1}^J a_{ik}^2 = 0 \quad \forall k \in [1 : p - K]$$

which in turn implies that

$$\begin{aligned} a_{ik} &= 0, \quad i \in [1 : J], \quad k \in [1 : p - K] \\ \longrightarrow (\bar{z}_j^i - d_j) &= 0 \quad \forall i \in [1 : J], \quad j \in [K + 1 : p]. \end{aligned}$$

It is evident that the J equations,

$$\begin{aligned} \bar{z}_j^1 - d_j &= 0 \\ \bar{z}_j^2 - d_j &= 0 \\ &\vdots \\ \bar{z}_j^J - d_j &= 0 \end{aligned} \tag{4.10.8}$$

hold for every $j \in [K + 1 : p]$. Subtracting the J th of the above equations from each of the previous $J - 1$ equations yields the following $J - 1$ equations:

$$\begin{aligned} \bar{z}_j^1 - \bar{z}_j^J &= 0 \\ \bar{z}_j^2 - \bar{z}_j^J &= 0 \\ &\vdots \\ \bar{z}_j^{(J-1)} - \bar{z}_j^J &= 0. \end{aligned}$$

It is evident that $\bar{z}_j^1 = \bar{z}_j^2 = \dots = \bar{z}_j^J \quad \forall j \in [K + 1 : p]$ i.e.

Chapter 5 - Quality of the CVA biplot

In Chapter 4 the construction and interpretation of the CVA biplot were discussed in detail. The relationships and predictions suggested by the CVA biplot are however not guaranteed to be valid. Knowledge about the quality of the different aspects of the CVA biplot is required in order to assess to what extent these relationships and predictions are representative of reality. A number of quality measures associated with the CVA biplot will be studied in this chapter. All of these quality measures are defined as ratios of sums of squared values. The orthogonality properties which underlie the validity of these ratios as quality measures, are outlined in Section 5.1.

5.1 Orthogonality properties underlying a CVA biplot

Recall from Chapter 4 that the construction of the r -dimensional CVA biplot is based on the best rank r (least squares) approximation to the transformed matrix of group centroids, $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$. Letting the svd of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ be given by

$$\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

the best rank r approximation to $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ is given by

$$\begin{aligned} \mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}_r' &= \mathbf{C}^{1/2}\widehat{\bar{\mathbf{X}}}\mathbf{L} & (5.1.1) \\ \text{where } \widehat{\bar{\mathbf{X}}} &= \bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}_r'. \end{aligned}$$

Consider the identity

$$\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L} = \mathbf{C}^{1/2}\widehat{\bar{\mathbf{X}}}\mathbf{L} + \mathbf{C}^{1/2}(\bar{\mathbf{X}} - \widehat{\bar{\mathbf{X}}})\mathbf{L}. \quad (5.1.2)$$

Since the matrix \mathbf{V}_r in equation (5.1.1) is an orthonormal matrix, the decomposition

of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ in (5.1.2) exhibits Type A orthogonality (see Section 3.1), that is:

$$\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}}'\mathbf{C}^{1/2} = \mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}\mathbf{L}'\widehat{\mathbf{X}}'\mathbf{C}^{1/2} + \mathbf{C}^{1/2}(\bar{\mathbf{X}} - \widehat{\mathbf{X}})\mathbf{L}\mathbf{L}'(\bar{\mathbf{X}} - \widehat{\mathbf{X}})'\mathbf{C}^{1/2}$$

or equivalently, the decomposition in (5.1.2) exhibits Type A orthogonality in the metric \mathbf{W} (Gardner-Lubbe *et al.*, 2008), that is:

$$\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{W}^{-1}\bar{\mathbf{X}}'\mathbf{C}^{1/2} = \mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}'\mathbf{C}^{1/2} + \mathbf{C}^{1/2}(\bar{\mathbf{X}} - \widehat{\mathbf{X}})\mathbf{W}^{-1}(\bar{\mathbf{X}} - \widehat{\mathbf{X}})'\mathbf{C}^{1/2}.$$

Given that the column vectors of the matrix \mathbf{V}_r in equation (5.1.1) are the first r right singular vectors of the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$, the decomposition in (5.1.2) exhibits Type B orthogonality (see Section 3.1), that is:

$$\mathbf{L}'\bar{\mathbf{X}}'\mathbf{C}\bar{\mathbf{X}}\mathbf{L} = \mathbf{L}'\widehat{\mathbf{X}}'\mathbf{C}\widehat{\mathbf{X}}\mathbf{L} + \mathbf{L}'(\bar{\mathbf{X}} - \widehat{\mathbf{X}})'\mathbf{C}(\bar{\mathbf{X}} - \widehat{\mathbf{X}})\mathbf{L}. \quad (5.1.3)$$

Multiplying equation (5.1.3) by $(\mathbf{L}')^{-1}$ from the left and by \mathbf{L}^{-1} from the right yields the equality,

$$\bar{\mathbf{X}}'\mathbf{C}\bar{\mathbf{X}} = \widehat{\mathbf{X}}'\mathbf{C}\widehat{\mathbf{X}} + (\bar{\mathbf{X}} - \widehat{\mathbf{X}})'\mathbf{C}(\bar{\mathbf{X}} - \widehat{\mathbf{X}})$$

indicating that the decomposition,

$$\mathbf{C}^{1/2}\bar{\mathbf{X}} = \mathbf{C}^{1/2}\widehat{\mathbf{X}} + \mathbf{C}^{1/2}(\bar{\mathbf{X}} - \widehat{\mathbf{X}}) \quad (5.1.4)$$

exhibits Type B orthogonality.

The fact that the decomposition in (5.1.2) exhibits Type A and Type B orthogonality validates all the quality measures concerning the canonical variables that will be discussed in this chapter. On the other hand, the fact that the decomposition in (5.1.4) exhibits Type A orthogonality in the metric \mathbf{W} as well as Type B orthogonality, validates all the quality measures concerning the measured variables that will be discussed.

5.2 The overall quality of the CVA biplot

5.2.1 The overall quality of the CVA biplot with respect to the canonical variables

5.2.1.1 Definition and properties

Gabriel (1972) proposed that the overall quality of the CVA biplot, like the overall quality of the PCA biplot, be measured by the ratio of fitted to total sums of squares associated with the matrix upon which the construction of the biplot is based. The expression of the overall quality of the r -dimensional CVA biplot is therefore given by

$$\begin{aligned}\Omega_{Can.var} &= \frac{\text{tr} \left\{ \mathbf{L}' \widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}} \mathbf{L} \right\}}{\text{tr} \left\{ \mathbf{L}' \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \mathbf{L} \right\}} \\ &= \frac{\text{tr} \left\{ \mathbf{V}_r \mathbf{\Lambda}_r \mathbf{V}_r' \right\}}{\text{tr} \left\{ \mathbf{V} \mathbf{\Lambda} \mathbf{V}' \right\}} \\ &= \frac{\sum_{j=1}^r \lambda_j}{\sum_{j=1}^K \lambda_j}\end{aligned}\tag{5.2.1}$$

where $\mathbf{\Lambda} = \mathbf{D}_p^2$ and $K = \min(J-1, p)$. Gower *et al.* (2011) refer to this ratio as the overall quality of the CVA biplot with respect to the canonical variables.

Note that the definitions of the overall quality with respect to the canonical variables corresponding to the three choices of \mathbf{C} are based on the orthogonal decompositions of three different between-group sums of squares. The overall quality with respect to the canonical variables corresponding to $\mathbf{C} = \mathbf{N}$ and $\mathbf{C} = \mathbf{I}$ are based on the orthogonal decompositions of $\text{tr} \left(\mathbf{L}' \overline{\mathbf{X}}' \mathbf{N} \overline{\mathbf{X}} \mathbf{L} \right)$ and $\text{tr} \left(\mathbf{L}' \overline{\mathbf{X}}' \overline{\mathbf{X}} \mathbf{L} \right)$ respectively, which are both between-groups sums of squares associated with the canonical observations that are corrected for the overall mean of the n canonical samples (which is $\mathbf{0}$) or equivalently, corrected for the weighted mean of the J canonical means where the weight of the j th canonical mean is proportional to n_j . The overall quality with respect to the canonical variables corresponding to $\mathbf{C} = \left(\mathbf{I} - \frac{1}{J} \mathbf{1} \mathbf{1}' \right)$ on the other hand is based on the decomposition of the between-groups sums of squares associated with the canonical observations corrected for the unweighted mean of the J canonical means, $\text{tr} \left(\mathbf{L}' \overline{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{J} \mathbf{1} \mathbf{1}' \right) \overline{\mathbf{X}} \mathbf{L} \right)$.

Recall from Chapter 4 that

$$\lambda_j = \frac{\mathbf{m}'_{(j)} \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \mathbf{m}_{(j)}}{\mathbf{m}'_{(j)} \mathbf{W} \mathbf{m}_{(j)}}$$

where $\mathbf{M} = \mathbf{L} \mathbf{V}$

and \mathbf{V} is the matrix of right singular vectors of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$, and hence λ_j measures the extent to which the groups are separated in the one-dimensional subspace of the p -dimensional canonical space that is spanned by the j th right singular vector of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$. This implies that the overall quality with respect to the canonical variables of an r -dimensional CVA biplot measures the extent to which the groups are separated in that r -dimensional CVA biplot compared to in the K -dimensional subspace of the canonical space that perfectly contains the K group centroids, \mathbb{C}^K . Since the J groups are maximally separated in \mathbb{C}^K , $\Omega_{Can.var}$ measures the quality of the separation of the groups in the r -dimensional CVA biplot.

Note that since the between-groups sample covariance matrix associated with the vector of canonical variables is proportional to $\text{tr}(\mathbf{L}'\overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}}\mathbf{L})$, $\Omega_{Can.var}$ can be interpreted as the proportion of the total between-groups variability associated with the vector of canonical variables that is accounted for in the CVA biplot when $\mathbf{C} = \mathbf{N}$. The fact that the decomposition of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ into $\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}$ and $\mathbf{C}^{1/2}(\overline{\mathbf{X}} - \widehat{\mathbf{X}})\mathbf{L}$ exhibits Type B orthogonality validates $\Omega_{Can.var}$ as a quality measure. Note that $\Omega_{Can.var}$ can also be expressed as

$$\Omega_{Can.var} = \frac{\text{tr}(\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}\mathbf{L}'\widehat{\mathbf{X}}'\mathbf{C}^{1/2})}{\text{tr}(\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}\mathbf{L}'\overline{\mathbf{X}}'\mathbf{C}^{1/2})}.$$

This indicates that the Type A orthogonality property of the decomposition of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ in equation (5.1.2) also ensures that $\Omega_{Can.var}$ is meaningful as a quality measure.

Being defined as a ratio of sums of squared values, $\Omega_{Can.var}$ can only take on non-negative values. From the expression of $\Omega_{Can.var}$ given in (5.2.1) it is evident that $\Omega_{Can.var}$ is a non-decreasing function of the dimension of the CVA biplot space, r , and has a maximum value of one which it will necessary obtain when $r = K$. Since $\lambda_j > 0 \forall j \in [1 : K]$, $\Omega_{Can.var}$ will obtain its maximum value of one if and only $r = K$. This fact can also be derived as follows:

$$\begin{aligned} \Omega_{Can.var} &= \frac{\text{tr}\{\mathbf{L}'\widehat{\mathbf{X}}'\mathbf{C}\widehat{\mathbf{X}}\mathbf{L}\}}{\text{tr}\{\mathbf{L}'\overline{\mathbf{X}}'\mathbf{C}\overline{\mathbf{X}}\mathbf{L}\}} \\ \rightarrow \Omega_{Can.var} &= 1 - \frac{\text{tr}\left\{\left(\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}\right)' \left(\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}\right)\right\}}{\text{tr}\{\mathbf{L}'\overline{\mathbf{X}}'\mathbf{C}\overline{\mathbf{X}}\mathbf{L}\}} \\ \rightarrow \Omega_{Can.var} &= 1 - \frac{\sum_{j=1}^J \sum_{i=1}^p \left[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}\right]_{ji}^2}{\sum_{j=1}^J \sum_{i=1}^p \left[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}\right]_{ji}^2} \end{aligned}$$

$$\begin{aligned}
\longrightarrow \Omega_{Car.var} = 1 &\iff \frac{\sum_{j=1}^J \sum_{i=1}^p \left[\mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right]_{ji}^2}{\sum_{j=1}^J \sum_{i=1}^p \left[\mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} \right]_{ji}^2} = 0 \\
\therefore \Omega_{Car.var} = 1 &\iff \sum_{j=1}^J \sum_{i=1}^p \left[\mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right]_{ji}^2 = 0 \\
\therefore \Omega_{Car.var} = 1 &\iff \left[\mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right]_{ji} = 0 \quad \forall j \in [1 : J], i \in [1 : p] \\
\therefore \Omega_{Car.var} = 1 &\iff \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} = \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} \\
\therefore \Omega_{Car.var} = 1 &\iff r = K.
\end{aligned}$$

The overall quality of the CVA biplot, $\Omega_{Can.var}$, has a minimum value of zero which it will attain if and only if $\text{tr} \left\{ \mathbf{L}' \widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}}\mathbf{L} \right\} = 0$ i.e.

$$\begin{aligned}
\Omega_{Can.var} = 0 &\iff \sum_{j=1}^J \sum_{i=1}^p \left[\mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right]_{ji}^2 = 0 \\
\therefore \Omega_{Can.var} = 0 &\iff \left[\mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right]_{ji}^2 = 0 \quad \forall j \in [1 : J], i \in [1 : p] \\
\therefore \Omega_{Can.var} = 0 &\iff \mathbf{e}_j' \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \in \mathcal{V}^\perp(\mathbf{V}_r) \quad \forall j \in [1 : J].
\end{aligned}$$

This is however only possible if $r = 0$ and hence $\Omega_{Can.var}$ will never attain the value zero. This is also evident from the fact that even when $r = 1$, $\Omega_{Can.var} = \frac{\lambda_1}{\sum_{j=1}^K \lambda_j} > 0$.

5.2.1.2 Scale invariance

Recall from Chapter 4 that if \mathbf{A} denotes the $p \times p$ diagonal matrix with i th diagonal element equal to the sample standard deviation of the i th measured variable, x_i , then the matrices $\bar{\mathbf{X}}$, \mathbf{L} , \mathbf{M} , \mathbf{M}^{-1} , $\mathbf{\Lambda}$, \mathbf{W} , \mathbf{B} and $\widehat{\mathbf{X}}$, corresponding to the standardised measurements, $\mathbf{X}^* = \mathbf{X}\mathbf{A}^{-1}$, are given by

$$\begin{aligned}
\bar{\mathbf{X}}^* &= \bar{\mathbf{X}}\mathbf{A}^{-1}, \\
\mathbf{L}^* &= \mathbf{A}\mathbf{L}, \\
\mathbf{M}^* &= \mathbf{A}\mathbf{M}, \\
\mathbf{M}^{-1*} &= \mathbf{M}^{-1}\mathbf{A}^{-1}, \\
\mathbf{\Lambda}^* &= \mathbf{\Lambda}, \\
\mathbf{W}^* &= \mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-1}, \\
\mathbf{B}^* &= \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} \\
\text{and } \widehat{\mathbf{X}}^* &= \bar{\mathbf{X}}^* \mathbf{M}_r^* \mathbf{M}^{r*} = \widehat{\mathbf{X}}\mathbf{A}^{-1}
\end{aligned}$$

respectively. Due to the fact that $\mathbf{\Lambda}^* = \mathbf{\Lambda}$, the overall quality of the r -dimensional CVA biplot with respect to the canonical variables constructed from the standardised measurements,

$$\Omega_{Can.var}^* = \frac{\sum_{j=1}^r \lambda_j^*}{\sum_{j=1}^K \lambda_j^*},$$

is equal to the overall quality of the r -dimensional CVA biplot with respect to the canonical variables constructed from the unstandardised measurements, $\Omega_{Can.var}$. Substituting any $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in the above equations shows that the overall quality of the CVA biplot with respect to the canonical variables is invariant to all non-singular linear transformations of the scales of the measurements of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$.

5.2.2 The overall quality of the CVA biplot with respect to the original variables

5.2.2.1 Definition and properties

Given that the interest of the investigator of a CVA biplot typically lies with the measured variables, he/she is typically interested in the overall accuracy of the approximations to the elements of the matrix $\bar{\mathbf{X}}$ rather than the overall accuracy of the approximations to the elements of the matrix $\bar{\mathbf{X}}\mathbf{L}$. The quality measure that assesses the former set of approximations is called the overall quality of the CVA biplot with respect to the original measured variables and is defined as:

$$\Omega_{Orig.var} = \frac{\text{tr}\left(\widehat{\bar{\mathbf{X}}}' \mathbf{C} \widehat{\bar{\mathbf{X}}}\right)}{\text{tr}\left(\bar{\mathbf{X}}' \mathbf{C} \bar{\mathbf{X}}\right)} \quad (5.2.2)$$

(Gower *et al.*, 2011). The fact that the decomposition of $\mathbf{C}^{1/2}\bar{\mathbf{X}}$ into $\mathbf{C}^{1/2}\widehat{\bar{\mathbf{X}}}$ and $\mathbf{C}^{1/2}\left(\bar{\mathbf{X}} - \widehat{\bar{\mathbf{X}}}\right)$ exhibits Type B orthogonality validates $\Omega_{Orig.var}$ as a quality measure.

Note that the definitions of the overall quality with respect to the original variables corresponding to the three choices of \mathbf{C} are based on the orthogonal decompositions of three different between-group sums of squares. The overall quality with respect to the original variables corresponding to $\mathbf{C} = \mathbf{N}$ and $\mathbf{C} = \mathbf{I}$ are based on the orthogonal decompositions of $\text{tr}\left(\bar{\mathbf{X}}' \mathbf{N} \bar{\mathbf{X}}\right)$ and $\text{tr}\left(\bar{\mathbf{X}}' \bar{\mathbf{X}}\right)$ respectively, which are both between-groups sums of squares that are corrected for the overall mean of the n individual samples (which is $\mathbf{0}$) or equivalently, corrected for the weighted mean of the J group centroids where the weight of the j th group centroid is proportional to n_j . The overall quality with respect to the original variables corresponding to

$\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$ on the other hand is based on the decomposition of the between-groups sums of squares corrected for the unweighted mean of the J group centroids, $\text{tr}(\bar{\mathbf{X}}'(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}})$. It follows that for both $\mathbf{C} = \mathbf{N}$ and $\mathbf{C} = \mathbf{I}$ the overall quality with respect to the original variables is equal to the proportion of the between-groups sample variance that is account for in the CVA biplot, where variance is interpreted as deviation from the overall mean of the n individual samples (or equivalently the weighted mean of the J group centroids) i.e. $\mathbf{0}$. Note however that the between-group sample variance is measured by different quantities for $\mathbf{C} = \mathbf{N}$ and $\mathbf{C} = \mathbf{I}$. For $\mathbf{C} = \mathbf{N}$, the contributions of the J groups to the between-groups variance are weighted by their sizes while for $\mathbf{C} = \mathbf{I}$ the contributions of the J groups are equally weighted. When $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$, the overall quality with respect to the original variables is equal to the proportion of the between-groups sample variance accounted for in the CVA biplot, where variance is interpreted as deviation from the unweighted centroid of the J group means.

Substituting $\bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}^r$ and $\bar{\mathbf{X}}\mathbf{M}\mathbf{M}^{-1}$ for $\widehat{\bar{\mathbf{X}}}$ and $\bar{\mathbf{X}}$ respectively in equation (5.2.2) yields the following expression for $\Omega_{Orig.var}$:

$$\begin{aligned}
 \Omega_{Orig.var} &= \frac{\text{tr}\left\{(\mathbf{M}^r)'\mathbf{M}'_r\bar{\mathbf{X}}'\mathbf{C}\bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}^r\right\}}{\text{tr}\left\{(\mathbf{M}^{-1})'\mathbf{M}'\bar{\mathbf{X}}'\mathbf{C}\bar{\mathbf{X}}\mathbf{M}\mathbf{M}^{-1}\right\}} \\
 &= \frac{\text{tr}\left\{(\mathbf{M}^r)'\mathbf{\Lambda}_r\mathbf{M}^r\right\}}{\text{tr}\left\{(\mathbf{M}^{-1})'\mathbf{\Lambda}\mathbf{M}^{-1}\right\}} \\
 &= \frac{\text{tr}\left\{\mathbf{M}^r(\mathbf{M}^r)'\mathbf{\Lambda}_r\right\}}{\text{tr}\left\{\mathbf{M}^{-1}(\mathbf{M}^{-1})'\mathbf{\Lambda}\right\}} \\
 &= \frac{\text{tr}\left\{\mathbf{M}^r(\mathbf{M}^r)'\mathbf{\Lambda}_r\right\}}{\text{tr}\left\{(\mathbf{M}'\mathbf{M})^{-1}\mathbf{\Lambda}\right\}} \\
 &= \frac{\sum_{j=1}^r \lambda_j [(\mathbf{M}'\mathbf{M})^{-1}]_{jj}}{\sum_{j=1}^p \lambda_j [(\mathbf{M}'\mathbf{M})^{-1}]_{jj}} \\
 \Omega_{Orig.var} &= \frac{\sum_{j=1}^r \lambda_j [(\mathbf{M}'\mathbf{M})^{-1}]_{jj}}{\sum_{j=1}^K \lambda_j [(\mathbf{M}'\mathbf{M})^{-1}]_{jj}}. \tag{5.2.3}
 \end{aligned}$$

The last step in the above derivation follows since the last $p - K$ singular values of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ are equal to zero (see Chapter 4). Note that

$$[(\mathbf{M}'\mathbf{M})^{-1}]_{jj} = [\mathbf{M}^{-1}(\mathbf{M}^{-1})']_{jj} = 0 \iff [\mathbf{M}^{-1}]_{ji} = 0 \quad \forall i \in [1:p], j \in [1:p].$$

However, since \mathbf{M}^{-1} is a non-singular matrix, none of its row vectors can be equal

to the null vector. Since

$$\lambda_k > 0 \quad \forall k \in [1 : K]$$

$$\text{and } \left[\mathbf{M}^{-1} (\mathbf{M}^{-1})' \right]_{jj} > 0 \quad \forall j \in [1 : p]$$

$\Omega_{Orig.var}$ can only take on positive values and is an increasing function of r . Hence, $\Omega_{Orig.var}$ will necessarily attain its maximum value if $r = K$. It is evident from equation (5.2.3) that the maximum value of $\Omega_{Orig.var}$ is one. The following derivation shows that $r = K$ is not only a sufficient condition for Ω to attain the value of one, but also a necessary condition:

$$\Omega_{Orig.var} = \frac{\text{tr} \left\{ \widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}} \right\}}{\text{tr} \left\{ \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \right\}}$$

$$\rightarrow \Omega_{Orig.var} = 1 - \frac{\text{tr} \left\{ \left(\overline{\mathbf{X}}' \mathbf{C}^{1/2} - \widehat{\mathbf{X}}' \mathbf{C}^{1/2} \right) \left(\mathbf{C}^{1/2} \overline{\mathbf{X}} - \mathbf{C}^{1/2} \widehat{\mathbf{X}} \right) \right\}}{\text{tr} \left\{ \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \right\}}$$

$$\therefore \Omega_{Orig.var} = 1 \leftrightarrow \mathbf{C}^{1/2} \widehat{\mathbf{X}} = \mathbf{C}^{1/2} \overline{\mathbf{X}}$$

$$\therefore \Omega_{Orig.var} = 1 \leftrightarrow r = K .$$

5.2.2.2 Scale dependence

When the CVA biplot is constructed from the standardised measurements, $\mathbf{X}^* = \mathbf{X}\mathbf{A}$, the expression for the overall quality of the CVA biplot with respect to the original measured variables is given by

$$\Omega_{Orig.var}^* = \frac{\text{tr} \left\{ \left(\widehat{\mathbf{X}}^* \right)' \mathbf{C} \widehat{\mathbf{X}}^* \right\}}{\text{tr} \left\{ \left(\overline{\mathbf{X}}^* \right)' \mathbf{C} \overline{\mathbf{X}}^* \right\}}$$

$$\rightarrow \Omega_{Orig.var}^* = \frac{\text{tr} \left\{ \mathbf{A}^{-1} \widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}} \mathbf{A}^{-1} \right\}}{\text{tr} \left\{ \mathbf{A}^{-1} \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \mathbf{A}^{-1} \right\}} .$$

It is evident that unlike the overall quality with respect to the canonical variables, the overall quality with respect to the original variables corresponding to the CVA biplot constructed from the standardised measurements is not identical to that corresponding to the CVA biplot constructed from the unstandardised measurements.

Substituting an arbitrary $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in the expressions above shows that the overall quality with respect to the original variables is not invariant to all non-singular linear transformations of the matrix \mathbf{X} of the form $\mathbf{X} \rightarrow \mathbf{XF}$.

5.3 Adequacies

5.3.1 Definition and properties

By an argument similar to that used to define the adequacy of the k th biplot axis of the r -dimensional PCA biplot, the adequacy of the k th biplot axis of the r -dimensional CVA biplot, γ_k , is defined as the ratio of the square of the length representing one unit of the k th measured variable in the r -dimensional CVA biplot space to the square of the length representing one unit of the k th measured variable in the p -dimensional CVA biplot space (Gardner (2001), Gardner-Lubbe *et al.* (2008), Gower *et al.* (2011)), that is

$$\gamma_k = \frac{\|\mathbf{e}'_k \mathbf{M}_r\|^2}{\|\mathbf{e}'_k \mathbf{M}\|^2} \quad (5.3.1)$$

$$\begin{aligned} \rightarrow \gamma_k &= \frac{[\mathbf{M}_r \mathbf{M}'_r]_{kk}}{[\mathbf{M} \mathbf{M}']_{kk}} \\ \rightarrow \gamma_k &= \frac{[\mathbf{M}_r \mathbf{M}'_r]_{kk}}{[\mathbf{W}^{-1}]_{kk}} \\ \rightarrow \gamma_k &= \frac{\sum_{i=1}^r ([\mathbf{M}]_{ki})^2}{[\mathbf{W}^{-1}]_{kk}}. \end{aligned} \quad (5.3.2)$$

The p -component vector with k th element equal to the k th adequacy associated with the r -dimensional CVA biplot follows as

$$\begin{aligned} \gamma &= \text{diag}(\mathbf{M}_r \mathbf{M}'_r) [\text{diag}(\mathbf{M} \mathbf{M}')]^{-1} \\ \rightarrow \gamma &= \text{diag}(\mathbf{M}_r \mathbf{M}'_r) [\text{diag}(\mathbf{W}^{-1})]^{-1}. \end{aligned}$$

The k th adequacy can also be expressed as:

$$\gamma_k = \frac{\|\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r\|^2}{\|\mathbf{e}'_k \mathbf{L}\|^2} \quad (5.3.3)$$

$$\rightarrow \gamma_k = \frac{[\mathbf{L} \mathbf{V}_r \mathbf{V}'_r \mathbf{L}']_{kk}}{[\mathbf{L} \mathbf{L}']_{kk}}. \quad (5.3.4)$$

Note that since the matrix \mathbf{V} is an orthogonal matrix, the decomposition of \mathbf{L} into $\mathbf{LV}_r\mathbf{V}_r'$ and $(\mathbf{L} - \mathbf{LV}_r\mathbf{V}_r')$ exhibits Type A orthogonality, that is

$$\mathbf{LL}' = (\mathbf{LV}_r\mathbf{V}_r')(\mathbf{LV}_r\mathbf{V}_r')' + (\mathbf{L} - \mathbf{LV}_r\mathbf{V}_r')(\mathbf{L} - \mathbf{LV}_r\mathbf{V}_r')'. \quad (5.3.5)$$

It is evident from the expression of γ_k in (5.3.4), that equation (5.3.5) validates γ_k as a quality measure, $k \in [1 : p]$.

Consider the expression of the k th adequacy provided in (5.3.3). Note that the k th adequacy associated with the r -dimensional CVA biplot is equal to the square of the length of the orthogonal projection of $\mathbf{e}'_k\mathbf{L}$ onto the r -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_r)$, to the square of the length of $\mathbf{e}'_k\mathbf{L}$. This implies that the k th adequacy has a minimum value of zero which it will attain if and only if $\mathbf{e}'_k\mathbf{L}$ lies orthogonal to $\mathcal{V}(\mathbf{V}_r)$, and a maximum value of one which it will attain if and only if $\mathbf{e}'_k\mathbf{L}$ lies in $\mathcal{V}(\mathbf{V}_r)$. It follows that $r = p$ is necessary for all p the biplot axes to have unit adequacy. Note that the j th adequacy associated with the K -dimensional weighted CVA biplot and that associated with each of the two K -dimensional unweighted CVA biplots are identical due to the fact that the row spaces of $\mathbf{N}^{1/2}\overline{\mathbf{X}}\mathbf{L}$, $\overline{\mathbf{X}}\mathbf{L}$ and $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')\overline{\mathbf{X}}\mathbf{L}$ are identical (see Section 4.4) and hence that

$$\mathbf{V}_K^{\mathbf{N}}(\mathbf{V}_K^{\mathbf{N}})' = \mathbf{V}_K^{\mathbf{I}}(\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\mathbf{Cent}}(\mathbf{V}_K^{\mathbf{Cent}})'$$

It is shown below that the k th adequacy associated with the r -dimensional CVA biplot is a decreasing function of the angle between the k th interpolative biplot axis of the p -dimensional CVA biplot and the k th interpolative biplot axis of the r -dimensional CVA biplot:

$$\begin{aligned} \gamma_k &= \frac{\|\mathbf{e}'_k\mathbf{LV}_r\mathbf{V}_r'\|^2}{\|\mathbf{e}'_k\mathbf{L}\|^2} \\ &= \frac{\mathbf{e}'_k\mathbf{LV}_r\mathbf{V}_r'\mathbf{L}'\mathbf{e}_k}{\|\mathbf{e}'_k\mathbf{L}\|^2} \\ &= \left\{ \frac{(\mathbf{e}'_k\mathbf{LV}_r\mathbf{V}_r'\mathbf{L}'\mathbf{e}_k)^{1/2}}{\|\mathbf{e}'_k\mathbf{L}\|} \right\}^2 \\ &= \left\{ \frac{\mathbf{e}'_k\mathbf{LV}_r\mathbf{V}_r'\mathbf{L}'\mathbf{e}_k}{\|\mathbf{e}'_k\mathbf{LV}_r\mathbf{V}_r'\|\|\mathbf{e}'_k\mathbf{L}\|} \right\}^2 \\ &\rightarrow \gamma_k = \cos^2(\theta_k) \end{aligned}$$

where θ_k denotes the angle between the vectors $\mathbf{e}'_k \mathbf{L}$ and $\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$. Note that the coordinate vectors of $\mathbf{e}'_k \mathbf{L}$ and $\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$ relative to the basis of the p -dimensional CVA biplot space given by the column vectors of \mathbf{V} , are given by $\mathbf{e}'_k \mathbf{L} \mathbf{V} = \mathbf{e}'_k \mathbf{M}$ and $\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r \mathbf{V} = [\mathbf{e}'_k \mathbf{M}_r \quad \mathbf{0}']$ respectively. It follows that γ_k is a decreasing function of the size of the angle between the k th interpolative biplot axis of the p -dimensional CVA biplot and the k th interpolative biplot axis of the r -dimensional CVA biplot.

Given that the decomposition of \mathbf{L} into $\mathbf{L} \mathbf{V}_r \mathbf{V}'_r$ and $(\mathbf{L} - \mathbf{L} \mathbf{V}_r \mathbf{V}'_r)$ exhibits Type A orthogonality, it follows that the k th adequacy associated with the r -dimensional CVA biplot is a decreasing function of the Pythagorean distance between the two points, $\mathbf{e}'_k \mathbf{L}$ and $\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$:

$$\gamma_k = 1 - \frac{\|\mathbf{e}'_k \mathbf{L} - \mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r\|^2}{\|\mathbf{e}'_k \mathbf{L}\|^2}. \quad (5.3.6)$$

Equation (5.3.6) shows that γ_k is a decreasing function of the Pythagorean distance between the point representing one unit of the k th variable in the p -dimensional canonical space, $\mathbf{e}'_k \mathbf{L}$, and the point representing one unit of the k th variable in the r -dimensional CVA biplot space, $\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$. It is important to note that since $\|\mathbf{e}'_k \mathbf{L}\|^2$ is not necessarily equal to $\|\mathbf{e}'_j \mathbf{L}\|^2$,

$$\gamma_k > \gamma_j$$

does not necessarily imply that

$$\|\mathbf{e}'_k \mathbf{L} - \mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r\| < \|\mathbf{e}'_j \mathbf{L} - \mathbf{e}'_j \mathbf{L} \mathbf{V}_r \mathbf{V}'_r\|.$$

It is evident from the expression of γ_k in equation (5.3.2) that γ_k is a non-decreasing function of the dimension of the CVA biplot space, r . This implies that the size of the angle between the vectors $\mathbf{e}'_k \mathbf{L}$ and $\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$, as well as the Pythagorean distance between the two points $\mathbf{e}'_k \mathbf{L}$ and $\mathbf{e}'_k \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$, are non-increasing functions of the dimension of the CVA biplot space.

5.3.2 Scale invariance

Substituting \mathbf{AM} for \mathbf{M}^* in the expression of the adequacy of the k th biplot axis of the r -dimensional CVA biplot constructed from the standardised measurements,

$$\gamma_k^* = \frac{\|\mathbf{e}'_k \mathbf{M}_r^*\|^2}{\|\mathbf{e}'_k \mathbf{M}^*\|^2} \quad (5.3.7)$$

shows that γ_k^* is identical to the adequacy of the k th biplot axis of the r -dimensional CVA biplot constructed from the unstandardised measurements, γ_k :

$$\begin{aligned} \gamma_k^* &= \frac{\|\mathbf{e}'_k \mathbf{A} \mathbf{M}_r\|^2}{\|\mathbf{e}'_k \mathbf{A} \mathbf{M}\|^2} \\ &= \frac{\|a_{kk} \mathbf{e}'_k \mathbf{M}_r\|^2}{\|a_{kk} \mathbf{e}'_k \mathbf{M}\|^2} \\ &= \frac{\|\mathbf{e}'_k \mathbf{M}_r\|^2}{\|\mathbf{e}'_k \mathbf{M}\|^2} \\ &\longrightarrow \gamma_k^* = \gamma_k . \end{aligned}$$

Note however that unlike the overall quality with respect to the canonical variables, the adequacy measure is not invariant to all non-singular linear transformations of the form

$$\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x} \quad (5.3.8)$$

where \mathbf{F} is an arbitrary $p \times p$ non-singular matrix. This is evident upon substituting $\mathbf{F}^{-1}\mathbf{M}$ for \mathbf{M}^* in (5.3.7). The adequacy measure is only invariant to transformations of the form in (5.3.8) when the matrix \mathbf{F} is a non-singular diagonal matrix.

5.4 Axis predictivities

5.4.1 Definition and properties

The axis predictivity of the k th CVA biplot axis is a measure of the overall accuracy of the approximations to the measurements of the J group centroids on the k th measured variable read off from the k th (predictive) biplot axis (Gardner-Lubbe *et al.* (2008), Gower *et al.* (2011)). Gower *et al.* (2011) defined the axis predictivity

of the k th predictive biplot axis (or k th axis predictivity) of the r -dimensional CVA biplot as:

$$\begin{aligned}
 \pi_k &= \frac{\left[\widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}} \right]_{kk}}{\left[\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \right]_{kk}} \\
 \rightarrow \pi_k &= \frac{\left[(\mathbf{M}^r)' \boldsymbol{\Lambda}_r \mathbf{M}^r \right]_{kk}}{\left[(\mathbf{M}^{-1})' \boldsymbol{\Lambda} \mathbf{M}^{-1} \right]_{kk}} \\
 \rightarrow \pi_k &= \frac{\sum_{i=1}^r \lambda_i \left([\mathbf{M}^{-1}]_{ik} \right)^2}{\sum_{i=1}^K \lambda_i \left([\mathbf{M}^{-1}]_{ik} \right)^2}. \tag{5.4.1}
 \end{aligned}$$

The Type B orthogonality property of the decomposition $\mathbf{C}^{1/2} \overline{\mathbf{X}}$ into $\mathbf{C}^{1/2} \widehat{\mathbf{X}}$ and $\mathbf{C}^{1/2} (\overline{\mathbf{X}} - \widehat{\mathbf{X}})$ validates π_k as a quality measure. Let the p -component vector with k th element equal to the k th axis predictivity be denoted by $\boldsymbol{\pi}$, then

$$\boldsymbol{\pi} = \text{diag} \left(\widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}} \right) \left[\text{diag} \left(\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \right) \right]^{-1}. \tag{5.4.2}$$

Being defined as a ratio of sums of non-negative values, the axis predictivity measure can only take on non-negative values. Since the numerator of π_k is a non-decreasing function of the dimension of the CVA biplot, r , while the denominator is fixed, π_k is a non-decreasing function of r . It follows that π_k will necessarily equal its maximum value when $r = p$. It is evident from equation (5.4.1) that the maximum value of π_k is one. Although $r = p$ is a sufficient condition for all p biplot axes to have unit axis predictivities, it is not a necessary condition for one, or even all, of the biplot axes to have unit axis predictivity. A condition which is sufficient as well as necessary for all p the biplot axes to have unit axis predictivities, is $r = K$:

$$\begin{aligned}
 \pi_k = 1 &\leftrightarrow \frac{\sum_{i=1}^r \lambda_i \left([\mathbf{M}^{-1}]_{ik} \right)^2}{\sum_{i=1}^K \lambda_i \left([\mathbf{M}^{-1}]_{ik} \right)^2} = 1 \\
 \therefore \pi_k = 1 &\leftrightarrow r = K.
 \end{aligned}$$

Since the axis predictivity of a CVA biplot axis is a measure of the overall accuracy of the approximations to the measurements of the J group centroids on the corresponding variable that are read off from that biplot axis, the dissimilarities (and similarities) between groups with respect to a measured variable that is reflected in a CVA biplot can only be trusted to be an accurate representation of reality when the axis predictivity of the corresponding biplot axis is close to one. Hence,

when investigating the dissimilarities and similarities between groups with respect to the individual measured variables, the positions of the points representing the corresponding group centroids in the CVA biplot relative to the individual biplot axes need to be considered together with the axis predictivities of the biplot axes.

Axis predictivities can also be defined for new biplot axes that have been interpolated onto an existing CVA biplot (Gower *et al.*, 2011). This does however not lie within the scope of this work. Details on this can be found in Gower *et al.* (2011).

5.4.2 The relationship of the axis predictivities with the overall quality with respect to the original variables

It is shown below that $\Omega_{Orig.var}$ can be expressed as a weighted average of the p axis predictivities:

$$\begin{aligned}\pi_k &= \frac{[\widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}}]_{kk}}{[\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}]_{kk}} \\ \rightarrow \pi_k [\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}]_{kk} &= [\widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}}]_{kk} \\ \rightarrow \sum_{k=1}^p \pi_k [\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}]_{kk} &= \sum_{k=1}^p [\widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}}]_{kk} \\ \rightarrow \sum_{k=1}^p \pi_k [\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}]_{kk} &= \text{tr} \left\{ \widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}} \right\} \\ \rightarrow \Omega_{Orig.var} &= \sum_{k=1}^p \pi_k \frac{[\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}]_{kk}}{\text{tr}(\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}})}.\end{aligned}$$

Note that since $\mathbf{1}'\mathbf{X} = \mathbf{0}'$, $[\overline{\mathbf{X}}' \overline{\mathbf{X}}]_{kk}$ is proportional to the sample between-groups variance associated with the variable x_k . Hence, when $\mathbf{C} = \mathbf{I}$, the weight of the k th axis predictivity in $\Omega_{Orig.var}$ is an increasing function of the sample between-groups variance associated with x_k . When $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$,

$$[\overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}}]_{kk} = [\overline{\mathbf{X}}' \overline{\mathbf{X}}]_{kk} - J \left(\left[\frac{1}{J} \mathbf{1}' \overline{\mathbf{X}} \right]_k \right)^2$$

and hence the weight of π_k in $\Omega_{Orig.var}$ is also an increasing function of the sample between-groups variance associated with x_k when $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$. Note that $[\overline{\mathbf{X}}' \overline{\mathbf{X}}]_{jj}$ can be larger than $[\overline{\mathbf{X}}' \overline{\mathbf{X}}]_{kk}$, and hence the weight associated with π_j can be larger than that associated with π_k , either because the groups truly differ more with respect to x_j than with respect to x_k or simply because the (unconditional)

standard deviation of x_j is larger than the (unconditional) standard deviation of x_k , or a combination of both these reasons. When $\mathbf{C} = \mathbf{N}$ and $n_j = n_k$, then the weight associated with π_j will be larger than that associated with π_k if and only if the sample between-groups variance associated with x_j is larger than that associated with x_k . When however $\mathbf{C} = \mathbf{N}$ and n_j is larger than n_k , it is possible that the weight associated with π_j in $\Omega_{Orig.var}$ will be larger than the weight associated with π_k , even if the sample between-groups variance associated with x_j is smaller than that associated with x_k . When $\mathbf{C} = \mathbf{N}$ and the sample between-groups variance associated with x_j is identical to that associated with x_k , then the weight associated with π_j will be greater than the weight associated with π_k if and only if n_j is larger than n_k .

5.4.3 Scale invariance

Substituting $\widehat{\mathbf{X}}\mathbf{A}^{-1}$ for $\widehat{\mathbf{X}}^*$ in the expression for the k th axis predictivity of the predictive CVA biplot constructed from the standardised measurements, that is

$$\pi_k^* = \frac{\left[\left(\widehat{\mathbf{X}}^* \right)' \mathbf{C} \widehat{\mathbf{X}}^* \right]_{kk}}{\left[\left(\overline{\mathbf{X}}^* \right)' \mathbf{C} \overline{\mathbf{X}}^* \right]_{kk}}$$

shows that $\pi_k^* = \pi_k$:

$$\begin{aligned} \pi_k^* &= \frac{\left[\left(\widehat{\mathbf{X}}^* \right)' \mathbf{C} \widehat{\mathbf{X}}^* \right]_{kk}}{\left[\left(\overline{\mathbf{X}}^* \right)' \mathbf{C} \overline{\mathbf{X}}^* \right]_{kk}} \\ &= \frac{\left[\mathbf{A}^{-1} \widehat{\mathbf{X}}' \mathbf{C} \widehat{\mathbf{X}} \mathbf{A}^{-1} \right]_{kk}}{\left[\mathbf{A}^{-1} \overline{\mathbf{X}}' \mathbf{C} \overline{\mathbf{X}} \mathbf{A}^{-1} \right]_{kk}} \\ &= \frac{\frac{1}{a_{kk}} \widehat{\mathbf{x}}'_{(k)} \mathbf{C} \widehat{\mathbf{x}}_{(k)} \frac{1}{a_{kk}}}{\frac{1}{a_{kk}} \overline{\mathbf{x}}'_{(k)} \mathbf{C} \overline{\mathbf{x}}_{(k)} \frac{1}{a_{kk}}} \\ &= \frac{\widehat{\mathbf{x}}'_{(k)} \mathbf{C} \widehat{\mathbf{x}}_{(k)}}{\overline{\mathbf{x}}'_{(k)} \mathbf{C} \overline{\mathbf{x}}_{(k)}} \\ &\longrightarrow \pi_k^* = \pi_k . \end{aligned}$$

Like the adequacy measure, the axis predictivity measure is only invariant to transformations of the form $\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x}$ if \mathbf{F} is a $p \times p$ non-singular diagonal matrix.

Since $\pi_k^* = \pi_k$, it follows that when the CVA biplot is constructed from the

standardised measurements, the overall quality of the CVA biplot with respect to the original variables is given by

$$\begin{aligned}
 \Omega_{Orig.var}^* &= \sum_{k=1}^p \pi_k^* \frac{\left[(\bar{\mathbf{X}}^*)' \mathbf{C} \bar{\mathbf{X}}^* \right]_{kk}}{\text{tr} \left((\bar{\mathbf{X}}^*)' \mathbf{C} \bar{\mathbf{X}}^* \right)} \\
 \longrightarrow \Omega_{Orig.var}^* &= \sum_{k=1}^p \pi_k \frac{\left[(\bar{\mathbf{X}}^*)' \mathbf{C} \bar{\mathbf{X}}^* \right]_{kk}}{\text{tr} \left((\bar{\mathbf{X}}^*)' \mathbf{C} \bar{\mathbf{X}}^* \right)} \\
 \longrightarrow \Omega_{Orig.var}^* &= \sum_{k=1}^p \pi_k \left(\frac{\frac{1}{a_{kk}^2} \left[\bar{\mathbf{X}}' \mathbf{C} \bar{\mathbf{X}} \right]_{kk}}{\sum_{i=1}^p \frac{1}{a_{ii}^2} \left[\bar{\mathbf{X}}' \mathbf{C} \bar{\mathbf{X}} \right]_{ii}} \right). \tag{5.4.3}
 \end{aligned}$$

Equation (5.4.3) confirms the fact that $\Omega_{Orig.var}$ is not invariant to the scales in which the original variables were measured. Note that unlike the weights associated with the axis predictivities in $\Omega_{Orig.var}$, the weights associated with the axis predictivities in $\Omega_{Orig.var}^*$ take the different standard deviations of the measured variables into account. When $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = (\mathbf{I} - \frac{1}{j} \mathbf{1}\mathbf{1}')$ (or $\mathbf{C} = \mathbf{N}$ with $n_j = n_k$), the weight associated with π_j in $\Omega_{Orig.var}^*$ will only be greater than that associated with π_k if the groups truly differ more with respect to x_j than with respect to x_k , and not simply because the (unconditional) standard deviation of x_j is greater than that of x_k . Equation (5.4.3) also shows that unlike the overall quality of the PCA biplot, the overall quality of the CVA biplot with respect to the original variables does not simplify to the arithmetic average of the p axis predictivities when the CVA biplot is constructed from the standardised measurements.

5.5 Group predictivities

5.5.1 Definition and properties

Group predictivity (or class predictivity as it is referred to by Gower *et al.* (2011)) is the analogue of the sample predictivity measure associated with the PCA biplot, in the context of the CVA biplot. The group predictivity of the j th group measures the overall accuracy of the approximations to the elements of the j th group centroid that are read off from the predictive CVA biplot axes, that is the overall accuracy of the elements of

$$(\hat{\bar{\mathbf{x}}}^j)' = (\bar{\mathbf{x}}^j)' \mathbf{M}_r \mathbf{M}^r.$$

Gardner-Lubbe *et al.* (2008) proposed the group predictivity measure in the context of the unweighted CVA biplot constructed from the svd of $\bar{\mathbf{X}}\mathbf{L}$ (i.e. $\mathbf{C} = \mathbf{I}$) and also coined the name of the quality measure. They defined the group predictivity of the j th group, or j th group predictivity for short, as:

$$\psi^j = \frac{\left[\widehat{\mathbf{X}}\mathbf{L}\mathbf{L}'\widehat{\mathbf{X}}' \right]_{jj}}{\left[\bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}}' \right]_{jj}}.$$

Gower *et al.* (2011) suggested that this definition be adjusted so as to make it applicable to the general case where the CVA biplot is constructed from the svd of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$, in the following way:

$$\psi^j = \frac{\left[\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}\mathbf{L}'\widehat{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}}{\left[\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}} \quad (5.5.1)$$

$$\rightarrow \psi^j = \frac{\left[\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}}{\left[\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{W}^{-1}\bar{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}}$$

$$\rightarrow \psi^j = \frac{\left[\mathbf{C}^{1/2}\widehat{\mathbf{X}}\widehat{\Sigma}_{\mathbf{W}}^{-1}\widehat{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}}{\left[\mathbf{C}^{1/2}\bar{\mathbf{X}}\widehat{\Sigma}_{\mathbf{W}}^{-1}\bar{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}}. \quad (5.5.2)$$

Note that the definition of the j th group predictivity associated with the unweighted CVA biplot constructed from the svd of $\bar{\mathbf{X}}\mathbf{L}$ that was proposed by Gower *et al.* (2011), is identical to that proposed by Gardner-Lubbe *et al.* (2008). The Type A orthogonality property of the decomposition of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ into $\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{L}$ and $\mathbf{C}^{1/2}(\bar{\mathbf{X}} - \widehat{\mathbf{X}})\mathbf{L}$ validates ψ^j as a quality measure.

Let the J -component vector of group predictivities with j th element equal to the j th group predictivity, be denoted by $\boldsymbol{\psi}$, then

$$\boldsymbol{\psi} = \text{diag} \left(\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}' \mathbf{C}^{1/2} \right) \left[\text{diag} \left(\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{W}^{-1}\bar{\mathbf{X}}' \mathbf{C}^{1/2} \right) \right]^{-1}.$$

Note that when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$, the expression of the j th group predictivity

simplifies to

$$\psi^j = \frac{\left[\widehat{\mathbf{X}} \mathbf{L} \mathbf{L}' \widehat{\mathbf{X}}' \right]_{jj}}{\left[\overline{\mathbf{X}} \mathbf{L} \mathbf{L}' \overline{\mathbf{X}}' \right]_{jj}} \quad (5.5.3)$$

$$\begin{aligned} \rightarrow \psi^j &= \frac{\left[\widehat{\mathbf{X}} \mathbf{W}^{-1} \widehat{\mathbf{X}}' \right]_{jj}}{\left[\overline{\mathbf{X}} \mathbf{W}^{-1} \overline{\mathbf{X}}' \right]_{jj}} \\ \rightarrow \psi^j &= \frac{\left[\widehat{\mathbf{X}} \widehat{\Sigma}_W^{-1} \widehat{\mathbf{X}}' \right]_{jj}}{\left[\overline{\mathbf{X}} \widehat{\Sigma}_W^{-1} \overline{\mathbf{X}}' \right]_{jj}}. \end{aligned} \quad (5.5.4)$$

It is evident from equation (5.5.3) that when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$, the j th group predictivity is equal to the ratio of the squared Pythagorean distance between the point $\mathbf{e}'_j \widehat{\mathbf{X}} \mathbf{L}$ (that is the point that represents the j th group centroid in the CVA biplot space) and the origin to the squared Pythagorean distance between the point $\mathbf{e}'_j \overline{\mathbf{X}} \mathbf{L}$ (that is the point that represents the j th group centroid in the p -dimensional canonical space) and the origin. Equivalently, ψ^j is equal to the ratio of the squared Mahalanobis distance between $\widehat{\mathbf{x}}^j$ and the origin to the squared Mahalanobis distance between $\overline{\mathbf{x}}^j$ and the origin (equation (5.5.4)). On the other hand, when $\mathbf{C} = (\mathbf{I} - \frac{1}{J} \mathbf{1} \mathbf{1}')$, the j th group predictivity is equal to the ratio of the squared Pythagorean distance between the point $\mathbf{e}'_j \widehat{\mathbf{X}} \mathbf{L}$ and the centroid of the J points representing the J group centroids in the CVA biplot space to the squared Pythagorean distance between the point $\mathbf{e}'_j \overline{\mathbf{X}} \mathbf{L}$ and the centroid of the J points representing the J group centroids in the p -dimensional canonical space (equation (5.5.1)). Equivalently, when $\mathbf{C} = (\mathbf{I} - \frac{1}{J} \mathbf{1} \mathbf{1}')$, ψ^j is equal to the ratio of the squared Mahalanobis distance between the point $\widehat{\mathbf{x}}^j$ and the centroid of the J points, $\{\widehat{\mathbf{x}}^j\}$, to the squared Mahalanobis distance between the point $\overline{\mathbf{x}}^j$ and the centroid of the J points, $\{\overline{\mathbf{x}}^j\}$ (equation (5.5.2) with $\mathbf{C} = (\mathbf{I} - \frac{1}{J} \mathbf{1} \mathbf{1}')$).

Consider the following expression of the j th group predictivity:

$$\begin{aligned} \psi^j &= \frac{\left[\mathbf{C}^{1/2} \widehat{\mathbf{X}} \mathbf{L} \mathbf{L}' \widehat{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}}{\left[\mathbf{C}^{1/2} \overline{\mathbf{X}} \mathbf{L} \mathbf{L}' \overline{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}} \\ &= \frac{\left\| \mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}} \mathbf{L} \right\|^2}{\left\| \mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}} \mathbf{L} \right\|^2} \end{aligned}$$

$$= \frac{\|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r \mathbf{V}'_r\|^2}{\|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}\|^2} \quad (5.5.5)$$

$$= \frac{\|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r\|^2}{\|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}\|^2}$$

$$\rightarrow \psi^j = \frac{\sum_{i=1}^r \left([\mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}} \mathbf{L} \mathbf{V}'_r]_{ji} \right)^2}{\sum_{i=1}^p \left([\mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}} \mathbf{L} \mathbf{V}]_{ji} \right)^2}. \quad (5.5.6)$$

Being defined as a ratio of sums of squared values, the group predictivity measure can only take on non-negative values. Since the numerator of ψ^j is a non-decreasing function of the dimension of the CVA biplot space, r , while the denominator is fixed, it follows that the group predictivity measure is a non-decreasing function of the dimension of the CVA biplot space. It is evident that ψ^j has a maximum value of one which will necessarily be attained when $r = p$ (equation (5.5.6)). From the expression of ψ^j in (5.5.5) it is evident that the j th group predictivity is equal to the ratio of the squared length of the projection of the vector $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$ onto the r -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_r)$, to the squared length of the vector $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$. Consequently the j th group predictivity is a decreasing function of the size of the angle between the vectors $\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}} \mathbf{L}$ and $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$:

$$\psi^j = \frac{\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}} \mathbf{L} \mathbf{L}' \widehat{\bar{\mathbf{X}}}' \mathbf{C}^{1/2} \mathbf{e}_j}{\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{L}' \bar{\mathbf{X}}' \mathbf{C}^{1/2} \mathbf{e}_j}$$

$$= \frac{\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r \mathbf{V}'_r \mathbf{V}'_r \mathbf{V}'_r \mathbf{L}' \bar{\mathbf{X}}' \mathbf{C}^{1/2} \mathbf{e}_j}{\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{L}' \bar{\mathbf{X}}' \mathbf{C}^{1/2} \mathbf{e}_j}$$

$$= \left(\frac{\left(\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r \mathbf{V}'_r \mathbf{V}'_r \mathbf{V}'_r \mathbf{L}' \bar{\mathbf{X}}' \mathbf{C}^{1/2} \mathbf{e}_j \right)}{\|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r \mathbf{V}'_r\| \|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}\|} \right)^2$$

$$= \left(\frac{\left(\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r \mathbf{V}'_r \right) \left(\mathbf{L}' \bar{\mathbf{X}} \mathbf{C}^{1/2} \mathbf{e}_j \right)}{\|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r \mathbf{V}'_r\| \|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}\|} \right)^2$$

$$\rightarrow \psi^j = \cos^2(\theta_j) \quad (5.5.7)$$

where θ_j denotes the angle between the vectors, $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L} \mathbf{V}'_r \mathbf{V}'_r$ and $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$. It follows that the j th group predictivity has a maximum value of one which it will attain if and only if the vector $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$ lies in the CVA biplot space, $\mathcal{L} = \mathcal{V}(\mathbf{V}_r)$, and a minimum value of zero which it will attain if and only if the vector $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$

lies in the orthogonal complement of the CVA biplot space, that is:

$$\begin{aligned}\psi^j = 1 &\leftrightarrow \mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} \in \mathcal{L} \\ \psi^j = 0 &\leftrightarrow \mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} \in \mathcal{L}^\perp.\end{aligned}$$

It follows that $r = K$ is both sufficient as well as necessary for all J the group predictivities to equal one:

$$\begin{aligned}\psi^j = 1 \quad \forall j \in [1 : J] &\leftrightarrow \mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} \in \mathcal{V}(\mathbf{V}_r) \quad \forall j \in [1 : J] \\ \therefore \psi^j = 1 \quad \forall j \in [1 : J] &\leftrightarrow r = K.\end{aligned}$$

Note that when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$, ψ^j is equal to the squared cosine of the angle between the vectors, $\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L} \mathbf{V}'_r \mathbf{V}'_r$. This is however not true when $\mathbf{C} = (\mathbf{I} - \frac{1}{J} \mathbf{1}\mathbf{1}')$. Hence, when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$, ψ^j will equal one if and only if the point representing the j th group centroid in the p -dimensional canonical space, that is $\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}$, lies in the CVA biplot space. It is evident from equation (5.5.7) that

$$\psi^j > \psi^k$$

necessarily implies that the angle between the vectors $\mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L}$ is smaller than the angle between the vectors $\mathbf{e}'_k \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_k \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L}$. It follows that when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$ and $\psi^j > \psi^k$, the angle between the vectors $\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}$ is smaller than the angle between the vectors $\mathbf{e}'_k \overline{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_k \widehat{\mathbf{X}}\mathbf{L}$.

The fact that the decomposition of $\mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L}$ into $\mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L}$ and $\mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} - \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L}$ exhibits Type A orthogonality implies that the j th group predictivity can also be expressed as a decreasing function of the Pythagorean distance between the two points, $\mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L}$:

$$\begin{aligned}\psi^j &= \frac{\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \mathbf{L}' \widehat{\mathbf{X}}' \mathbf{C}^{1/2} \mathbf{e}_j}{\mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} \mathbf{L}' \overline{\mathbf{X}} \mathbf{C}^{1/2} \mathbf{e}_j} \\ \rightarrow \psi^j &= 1 - \frac{(\mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L})(\mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L})'}{\mathbf{e}'_j \mathbf{C}^{1/2} \overline{\mathbf{X}}\mathbf{L} \mathbf{L}' \overline{\mathbf{X}} \mathbf{C}^{1/2} \mathbf{e}_j}.\end{aligned}\quad (5.5.8)$$

Note that when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$, ψ^j is a decreasing function of the Pythagorean

distance between the point representing the j th group centroid in the p -dimensional canonical space, that is $\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}$, and the point representing that group centroid in the lower-dimensional CVA biplot space, that is $\mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}$:

$$\begin{aligned} \psi^j &= \frac{\mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}\mathbf{L}'\widehat{\mathbf{X}}' \mathbf{e}_j}{\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}} \mathbf{e}_j} \\ \rightarrow \psi^j &= 1 - \frac{(\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L})(\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L})'}{\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}}' \mathbf{e}_j}. \end{aligned} \quad (5.5.9)$$

This is however not true when $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$ since $\mathbf{1}'\bar{\mathbf{X}}\mathbf{L}$ is in general not equal to $\mathbf{1}'\widehat{\mathbf{X}}\mathbf{L} = \mathbf{1}'\bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$, where \mathbf{V} is the matrix of right singular vectors of $(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')\bar{\mathbf{X}}\mathbf{L}$. It seems reasonable that $\mathbf{1}'\widehat{\mathbf{X}}\mathbf{L}$ should at least be similar to $\mathbf{1}'\bar{\mathbf{X}}\mathbf{L}$ and hence that if the j th group predictivity corresponding to the r -dimensional CVA biplot constructed from $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$ is very high, then the point $(\bar{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$ should be close to the point $(\bar{\mathbf{x}}^j)'\mathbf{L}$.

Note that since the exact (K -dimensional) CVA classification regions are based on the positions of the points representing the J group means in the K -dimensional subspace of the p -dimensional canonical space that perfectly contains the J group means, \mathbb{C}^K , i.e. the J points $\{(\bar{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_K\mathbf{V}'_K\} = \{(\bar{\mathbf{x}}^j)'\mathbf{L}\}$, while the (approximate) classification regions of the r -dimensional CVA biplot are based on the positions of the J points $\{(\bar{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}'_r\}$, it follows that the closer the J points $\{(\bar{\mathbf{x}}^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}'_r\}$ are to the corresponding points in the set $\{(\bar{\mathbf{x}}^j)'\mathbf{L}\}$, the more accurately the classification regions in the CVA biplot will represent the exact CVA classification regions. It follows that if all J the group predictivities of the r -dimensional CVA biplot are very high, then the classification regions in the r -dimensional CVA biplot accurately represent the exact CVA classification regions.

Given that ψ^j is a non-decreasing function of the dimension of the CVA biplot space, r , equation (5.5.9) implies that when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$, the Pythagorean distance between the point representing the j th group centroid in the CVA biplot space and that representing the j th group centroid in the p -dimensional canonical space is a non-increasing function of the dimension of the CVA biplot space. This is true irrespective of the \mathbf{C} matrix used in the construction of the CVA biplot since the orthonormal matrix \mathbf{V}_r in

$$\widehat{\mathbf{X}}\mathbf{L} = \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$$

ensures that the decomposition,

$$\bar{\mathbf{X}}\mathbf{L} = \widehat{\mathbf{X}}\mathbf{L} + (\bar{\mathbf{X}}\mathbf{L} - \widehat{\mathbf{X}}\mathbf{L})$$

exhibits Type A orthogonality irrespective of the \mathbf{C} -matrix used in the construction of the CVA biplot:

$$\begin{aligned} \left\| \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_{r+1} \mathbf{V}'_{r+1} \right\|^2 &= (\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_{r+1} \mathbf{V}'_{r+1}) (\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_{r+1} \mathbf{V}'_{r+1})' \\ &= \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}}' \mathbf{e}_j - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_{r+1} \mathbf{V}'_{r+1} \mathbf{L}'\bar{\mathbf{X}}' \mathbf{e}_j \\ &= \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}}' \mathbf{e}_j - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r \mathbf{V}'_r \mathbf{L}'\bar{\mathbf{X}}' \mathbf{e}_j \\ &\quad - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{v}_{(r+1)} \mathbf{v}'_{(r+1)} \mathbf{L}'\bar{\mathbf{X}}' \mathbf{e}_j \\ &= (\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r \mathbf{V}'_r) (\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r \mathbf{V}'_r)' \\ &\quad - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{v}_{(r+1)} \mathbf{v}'_{(r+1)} \mathbf{L}'\bar{\mathbf{X}}' \mathbf{e}_j \\ \rightarrow \left\| \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_{r+1} \mathbf{V}'_{r+1} \right\|^2 &< \left\| \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r \mathbf{V}'_r \right\|^2 \\ \rightarrow \left\| \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_{r+1} \mathbf{V}'_{r+1} \right\| &< \left\| \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r \mathbf{V}'_r \right\|. \end{aligned}$$

It is evident from the expression of ψ^j in (5.5.8) that it is only when

$$\left\| \mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} \right\| = \left\| \mathbf{e}'_k \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} \right\| \quad (5.5.10)$$

that

$$\psi^j > \psi^k \quad (5.5.11)$$

necessarily implies that

$$\left\| \mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right\| < \left\| \mathbf{e}'_k \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right\|. \quad (5.5.12)$$

Conclusions about the relative magnitudes of the Pythagorean distances,

$$\left\| \mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\mathbf{X}}\mathbf{L} \right\| \quad \text{for } j \in [1 : J]$$

can therefore not be made from the group predictivities alone. When $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$, the expression of ψ^j in (5.5.9) shows that it is only when $\|\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\| = \|\mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}\|$ that (5.5.11) necessarily implies that

$$\|\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\bar{\mathbf{X}}}\mathbf{L}\| < \|\mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \widehat{\bar{\mathbf{X}}}\mathbf{L}\| .$$

When $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$ the inequality in (5.5.11) does not imply (5.5.12) even if equation (5.5.10) holds due to the fact that in general $\mathbf{1}'\bar{\mathbf{X}}\mathbf{L}$ is not equal to $\mathbf{1}'\widehat{\bar{\mathbf{X}}}\mathbf{L}$. It follows that for all three choices of \mathbf{C} , conclusions about the relative magnitudes of the Pythagorean distances,

$$\|\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\bar{\mathbf{X}}}\mathbf{L}\| \quad \text{for } j \in [1 : J]$$

cannot be made from the group predictivities alone.

5.5.2 Group predictivities and the accuracy of distances represented in the CVA biplot

When the j th and k th group predictivities associated with an r -dimensional CVA biplot are close to one and hence $\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}}\mathbf{L}$ and $\mathbf{e}'_k \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}}\mathbf{L}$ are accurate approximations of $\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_k \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L}$ respectively, it follows that $\|\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}}\mathbf{L} - \mathbf{e}'_k \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}}\mathbf{L}\|$ will most likely be an accurate approximation of $\|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L}\|$. Note that for $\mathbf{C} = \mathbf{I}$ and $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$,

$$\begin{aligned} \|\mathbf{e}'_j \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}}\mathbf{L} - \mathbf{e}'_k \mathbf{C}^{1/2} \widehat{\bar{\mathbf{X}}}\mathbf{L}\| &= \|\mathbf{e}'_j \widehat{\bar{\mathbf{X}}}\mathbf{L} - \mathbf{e}'_k \widehat{\bar{\mathbf{X}}}\mathbf{L}\| \\ \text{and } \|\mathbf{e}'_j \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \mathbf{C}^{1/2} \bar{\mathbf{X}}\mathbf{L}\| &= \|\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}\| . \end{aligned}$$

Hence, when the j th and k th group predictivities associated with an r -dimensional unweighted CVA biplot are close to one, $\|\mathbf{e}'_j \widehat{\bar{\mathbf{X}}}\mathbf{L} - \mathbf{e}'_k \widehat{\bar{\mathbf{X}}}\mathbf{L}\|$ will most likely be an accurate approximation of $\|\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}\|$. When $\mathbf{C} = \mathbf{N}$ and the j th group predictivity associated with the r -dimensional CVA biplot is close to one, that is $\mathbf{e}'_j \mathbf{N}^{1/2} \widehat{\bar{\mathbf{X}}}_r \mathbf{L}$ is an accurate approximation of $\mathbf{e}'_j \mathbf{N}^{1/2} \bar{\mathbf{X}}\mathbf{L}$, then $\mathbf{e}'_j \widehat{\bar{\mathbf{X}}}\mathbf{L}$ will necessarily be an accurate approximation of $\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}$, $j \in [1 : J]$. Consequently, when $\mathbf{C} = \mathbf{N}$ and the j th and k th group predictivities are close to one, $\|\mathbf{e}'_j \widehat{\bar{\mathbf{X}}}\mathbf{L} - \mathbf{e}'_k \widehat{\bar{\mathbf{X}}}\mathbf{L}\|$ will most likely be an accurate approximation of $\|\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}\|$. It follows that whenever both ψ^j and ψ^k

are close to one, the Pythagorean distance between the two points representing the j th and k th group centroids in the r -dimensional CVA biplot will most likely be an accurate approximation of the Pythagorean distance between the corresponding two points in the p -dimensional canonical space, irrespective of the \mathbf{C} matrix used in the construction of the CVA biplot. Equivalently, when ψ^j and ψ^k are close to one, the Pythagorean distance between the two points representing the j th and k th group centroids in the r -dimensional CVA biplot will most likely be an accurate approximation of $\frac{1}{\sqrt{n}}$ times the Mahalanobis distance between the two points representing the j th and k th group centroids in the p -dimensional measurement space, irrespective of the \mathbf{C} -matrix used in the construction of the CVA biplot. When however one or both of the j th and k th group predictivities are not close to one, $\|\mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \widehat{\mathbf{X}}\mathbf{L}\|$ will likely not be an accurate approximation of $\|\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \overline{\mathbf{X}}\mathbf{L}\|$. It is however not guaranteed that the approximation will be poor - it is possible that the Pythagorean distance $\|\mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_k \widehat{\mathbf{X}}\mathbf{L}\|$ is accurately approximated in the CVA biplot even if one or both of the individual group centroids are poorly approximated. It follows that the relative magnitudes of the Pythagorean distances between points representing group centroids in the CVA biplot will most likely be accurate approximations of the relative magnitudes of the Mahalanobis distances between the points representing those group centroids in the p -dimensional measurement space, only when the group predictivities of those groups are close to one. Remember that the smaller the Mahalanobis distance between two group centroids in the measurement space is, the more similar the two groups are with respect to the measured variables. The relative magnitudes of the Pythagorean distances between points representing group centroids in the CVA biplot can therefore only be expected to accurately reflect the true differences between the group centroids when the group predictivities of those groups are close to one. Recall that of the three types of CVA biplots, the unweighted CVA biplot constructed from the svd of $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')\overline{\mathbf{X}}\mathbf{L}$ yields the most accurate visualisation of the relative magnitudes of the Mahalanobis distances between the group centroids in the p -dimensional measurement space (Section 4.4).

A measure of the accuracy of the Pythagorean distances between the group centroids in the CVA biplot as approximations to the corresponding Pythagorean distances in the p -dimensional canonical space is proposed in Section 5.6.

5.5.3 The effect of accounting for the group sizes in the construction of the CVA biplot on the group predictivities

Recall from Section 4.4 that the r -dimensional weighted CVA biplot space tends to lie closer to the points representing the group centroids of the larger groups in the p -dimensional canonical space than the two unweighted CVA biplots of the same dimension, $r \in [1 : K - 1]$. This means that if the j th group is one of the larger groups, then the orthogonal projection of the point $\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}$ onto the r -dimensional weighted CVA biplot space tends to lie closer to $\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}$ (in terms of Pythagorean distance) than does the orthogonal projections of $\mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}$ onto the r -dimensional unweighted CVA biplot spaces associated with $\mathbf{C} = \mathbf{I}$ and $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$ respec-

tively, $r \in [1 : K - 1]$. Similarly, if the k th group is one of the smaller groups then the orthogonal projection of $\mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}$ onto the r -dimensional weighted CVA biplot space tends to lie further from $\mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}$ than does the orthogonal projections of $\mathbf{e}'_k \bar{\mathbf{X}}\mathbf{L}$ onto the r -dimensional unweighted CVA biplot spaces associated with $\mathbf{C} = \mathbf{I}$ and $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$ respectively, $r \in [1 : K - 1]$. Since both the j th group predictivity associated with the weighted CVA biplot and that associated with the unweighted CVA biplot corresponding to $\mathbf{C} = \mathbf{I}$ are decreasing functions of the Pythagorean distance between the point $\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}$ and its orthogonal projection onto the CVA biplot space, this implies that the group predictivities of the larger groups associated with the r -dimensional weighted CVA biplot will tend to be greater than those associated with the r -dimensional unweighted CVA biplot corresponding to $\mathbf{C} = \mathbf{I}$, $r \in [1 : K - 1]$. Similarly, the group predictivities of the smaller groups associated with the r -dimensional weighted CVA biplot will tend to be smaller than those associated with the r -dimensional unweighted CVA biplot corresponding to $\mathbf{C} = \mathbf{I}$, $r \in [1 : K - 1]$. It is however important to note that these relationships between the group predictivities associated with the weighted and unweighted CVA biplots are not guaranteed to hold. Consider for instance the case where the point representing the group centroid of a very small group in the p -dimensional canonical space lies close to a point representing the group centroid of a very large group. In this scenario, if the weighted CVA biplot space lies closer to the point representing the group centroid of the large group than does the unweighted CVA biplot space, then the same will most likely be true for the point representing the group centroid of the small group.

Consider the expression of the j th group predictivity corresponding to the unweighted CVA biplot constructed from the svd of $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}$:

$$\psi^j = 1 - \frac{\left(\mathbf{e}'_j (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \widehat{\bar{\mathbf{X}}}\mathbf{L}\right) \left(\mathbf{e}'_j (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \widehat{\bar{\mathbf{X}}}\mathbf{L}\right)'}{\mathbf{e}'_j (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}\mathbf{L}'\bar{\mathbf{X}}' (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \mathbf{e}_j}.$$

Since the j th group predictivity associated with the unweighted CVA biplot constructed from the svd of $(\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}') \bar{\mathbf{X}}\mathbf{L}$ is not a decreasing function of the Pythagorean distance between $\mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}$ and $\mathbf{e}'_j \widehat{\bar{\mathbf{X}}}\mathbf{L}$, it is not as directly comparable to the j th group predictivity associated with the weighted CVA biplot as the j th group predictivity associated with the unweighted CVA biplot corresponding to $\mathbf{C} = \mathbf{I}$.

Three of the simulated data sets introduced in Section 4.8 will now be used to illustrate the effect of taking the group sizes into account in the construction of a CVA biplot on the group predictivities. The group predictivities corresponding to the one-dimensional weighted and unweighted CVA biplots of the third, fifth and seventh simulated data sets are provided in Tables 5.1, 5.2 and 5.3 respectively.

Upon consideration of the group sizes and the group predictivities in Table 5.1, note that for each of the two larger groups the group predictivity associated with the weighted CVA biplot is not only larger than that associated with the unweighted CVA biplot constructed with $\mathbf{C} = \mathbf{I}$ but also larger than that corresponding to the

unweighted CVA biplot constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$. Similarly, for each of the two smaller groups the group predictivity corresponding to the weighted CVA biplot is not only smaller than that corresponding to the unweighted CVA biplot corresponding to $\mathbf{C} = \mathbf{I}$ but also smaller than that corresponding to the unweighted CVA biplot corresponding to $\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$. That is, taking the group sizes into account in the construction of the one-dimensional CVA biplot, resulted in an increase in the group predictivities of the two larger groups (i.e. Group 3 and Group 4) and a decrease in the group predictivities of the two smaller groups (i.e. Group 1 and Group 2). Since for each of the two larger groups the group predictivity corresponding to the weighted CVA biplot is larger than that corresponding to the unweighted CVA biplot associated with $\mathbf{C} = \mathbf{I}$, it can be concluded that the one-dimensional weighted CVA biplot space lies closer to the points representing the group centroids of the two larger groups in the canonical space (or equivalently in \mathbb{C}^3) than does the one-dimensional unweighted CVA biplot space corresponding to $\mathbf{C} = \mathbf{I}$. Similarly, since for each of the two smaller groups the group predictivity corresponding to the weighted CVA biplot is smaller than that corresponding to the unweighted CVA biplot associated with $\mathbf{C} = \mathbf{I}$, it can be concluded that the one-dimensional weighted CVA biplot space lies further away from the points representing the group centroids of the two smaller groups in \mathbb{C}^3 than does the one-dimensional unweighted CVA biplot space corresponding to $\mathbf{C} = \mathbf{I}$.

Table 5.1: *The group predictivities of the one-dimensional weighted and unweighted CVA biplots of the third simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	50	50	150	150
$\mathbf{C} = \mathbf{N}$	0.035	0.067	0.879	0.789
$\mathbf{C} = \mathbf{I}$	0.359	0.836	0.229	0.081
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.047	0.828	0.609	0.159

The same relationship that is observed between the group sizes and the group predictivities associated with the one-dimensional weighted and unweighted CVA biplots for the third simulated data set, is observed for the fifth simulated data set in Table 5.2.

Table 5.2: *The group predictivities corresponding to the one-dimensional CVA biplots of the fifth simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	50	150	50	150
$\mathbf{C} = \mathbf{N}$	0.145	0.984	0.541	0.358
$\mathbf{C} = \mathbf{I}$	0.451	0.616	0.858	0.007
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.165	0.882	0.774	0.041

The group predictivities of Group 4 of the seventh simulated data set (Table 5.3) demonstrate that the group predictivity of a small group (that is, small relative to the other groups) associated with a weighted CVA biplot will not necessarily be smaller than those associated with the corresponding unweighted CVA biplots.

Table 5.3: *The group predictivities of the one-dimensional weighted and unweighted CVA biplots of the seventh simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	160	80	80	80
$\mathbf{C} = \mathbf{N}$	0.554	0.124	0.791	0.226
$\mathbf{C} = \mathbf{I}$	0.103	0.534	0.850	0.040
$\mathbf{C} = \left(\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)$	0.072	0.712	0.865	0.016

5.5.4 The relationship between group predictivities and the overall quality with respect to the canonical variables

Recall that the overall quality of the CVA biplot with respect to the canonical variables, $\Omega_{Can.var}$, can be expressed as

$$\Omega_{Can.var} = \frac{\text{tr}\left(\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}'\mathbf{C}^{1/2}\right)}{\text{tr}\left(\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}\right)}.$$

The similarity between the expressions of $\Omega_{Can.var}$ and ψ indicates the possibility of a close relationship between $\Omega_{Can.var}$ and the J group predictivities. It is shown below that $\Omega_{Can.var}$ can be expressed as the weighted average of the J group predictivities:

$$\begin{aligned} \psi^i &= \frac{\left[\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii}}{\left[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii}} \\ \rightarrow \psi^i \left[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii} &= \left[\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii} \\ \rightarrow \sum_{i=1}^J \psi^i \left[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii} &= \sum_{i=1}^J \left[\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii} \\ \rightarrow \sum_{i=1}^J \psi^i \left[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii} &= \text{tr}\left(\mathbf{C}^{1/2}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}'\mathbf{C}^{1/2}\right) \\ \rightarrow \Omega_{Can.var} &= \sum_{i=1}^J \psi^i \frac{\left[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}\right]_{ii}}{\text{tr}\left(\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}\right)}. \end{aligned}$$

Note that when $\mathbf{C} = \mathbf{I}$, the weight of ψ^j in $\Omega_{Can.var}$ will be greater than that of ψ^k if and only if the Pythagorean distance between the point representing the j th group centroid in the p -dimensional canonical space and the origin, $[\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}']_{jj}$, is greater than $[\overline{\mathbf{X}}\mathbf{W}^{-1}\overline{\mathbf{X}}']_{kk}$. When $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$, the weight of ψ^j in $\Omega_{Can.var}$ will be greater than that of ψ^k if and only if the Pythagorean distance between the point representing the j th group centroid in the p -dimensional canonical space and the centroid of the points representing the J group centroids, is greater than the Pythagorean distance between the point representing the k th group centroid and that centroid. Note that when $\mathbf{C} = \mathbf{N}$, then, if the sizes of the j th and k th groups are identical, then the weight of ψ^j in $\Omega_{Can.var}$ will be greater than that of ψ^k if and only if the Pythagorean distance between the point representing the j th group centroid in the p -dimensional canonical space and the origin is greater than that between the point representing the k th group centroid and the origin. On the other hand, when $\mathbf{C} = \mathbf{N}$ and the Pythagorean distance between the point representing the j th group centroid in the p -dimensional canonical space and the origin and the Pythagorean distance between the point representing the k th group centroid in the p -dimensional canonical space and the origin are identical, then the weight of ψ^j in $\Omega_{Can.var}$ will be greater than that of ψ^k if and only if $n_j > n_k$. When $\mathbf{C} = \mathbf{N}$, the group predictivity corresponding to a group which is very large relative to the other groups, will likely have a much greater impact on the value of $\Omega_{Can.var}$ than the group predictivities of the smaller groups and consequently the value of $\Omega_{Can.var}$ may not be a good indication of the overall accuracy of the approximations to the J individual canonical centroids.

5.5.5 Scale invariance

Upon substituting $\overline{\mathbf{X}}\mathbf{A}^{-1}$ for $\overline{\mathbf{X}}^*$, $\widehat{\overline{\mathbf{X}}}\mathbf{A}^{-1}$ for $\widehat{\overline{\mathbf{X}}}^*$ and $\mathbf{A}\mathbf{W}^{-1}\mathbf{A}$ for $(\mathbf{W}^*)^{-1}$ in the expression for the j th group predictivity of the CVA biplot constructed from the standardised measurements,

$$\psi^{j*} = \frac{[\mathbf{C}^{1/2}\widehat{\overline{\mathbf{X}}}^* (\mathbf{W}^*)^{-1} (\widehat{\overline{\mathbf{X}}}^*)' \mathbf{C}^{1/2}]_{jj}}{[\mathbf{C}^{1/2}\overline{\mathbf{X}}^* (\mathbf{W}^*)^{-1} (\overline{\mathbf{X}}^*)' \mathbf{C}^{1/2}]_{jj}},$$

it is evident that $\psi^{j*} = \psi^j$:

$$\psi^{j*} = \frac{[\mathbf{C}^{1/2}\widehat{\overline{\mathbf{X}}}\mathbf{A}^{-1}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}\mathbf{A}^{-1}\widehat{\overline{\mathbf{X}}}'\mathbf{C}^{1/2}]_{jj}}{[\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}\mathbf{A}^{-1}\overline{\mathbf{X}}'\mathbf{C}^{1/2}]_{jj}},$$

$$\begin{aligned}
&= \frac{\left[\mathbf{C}^{1/2} \widehat{\mathbf{X}} \mathbf{W}^{-1} \widehat{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}}{\left[\mathbf{C}^{1/2} \overline{\mathbf{X}} \mathbf{W}^{-1} \overline{\mathbf{X}}' \mathbf{C}^{1/2} \right]_{jj}} \\
&\longrightarrow \psi^{j*} = \psi^j.
\end{aligned}$$

Substituting any $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in the equations above shows that the group predictivity measure is invariant to all non-singular linear transformations of the scales of the measurements that are of the form $\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x}$.

5.6 Group contrast predictivities

5.6.1 Definition and Properties

Given that the Pythagorean distances between the points representing the group centroids in the p -dimensional canonical space are proportional to the Mahalanobis distances between the group centroids in the p -dimensional measurement space, it is of interest to know how well the Pythagorean distances between the group centroids in the CVA biplot approximates the corresponding Pythagorean distances in the canonical space. The accuracy with which the Pythagorean distance between the i th and j th group centroids in the CVA biplot approximates the corresponding Pythagorean distance in the p -dimensional canonical space can (for all three choices of \mathbf{C}) be measured by the ratio,

$$\psi^{ij} = \frac{\left\| \mathbf{e}'_i \widehat{\mathbf{X}} \mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}} \mathbf{L} \right\|^2}{\left\| \mathbf{e}'_i \overline{\mathbf{X}} \mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}} \mathbf{L} \right\|^2} \quad (5.6.1)$$

where

$$\widehat{\mathbf{X}} \mathbf{L} = \overline{\mathbf{X}} \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$$

and \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2} \overline{\mathbf{X}} \mathbf{L}$. The ratio in (5.6.1) will henceforth be referred to as the (ij) th group contrast predictivity, $i, j \in [1 : J]$, $i \neq j$.

The validity of the ratio in (5.6.1) as a measure of the accuracy of the approximation $\left\| \mathbf{e}'_i \widehat{\mathbf{X}} \mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}} \mathbf{L} \right\|$ to $\left\| \mathbf{e}'_i \overline{\mathbf{X}} \mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}} \mathbf{L} \right\|$ is based on the orthogonal decomposition

of the total squared distance $\|\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\|^2$,

$$\|\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L}\|^2 = \|\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}\|^2 + \left\| \left(\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \right) \right\|^2. \quad (5.6.2)$$

The decomposition in (5.6.2) holds due to the fact that the cross product term

$$\left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \right)'$$

is equal to zero:

$$\begin{aligned} & \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \right)' \\ &= \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} \right)' - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right)' \\ &= \left(\mathbf{e}'_i \bar{\mathbf{X}} - \mathbf{e}'_j \bar{\mathbf{X}} \right) \mathbf{L} \mathbf{V}_r \mathbf{V}'_r \mathbf{L}' \left(\mathbf{e}'_i \bar{\mathbf{X}} - \mathbf{e}'_j \bar{\mathbf{X}} \right)' - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right)' \\ &= \left(\mathbf{e}'_i \bar{\mathbf{X}} - \mathbf{e}'_j \bar{\mathbf{X}} \right) \mathbf{L} \mathbf{V}_r \mathbf{V}'_r \mathbf{V}_r \mathbf{V}'_r \mathbf{L}' \left(\mathbf{e}'_i \bar{\mathbf{X}} - \mathbf{e}'_j \bar{\mathbf{X}} \right)' - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right)' \\ &= \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right)' - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right)' \\ &= 0. \end{aligned}$$

Due to the fact that the group contrast predictivity measure is defined as the ratio of squared values it can only take on non-negative values. The (ij) th group contrast predictivity has a minimum value of zero which it will attain if and only if

$$\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} = \mathbf{0}'.$$

From equation (5.6.2) it is evident that ψ^{ij} can also be expressed as

$$\psi^{ij} = 1 - \frac{\left\| \mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \right\|^2}{\left\| \mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} \right\|^2}. \quad (5.6.3)$$

It is evident from equation (5.6.3) that ψ^{ij} will attain its maximum value when

$$\left\| \mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) \right\|^2 \quad (5.6.4)$$

attains its minimum value. Since (5.6.4) has a minimum value of zero which it will attain if and only if

$$\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} - \left(\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right) = \mathbf{0}$$

ψ^{ij} has a maximum value of one which it will attain if and only if

$$\mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} = \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}.$$

Note that the conditions

$$\begin{aligned} \mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} &= \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} \\ \text{and } \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} &= \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \end{aligned}$$

are sufficient however not necessary for (5.6.4) to attain the value of zero. This implies that although accurate approximations of both the i th and j th group centroids suggest that it can be expected that $\left\| \mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} \right\|$ will be accurately approximated in the CVA biplot, an accurate approximation of $\left\| \mathbf{e}'_i \bar{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \bar{\mathbf{X}}\mathbf{L} \right\|$ can be achieved without accurate approximations of both or even one of the two group centroids. That is, although high values for both ψ^i and ψ^j imply that ψ^{ij} will most likely attain a value close to one, it is not a necessary condition for ψ^{ij} to be close to one. An example that demonstrates this is provided in Section 5.7. Remember that since

$$\bar{\mathbf{X}}\mathbf{L}\mathbf{V}'_K \mathbf{V}'_K = \bar{\mathbf{X}}\mathbf{L}$$

all the group contrast predictivities associated with the K -dimensional weighted and the two K -dimensional unweighted CVA biplots will be equal to one.

Recall from Section 4.4 that the r -dimensional unweighted CVA biplot space constructed with $\mathbf{C} = \left(\mathbf{I} - \frac{1}{j} \mathbf{1}\mathbf{1}' \right)$ is the r -dimensional subspace of the p -dimensional canonical space in which the Pythagorean distances between the group centroids in the canonical space are optimally approximated, $r \in [1 : K - 1]$. Consequently, it can

be expected that the group contrast predictivities associated with the r -dimensional unweighted CVA biplot constructed with $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$ will on average be higher than those corresponding to the r -dimensional weighted CVA biplot and the r -dimensional unweighted CVA biplot constructed with $\mathbf{C} = \mathbf{I}$, $r \in [1 : K - 1]$. Furthermore, recall that the points representing the group centroids of the larger groups in the p -dimensional canonical space tend to lie closer to their orthogonal projections onto the r -dimensional weighted CVA biplot space than to their orthogonal projections onto the corresponding r -dimensional unweighted CVA biplots, $r \in [1 : K - 1]$. It follows that for the subset of groups that are large relative to the other groups, the Pythagorean distances between the points representing their centroids in the canonical space are likely to be approximated more accurately in the r -dimensional weighted CVA biplot than in the corresponding r -dimensional unweighted CVA biplots. That is, the group contrast predictivities associated with the larger groups are likely to be higher for the r -dimensional weighted CVA biplot than for the corresponding r -dimensional unweighted CVA biplots. By a similar argument, the group contrast predictivities associated with the smaller groups are likely to be higher for the r -dimensional unweighted CVA biplots than for the corresponding r -dimensional weighted CVA biplot.

Note that since

$$\| \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \|^2 = \frac{1}{n} (\mathbf{e}'_i \widehat{\mathbf{X}} - \mathbf{e}'_j \widehat{\mathbf{X}}) \widehat{\Sigma}_W^{-1} (\mathbf{e}'_i \widehat{\mathbf{X}} - \mathbf{e}'_j \widehat{\mathbf{X}})'$$

ψ^{ij} is equal to the squared approximated Mahalanobis distance between the i th and j th group centroids, expressed as a proportion of the squared true Mahalanobis distance between those group centroids. It follows that if all the group contrast predictivities corresponding to some set of group centroids are very high, then the relative magnitudes of the Pythagorean distances between those group centroids in the CVA biplot accurately represents the relative magnitudes of the Mahalanobis distances between those groups in the measurement space.

Recall that the K -dimensional subspace of the canonical space that perfectly contains the J canonical means i.e. \mathbb{C}^K is the space in which the J groups are optimally represented or equivalently the space in which the Pythagorean distance between two points representing two group means is proportional to the Mahalanobis distance between the two group means in the measurement space. It follows that when the relative magnitudes of the Mahalanobis distances between the group means in the measurement space are poorly approximated by the corresponding Pythagorean distances in an r -dimensional CVA biplot, the separation of groups in that r -dimensional CVA biplot is likely to be poor compared to the separation of the groups in \mathbb{C}^K . Hence, if on average the group contrast predictivities are low, then the separation of the groups in the CVA biplot is likely to be poor relative to the separation of the groups in \mathbb{C}^K . Similarly, if on all of the group contrast predictivities are very high, then the separation of the groups in the CVA biplot is likely similar to that in \mathbb{C}^K . Furthermore, since the classification regions in \mathbb{C}^K are exact and determined by the positions of the points representing the group means in

\mathbb{C}^K , it is also likely that the classification regions in the r -dimensional CVA biplot are poor approximations of the exact classification regions if the some (or all) of the group contrast predictivities are low. Similarly, if all the group contrast predictivities are very high, then the classification regions in the r -dimensional CVA biplot most likely accurately represent the exact CVA classification regions.

5.6.2 Scale invariance

Upon substituting $\overline{\mathbf{X}}\mathbf{A}^{-1}$ for $\overline{\mathbf{X}}^*$, $\widehat{\mathbf{X}}\mathbf{A}^{-1}$ for $\widehat{\mathbf{X}}^*$ and $\mathbf{A}\mathbf{L}$ for \mathbf{L}^* in the expression of the (ij) th group contrast predictivity associated with the CVA biplot constructed from the standardised measurements,

$$\psi^{ij*} = \frac{\left\| \mathbf{e}'_i \widehat{\mathbf{X}}^* \mathbf{L}^* - \mathbf{e}'_j \widehat{\mathbf{X}}^* \mathbf{L}^* \right\|^2}{\left\| \mathbf{e}'_i \overline{\mathbf{X}}^* \mathbf{L}^* - \mathbf{e}'_j \overline{\mathbf{X}}^* \mathbf{L}^* \right\|^2}$$

it is evident that $\psi^{ij*} = \psi^{ij}$:

$$\begin{aligned} \psi^{ij*} &= \frac{\left\| \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{L} \right\|^2}{\left\| \mathbf{e}'_i \overline{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{L} \right\|^2} \\ &= \frac{\left\| \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \right\|^2}{\left\| \mathbf{e}'_i \overline{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L} \right\|^2} \\ \psi^{ij*} &= \psi^{ij}. \end{aligned}$$

Substituting any $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in the equations above shows that the group contrast predictivity is invariant to all non-singular linear transformations of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$.

5.7 Axis predictivities, group predictivities and group contrast predictivities: an illustrative example

Some of the concepts regarding axis predictivities, group predictivities and group contrast predictivities will now be illustrated at the hand of the seventh simulated data set that was introduced in Section 4.8. Consider the two-dimensional (predictive) unweighted (with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$) CVA biplot of the seventh simulated data

5.7. AXIS PREDICTIVITIES, GROUP PREDICTIVITIES AND GROUP CONTRAST PREDICTIVITIES: AN ILLUSTRATIVE EXAMPLE

291

set provided in Figure 5.1. In this biplot only the group centroid and 50% bag is shown for each of the four groups.

The relative magnitudes of the Pythagorean distances between the points representing the group centroids in the biplot in Figure 5.1 suggest that the centroid of Group 4 differs less from the centroids of Groups 2 and 3 than from the centroid of Group 1 while it differs to approximately the same extent from the centroids of Groups 2 and 3. These distances are provided in Table 5.4. However, given the very low group predictivity of Group 4 shown in Table 5.5, it is likely that the Pythagorean distances between the point representing the group centroid of Group 4 in the five-dimensional canonical space and the points representing the group centroids of the other groups are not accurately represented in the two-dimensional CVA biplot space. It can therefore be expected that the relative magnitudes of the Mahalanobis distances between the group centroid of Group 4 and each of the other group centroids in the five-dimensional measurement space will not be accurately represented by the relative magnitudes of the corresponding Pythagorean distances in the two-dimensional CVA biplot space.

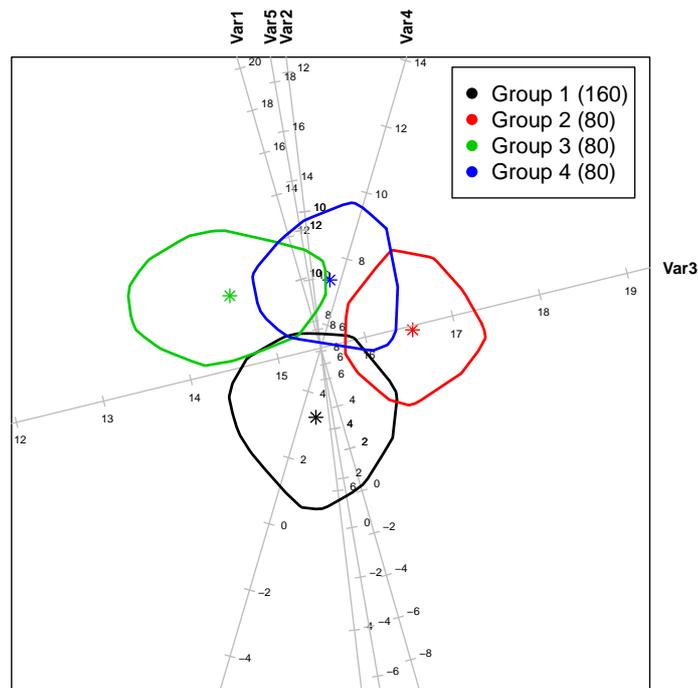


Figure 5.1: *The two-dimensional (predictive) unweighted (with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$) CVA biplot of the seventh simulated data set showing the group centroid (asterisk) and 50% bag for each of the four groups.*

In an attempt to confirm some of these expectations, consider the group contrast predictivities corresponding to the two-dimensional unweighted CVA biplot provided in Table 5.6 - the element in the i th row and j th column of Table 5.6 is equal to the

(ij) th group contrast predictivity, $i, j \in [1 : J], i \neq j$. Inspection of Table 5.6 shows that as far as the centroid of Group 4 is concerned, only its Pythagorean distance from the centroid of Group 1 is accurately approximated in the two-dimensional CVA biplot. The relative magnitudes of the true Pythagorean distances provided in Table 5.7 show that the centroid of Group 4 differs to approximately the same extent from those of Groups 1, 2 and 3 and does not differ more from the centroid of Group 1 than from the centroids of Groups 2 and 3 as suggested by CVA biplot in Figure 5.1. The fact that the group contrast predictivity associated with Group 4 and Group 1 is so high even though the group predictivity of Group 4 is very low demonstrates that it is possible that the Pythagorean distance between two group centroids in the canonical space can be accurately approximated in the CVA biplot without both group centroids being accurately approximated.

Table 5.4: *The Pythagorean distances between the points representing the group centroids of the seventh simulated data set in the two-dimensional CVA biplot space.*

	Group 1	Group 2	Group 3	Group 4
Group 1	0.0000	0.0957	0.1095	0.1018
Group 2	0.0957	0.0000	0.1360	0.0708
Group 3	0.1095	0.1360	0.0000	0.0740
Group 4	0.1018	0.0708	0.0740	0.0000

Table 5.5: *The group predictivities of the two-dimensional unweighted CVA biplot (with $\mathbf{C} = (\mathbf{I} - \frac{1}{4}\mathbf{1}\mathbf{1}')$) of the seventh simulated data set.*

Group 1	Group 2	Group 3	Group 4
0.95	0.81	0.87	0.41

Table 5.6: *The group contrast predictivities corresponding to the two-dimensional unweighted CVA biplot of the seventh simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group 1	–	0.807	0.864	0.917
Group 2	0.807	–	0.999	0.456
Group 3	0.864	0.999	–	0.500
Group 4	0.917	0.456	0.500	–

Given the very high group predictivity of Group 1 and the high group predictivity of Group 3, it can be expected that the Pythagorean distance between the two points representing the group centroids of these two groups in the canonical space will be accurately approximated by the Pythagorean distance between the corresponding two points in the two-dimensional CVA biplot space. This expectation is confirmed

5.7. AXIS PREDICTIVITIES, GROUP PREDICTIVITIES AND GROUP CONTRAST PREDICTIVITIES: AN ILLUSTRATIVE EXAMPLE 293

by the high group contrast predictivity associated with these two groups shown in Table 5.6. Given that the group predictivity of Group 2 is fairly high, it can be expected that the Pythagorean distances between the point representing this group's centroid in the canonical space and the points representing the centroids of Groups 1 and 3, will be approximated at least fairly accurately in the two-dimensional CVA biplot. The high group contrast predictivities associated with these two pairs of groups show that these approximations are indeed accurate. This implies that the relative magnitudes of the Pythagorean distances between the points representing the centroids of Groups 1, 2 and 3 in the two-dimensional CVA biplot accurately represent the relative magnitudes of the corresponding Mahalanobis distances in the measurement space.

Table 5.7: *The Pythagorean distances between the points representing the group centroids of the seventh simulated data set in the canonical space. (These distances are proportional to the Mahalanobis distances between the group centroids in the measurement space.)*

	Group 1	Group 2	Group 3	Group 4
Group 1	0.0000	0.1065	0.1177	0.1063
Group 2	0.1065	0.0000	0.1360	0.1049
Group 3	0.1177	0.1360	0.0000	0.1047
Group 4	0.1063	0.1049	0.1047	0.0000

In an attempt to gather information regarding with respect to which variables the group centroids are similar (or dissimilar), the positions of the points representing the group centroids in the CVA biplot relative to the individual biplot axes are considered together with the axis predictivities, which are provided in Table 5.8.

Table 5.8: *The axis predictivities of the two-dimensional unweighted CVA biplot (with $\mathbf{C} = (\mathbf{I} - \frac{1}{J}\mathbf{1}\mathbf{1}')$) of the seventh simulated data set.*

Var1	Var2	Var3	Var4	Var5
0.9539	0.9644	0.7564	0.6495	0.8768

Given that the axis predictivities of the variables *Var1*, *Var2* and *Var5* are high, the true similarities (and dissimilarities) between the group centroids with respect to these variables can be expected to be accurately represented in the two-dimensional CVA biplot in Figure 5.1. Inspection of Figure 5.1 suggests that the two group centroids associated with every pair of groups, except Groups 3 and 4, differ substantially with respect to *Var1*, *Var2* and *Var5*. The centroids of Groups 3 and 4 appear to be very similar with respect to all three the variables, *Var1*, *Var2* and *Var5* and consequently, the centroids of both Group 1 and Group 2 appear to be equally dissimilar from the centroids of Group 3 and Group 4. Both the centroid of Group 3 and that of Group 4 appear to differ to a greater extent from the centroid of Group 1 than from the centroid of Group 2 with respect to *Var1*,

Var2 and *Var3*. Upon consideration of the observed group centroids provided in Table 5.9 it is evident that this is an accurate representation of the true similarities and dissimilarities between the group centroids with respect to *Var2*. It also seems to be an accurate representation of the similarities and dissimilarities with respect to *Var1*, except for the fact that in reality the centroids of Groups 3 and 4 are not as similar as they appear in the biplot. As indicated by the relative positions of the points representing the group centroids with respect to the biplot axis representing *Var5*, the centroids of both Group 3 and Group 4 differ substantially from those of Group 1 and Group 2, although differing to a greater extent from Group 1. The difference between the centroids of Groups 1 and 2 with respect to *Var5* seems to be represented accurately in Figure 5.1. From the values in Table 5.9 it is evident that the centroids of Groups 3 and 4 differ from each other to approximately the same extent as those of Groups 1 and 2. This stands in contrast with what is suggested in Figure 5.1, namely that compared to the difference between the centroids of Group 1 and Group 2 with respect to *Var5*, the difference between the centroids of Groups 3 and 4 seems almost non-existent.

Table 5.9: *The observed group centroids of the seventh simulated data set.*

	Var1	Var2	Var3	Var4	Var5
Group 1	3.57	7.13	15.10	3.75	4.87
Group 2	7.02	7.99	16.96	5.37	6.46
Group 3	10.72	8.79	15.03	4.96	9.11
Group 4	8.87	9.19	15.25	8.73	10.96

Given the relatively low axis predictivities of the biplot axes representing the variables *Var3* and *Var4*, it is likely that at least some of the true similarities (and dissimilarities) between the group centroids with respect to these variables will not be represented accurately in the two-dimensional CVA biplot. In reality the centroid of Group 4 differs substantially from that of Group 2 with respect to *Var3* while it is very similar to the centroids of Group 1 and Group 3. The similarity between the centroids of Group 1 and 4 and the substantial difference between the centroids of Groups 2 and 4 with respect to *Var3* are accurately represented in Figure 5.1. However, the relative positions of the points representing the group centroids in the CVA biplot with respect to the biplot axis representing *Var3* suggest that Group 4 also differs substantially from Group 3 - it actually seems as though Group 4 differs slightly more from Group 3 with respect to *Var3* than it does from Group 2. Furthermore, Figure 5.1 incorrectly suggests that the centroids of Groups 1 and 3 differ to a moderate extent with respect to *Var3* and that the centroids of Groups 2 and 4 are very similar with respect to *Var4*. In reality the centroids of Groups 1 and 3 are very similar with respect to *Var3* and those of Groups 2 and 4 differ substantially with respect to *Var4*. The substantial difference between the centroids of Group 1 and Group 4 and the similarity between the centroids of Group 2 and Group 3 with respect to *Var4* (see Table 5.9) are however accurately represented in the CVA biplot. This demonstrates that the difference between the centroids

of two groups with respect to a particular variable can be accurately represented in the CVA biplot even if the biplot axis representing that variable has low axis predictivity.

Overall the results of this example were as expected - for the subset of groups with very high group predictivities, the true Pythagorean distances between the points representing the corresponding group centroids in the p -dimensional canonical space were accurately represented in the lower-dimensional CVA biplot whereas for the subset of groups with low group predictivities, at least some of the true Pythagorean distances between the group centroids were poorly represented. Furthermore, the true differences between the group centroids with respect to variables represented by biplot axes with high axis predictivities were accurately represented in the CVA biplot whereas at least some of the differences with respect to those variables represented by biplot axes with low axis predictivities were poorly represented in the biplot.

5.8 Within-group sample predictivities

5.8.1 Definition and properties

When investigating the group structure underlying a data set in a CVA biplot it is not only the accurate representation of the group centroids that is of importance but also the accurate representation of the individual samples within the groups that have been interpolated onto the CVA biplot. These samples allow for visualisation of the within-group dispersion and hence a more informative representation of the group structure.

Consider the decomposition of the individual samples into their respective group means and deviations from those group means:

$$\begin{aligned} \mathbf{X} &= \mathbf{G}\bar{\mathbf{X}} + (\mathbf{X} - \mathbf{G}\bar{\mathbf{X}}) \\ \therefore \mathbf{X} &= \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{X} + (\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{X} \end{aligned} \quad (5.8.1)$$

where \mathbf{G} is the $n \times J$ group indicator matrix defined in (4.2.7). It was explained in Chapter 4 that CVA (and the CVA biplot) is based on the decomposition of the total sum of squares,

$$\begin{aligned} \text{tr}\{\mathbf{X}'\mathbf{X}\} &= \text{tr}\{\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\} + \text{tr}\{\mathbf{W}\} \\ \therefore \text{tr}\{\mathbf{X}'\mathbf{X}\} &= \text{tr}\{\bar{\mathbf{X}}'\mathbf{G}'\mathbf{G}\bar{\mathbf{X}}\} + \text{tr}\{(\mathbf{X} - \mathbf{G}\bar{\mathbf{X}})'(\mathbf{X} - \mathbf{G}\bar{\mathbf{X}})\} . \end{aligned}$$

The validity of this decomposition of the total sum of squares follows from the fact that the decomposition in (5.8.1) exhibits Type A orthogonality. The fact that the decomposition in (5.8.1) exhibits Type A orthogonality follows from the

idempotency of the matrices $\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'$ and $\mathbf{I}-\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'$. The CVA biplot is designed to optimally represent the between-groups part of the total sum of squares, or equivalently, to optimally represent the between-groups variability in the data set at hand. It is however not designed to optimally represent the within-group dispersion - remember that the individual samples are not used in the construction of the CVA biplot but are interpolated onto the existing CVA biplot in which the group centroids are optimally represented. It is the accuracy of the representation of the within-group variability that forms the focus of this discussion.

To investigate the representation of the samples within their respective groups, the deviations of the samples from their corresponding group centroids are considered i.e the samples corrected for the group centroids are considered. Let \mathbf{K} denote the $n \times p$ matrix containing the deviations of the individual samples from their respective group centroids i.e.

$$\begin{aligned}\mathbf{K} &= \mathbf{X} - \mathbf{G}\bar{\mathbf{X}} \\ \therefore \mathbf{K} &= (\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{X}.\end{aligned}\tag{5.8.2}$$

Let \mathbf{H} denote the matrix $\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'$, then

$$\mathbf{K} = \mathbf{H}\mathbf{X}.$$

The approximation to \mathbf{K} which is produced by the r -dimensional predictive CVA biplot is given by:

$$\begin{aligned}\hat{\mathbf{K}} &= \hat{\mathbf{X}} - \mathbf{G}\hat{\bar{\mathbf{X}}} \\ &= \mathbf{X}\mathbf{M}_r\mathbf{M}^r - \mathbf{G}\bar{\mathbf{X}}\mathbf{M}_r\mathbf{M}^r \\ \longrightarrow \hat{\mathbf{K}} &= \mathbf{K}\mathbf{M}_r\mathbf{M}^r\end{aligned}$$

where

$$\mathbf{M} = \mathbf{L}\mathbf{V}$$

and \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$. Consider the following decomposition of the deviations of the individual samples from their

respective group means after transformation to the canonical space:

$$\mathbf{KL} = \widehat{\mathbf{K}}\mathbf{L} + (\mathbf{K} - \widehat{\mathbf{K}})\mathbf{L} \quad (5.8.3)$$

Due to the fact that the matrix \mathbf{V}_r in

$$\widehat{\mathbf{K}}\mathbf{L} = \mathbf{K}\mathbf{L}\mathbf{V}_r\mathbf{V}_r'$$

is a orthonormal matrix, the decomposition of \mathbf{KL} in (5.8.3) exhibits Type A orthogonality i.e.

$$\mathbf{KLL}'\mathbf{K}' = \widehat{\mathbf{K}}\mathbf{LL}'\widehat{\mathbf{K}}' + (\mathbf{K} - \widehat{\mathbf{K}})\mathbf{LL}'(\mathbf{K} - \widehat{\mathbf{K}})' \quad (5.8.4)$$

or equivalently, the decomposition of the deviation of the individual samples from their group centroids in the measurement space,

$$\mathbf{K} = \widehat{\mathbf{K}} + (\mathbf{K} - \widehat{\mathbf{K}})$$

exhibits Type A orthogonality in the metric \mathbf{W} (see Section 3.1) i.e.

$$\mathbf{KW}^{-1}\mathbf{K}' = \widehat{\mathbf{K}}\mathbf{W}^{-1}\widehat{\mathbf{K}}' + (\mathbf{K} - \widehat{\mathbf{K}})\mathbf{W}^{-1}(\mathbf{K} - \widehat{\mathbf{K}})'$$

Due to the validity of the orthogonal decomposition in (5.8.4) the ratio

$$\phi_i^{jW} = \frac{(\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)' \mathbf{W}^{-1} (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)}{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)} \quad (5.8.5)$$

$$\begin{aligned} &= \frac{\left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\bar{\mathbf{x}}}^j)' \mathbf{L} \right) \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\bar{\mathbf{x}}}^j)' \mathbf{L} \right)'}{\left((\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \right) \left((\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \right)'} \\ &= \frac{\left((\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \right) \mathbf{V}_r \mathbf{V}_r' \left((\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \right)'}{\left((\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \right) \mathbf{V} \mathbf{V}' \left((\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \right)'} \end{aligned} \quad (5.8.6)$$

$$\begin{aligned}
 &= \frac{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r' \mathbf{L}' (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)}{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{V} \mathbf{V}' \mathbf{L}' (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)} \\
 &= \frac{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{M}_r \mathbf{M}_r' (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)}{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{M} \mathbf{M}' (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)} \\
 \rightarrow \phi_i^{jW} &= \frac{\sum_{k=1}^r \left((\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{m}_{(k)} \right)^2}{\sum_{k=1}^p \left((\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{m}_{(k)} \right)^2} \quad (5.8.7)
 \end{aligned}$$

can be used to measure the accuracy of the approximation of the deviation $\mathbf{x}_i^j - \bar{\mathbf{x}}^j$ which is produced by the CVA biplot, $i \in [1 : n_j]$, $j \in [1 : J]$. Gardner-Lubbe *et al.* (2008) originally proposed the ratio in (5.8.5) for this purpose and called it the within-group sample predictivity of the i th sample belonging to the j th group or simply the within-group sample predictivity of \mathbf{x}_i^j .

The n -component vector of within-group sample predictivities is given by:

$$\begin{aligned}
 \boldsymbol{\phi}^W &= \text{diag}(\widehat{\mathbf{K}} \mathbf{W}^{-1} \widehat{\mathbf{K}}') [\text{diag}(\mathbf{K} \mathbf{W}^{-1} \mathbf{K}')]^{-1} \\
 \rightarrow \boldsymbol{\phi}^W &= \text{diag}(\widehat{\mathbf{K}} \mathbf{L} \mathbf{L}' \widehat{\mathbf{K}}') [\text{diag}(\mathbf{K} \mathbf{L} \mathbf{L}' \mathbf{K}')]^{-1} \\
 \rightarrow \boldsymbol{\phi}^W &= \text{diag}(\mathbf{K} \mathbf{M}_r \mathbf{M}_r' \mathbf{K}') [\text{diag}(\mathbf{K} \mathbf{M} \mathbf{M}' \mathbf{K}')]^{-1} .
 \end{aligned}$$

Equation (5.8.4) implies that ϕ_i^{jW} can also be expressed as:

$$\phi_i^{jW} = 1 - \frac{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j - (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j))' \mathbf{L} \mathbf{L}' (\mathbf{x}_i^j - \bar{\mathbf{x}}^j - (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j))}{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{L}' (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)} . \quad (5.8.8)$$

This expression of ϕ_i^{jW} shows that ϕ_i^{jW} is a decreasing function of the sum of the squared differences between $(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L}$ and $(\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)' \mathbf{L}$ or equivalently, ϕ_i^{jW} is a decreasing function of the difference between the true deviation, $\mathbf{x}_i^j - \bar{\mathbf{x}}^j$, and the predicted deviation, $\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j$, measured in the Mahalanobis distance metric. Note that since $(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L}$ is not necessarily equal to $(\mathbf{x}_k^l - \bar{\mathbf{x}}^l)' \mathbf{L}$, $\phi_i^{jW} > \phi_k^{lW}$ does not necessarily imply that

$$\left\| (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L} - (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)' \mathbf{L} \right\|^2 < \left\| (\mathbf{x}_k^l - \bar{\mathbf{x}}^l)' \mathbf{L} - (\hat{\mathbf{x}}_k^l - \hat{\bar{\mathbf{x}}}^l)' \mathbf{L} \right\|^2$$

The expression of ϕ_i^{jW} provided in (5.8.7) implies that ϕ_i^{jW} can only take on non-negative values and is a non-decreasing function of the dimension of the CVA biplot

space, r . This means that the sum of the squared differences between $(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L}$ and $(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r'$ is a non-increasing function of r . It is evident that the maximum value of ϕ_i^{jW} is one and that a sufficient condition for ϕ_i^{jW} to attain this maximum is $r = p$, $i \in [1 : n_j]$, $j \in [1 : J]$. Given that

$$\bar{\mathbf{X}} \mathbf{M}_K \mathbf{M}^K = \bar{\mathbf{X}}$$

it follows that

$$\mathbf{K} \mathbf{M}_K \mathbf{M}^K = \mathbf{K} \iff \mathbf{X} \mathbf{M}_K \mathbf{M}^K = \mathbf{X}.$$

However, since the column vectors of \mathbf{V}_K do not span the row space of $\mathbf{X} \mathbf{L}$, $\mathbf{X} \mathbf{M}_K \mathbf{M}^K$ is not equal to \mathbf{X} and hence $\mathbf{K} \mathbf{M}_K \mathbf{M}_K'$ is not equal to \mathbf{K} and the n within-group sample predictivities associated with the K -dimensional CVA biplot are not all equal to one. For the within-group sample predictivity of an individual sample, say \mathbf{x}_i^j , to equal one, a necessary and sufficient condition is that $(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L} \in \mathcal{V}(\mathbf{V}_r)$ i.e.

$$\begin{aligned} \phi_i^{jW} = 1 &\iff ((\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L}) \mathbf{V}_r \mathbf{V}_r' = (\mathbf{x}_i^j)' \mathbf{L} - (\bar{\mathbf{x}}^j)' \mathbf{L} \\ \therefore \phi_i^{jW} = 1 &\iff (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{L} \in \mathcal{V}(\mathbf{V}_r). \end{aligned}$$

It is evident that the minimum value of ϕ_i^{jW} is zero and that a sufficient condition for this minimum value to be attained is that both $\mathbf{x}_i^j' \mathbf{L}$ and $\bar{\mathbf{x}}^j' \mathbf{L}$ lie orthogonal to the biplot space, that is:

$$\begin{aligned} (\mathbf{x}_i^j)' \mathbf{L} \in \mathcal{V}^\perp(\mathbf{V}_r) &\longrightarrow (\hat{\mathbf{x}}_i^j)' \mathbf{L} = \mathbf{0}' \\ (\bar{\mathbf{x}}^j)' \mathbf{L} \in \mathcal{V}^\perp(\mathbf{V}_r) &\longrightarrow (\hat{\bar{\mathbf{x}}}^j)' \mathbf{L} = \mathbf{0}' \\ \therefore (\mathbf{x}_i^j)' \mathbf{L} \in \mathcal{V}^\perp(\mathbf{V}_r) \text{ and } (\bar{\mathbf{x}}^j)' \mathbf{L} \in \mathcal{V}^\perp(\mathbf{V}_r) &\longrightarrow \phi_i^{jW} = 0. \end{aligned}$$

A condition which is necessary and sufficient for ϕ_i^{jW} to equal zero is $\hat{\mathbf{x}}_i^j = \hat{\bar{\mathbf{x}}}^j$:

$$\begin{aligned} \phi_i^{jW} = 0 &\iff \hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j = \mathbf{0} \\ \therefore \phi_i^{jW} = 0 &\iff \hat{\mathbf{x}}_i^j = \hat{\bar{\mathbf{x}}}^j. \end{aligned}$$

Note that when $\mathbf{x}_i^j = \bar{\mathbf{x}}^j$, ϕ_i^{jW} is undefined.

Note that ϕ_i^{jW} associated with the K -dimensional weighted CVA biplot and that associated with each of two the K -dimensional unweighted CVA biplots will be identical due to the fact that

$$\mathbf{V}_K^{\mathbf{I}} (\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\mathbf{Cent}} (\mathbf{V}_K^{\mathbf{Cent}})' = \mathbf{V}_K^{\mathbf{N}} (\mathbf{V}_K^{\mathbf{N}})' .$$

Remember that since the last $p - K$ singular values of $\mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$ are all equal to zero, the vector defining the i th dimension of the CVA biplot, where $i > K$, is not unique. The within-group sample predictivity of the sample \mathbf{x}_k^j associated with the $(K + j)$ -dimensional CVA biplot, where $j \in [1 : p - K - 1]$, can therefore only be calculated once it has been decided which of the last $p - K$ right singular vectors of $\mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$ will be used to define the $(K + 1)$ th up to the $(K + j)$ th dimensions of the $(K + j)$ -dimensional CVA biplot, $k \in [1 : n_l]$, $\ell \in [1 : J]$ (all n the within-group sample predictivities of the p -dimensional CVA biplot are equal to one). What can however be calculated prior to making this decision is the within-group sample predictivity of the sample \mathbf{x}_k^l associated with the orthogonal complement of the K -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_K)$. The value of this measure is obtained by subtracting the within-group sample predictivity of \mathbf{x}_i^j associated with the K -dimensional CVA biplot space from one.

5.8.2 Within-group sample predictivities and the accuracy of distances represented in the CVA biplot.

Consider the squared Pythagorean distance between the two points representing the samples \mathbf{x}_i^j and \mathbf{x}_k^j in the CVA biplot space:

$$\begin{aligned} \left\| (\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^j)' \mathbf{L} \right\|^2 &= \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^j)' \mathbf{L} \right) \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^j)' \mathbf{L} \right) \\ &= \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} + (\hat{\mathbf{x}}^j)' \mathbf{L} - \left((\hat{\mathbf{x}}_k^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} + (\hat{\mathbf{x}}^j)' \mathbf{L} \right) \right) \\ &\quad \times \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} + (\hat{\mathbf{x}}^j)' \mathbf{L} - \left((\hat{\mathbf{x}}_k^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} + (\hat{\mathbf{x}}^j)' \mathbf{L} \right) \right)' \\ &= \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} - \left((\hat{\mathbf{x}}_k^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} \right) \right) \\ &\quad \times \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} - \left((\hat{\mathbf{x}}_k^j)' \mathbf{L} - (\hat{\mathbf{x}}^j)' \mathbf{L} \right) \right)' \\ &= (\hat{\mathbf{x}}_i^j - \hat{\mathbf{x}}^j - (\hat{\mathbf{x}}_k^j - \hat{\mathbf{x}}^j))' \mathbf{W}^{-1} (\hat{\mathbf{x}}_i^j - \hat{\mathbf{x}}^j - (\hat{\mathbf{x}}_k^j - \hat{\mathbf{x}}^j)) . \end{aligned}$$

It follows that if the deviations $\hat{\mathbf{x}}_i^j - \hat{\mathbf{x}}^j$ and $\hat{\mathbf{x}}_k^j - \hat{\mathbf{x}}^j$ are very accurate approximations of $\hat{\mathbf{x}}_i^j - \bar{\mathbf{x}}^j$ and $\hat{\mathbf{x}}_k^j - \bar{\mathbf{x}}^j$ respectively i.e. if both ϕ_i^{jW} and ϕ_k^{jW} are close to one, then

the Pythagorean distance in the CVA biplot,

$$\left\| (\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^j)' \mathbf{L} \right\|$$

will most likely be an accurate approximation of the corresponding Pythagorean distance in the p -dimensional canonical space,

$$\left\| (\mathbf{x}_i^j)' \mathbf{L} - (\mathbf{x}_k^j)' \mathbf{L} \right\| = \left\{ \frac{1}{n} (\mathbf{x}_i^j - \mathbf{x}_k^j)' \widehat{\Sigma}_W^{-1} (\mathbf{x}_i^j - \mathbf{x}_k^j) \right\}^{1/2}.$$

This implies that when the within-group sample predictivities of a number of samples belonging to the same group are very high, the relative magnitudes of the intersample Pythagorean distances in the CVA biplot will most likely be accurate approximations of the relative magnitudes of the corresponding intersample Mahalanobis distances in the p -dimensional measurement space. That is, when the within-group sample predictivities of a number of samples belonging to the same group are very high the true intersample relationships, as measured by the Mahalanobis distance metric, will most likely be accurately represented by the relative magnitudes of the intersample Pythagorean distances in the CVA biplot. This is however not true when the samples belong to different groups. This is evident upon consideration of the following expression of the squared Pythagorean distance between the points representing \mathbf{x}_i^j and \mathbf{x}_k^l , where $j \neq k$, in the CVA biplot:

$$\begin{aligned} \left\| (\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^l)' \mathbf{L} \right\|^2 &= \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^l)' \mathbf{L} \right) \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^l)' \mathbf{L} \right)' \\ &= \left(\left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\bar{\mathbf{x}}}^j)' \mathbf{L} \right) + (\hat{\bar{\mathbf{x}}}^j)' \mathbf{L} - \left((\hat{\mathbf{x}}_k^l)' \mathbf{L} - (\hat{\bar{\mathbf{x}}}^l)' \mathbf{L} \right) - (\hat{\bar{\mathbf{x}}}^l)' \mathbf{L} \right) \\ &\quad \times \left(\left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\bar{\mathbf{x}}}^j)' \mathbf{L} \right) + (\hat{\bar{\mathbf{x}}}^j)' \mathbf{L} - \left((\hat{\mathbf{x}}_k^l)' \mathbf{L} - (\hat{\bar{\mathbf{x}}}^l)' \mathbf{L} \right) - (\hat{\bar{\mathbf{x}}}^l)' \mathbf{L} \right)' \\ &= \left((\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j) + \hat{\bar{\mathbf{x}}}^j - (\hat{\mathbf{x}}_k^l - \hat{\bar{\mathbf{x}}}^l) - \hat{\bar{\mathbf{x}}}^l \right)' \mathbf{W}^{-1} \left((\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j) + \hat{\bar{\mathbf{x}}}^j \right. \\ &\quad \left. - (\hat{\mathbf{x}}_k^l - \hat{\bar{\mathbf{x}}}^l) - \hat{\bar{\mathbf{x}}}^l \right). \end{aligned}$$

It is evident that the accurate approximation of $\mathbf{x}_i^j - \bar{\mathbf{x}}^j$ and $\mathbf{x}_k^l - \bar{\mathbf{x}}^l$, as measured by the within-group sample predictivity measure, is not sufficient evidence to suggest that the Pythagorean distance $\left\| (\mathbf{x}_i^j)' \mathbf{L} - (\mathbf{x}_k^l)' \mathbf{L} \right\|$ will most likely be accurately represented in the CVA biplot. If in addition to the accurate approximation of $\mathbf{x}_i^j - \bar{\mathbf{x}}^j$ and $\mathbf{x}_k^l - \bar{\mathbf{x}}^l$, the group centroids $\bar{\mathbf{x}}^j$ and $\bar{\mathbf{x}}^l$ are accurately approximated as measured by the group predictivity measure, then it can be expected that the Pythagorean distance $\left\| (\mathbf{x}_i^j)' \mathbf{L} - (\mathbf{x}_k^l)' \mathbf{L} \right\|$ will be accurately represented in the CVA biplot. That is, if ϕ_i^{jW} , ϕ_k^{lW} , ψ^j and ψ^l are very high then it can be expected that the

Pythagorean distance $\left\| (\mathbf{x}_i^j)' \mathbf{L} - (\mathbf{x}_k^l)' \mathbf{L} \right\|$ will most likely be accurately represented in the CVA biplot. It should however be noted that, irrespective of whether two samples belong to the same group or not, the Pythagorean distance between the two samples in the p -dimensional canonical space can be accurately approximated in the CVA biplot even if one or both of the samples are poorly approximated.

5.8.3 Scale invariance

Let the $n \times p$ matrix containing the deviations of the individual samples from their corresponding group centroids that corresponds to the standardised measurements, $\mathbf{X}^* = \mathbf{X}\mathbf{A}^{-1}$, be denoted by \mathbf{K}^* and the approximation to this matrix be denoted by $\widehat{\mathbf{K}}^*$, then

$$\begin{aligned} \mathbf{K}^* &= \mathbf{H}\mathbf{X}^* = \mathbf{H}\mathbf{X}\mathbf{A}^{-1} \\ \text{and } \widehat{\mathbf{K}}^* &= \mathbf{H}\widehat{\mathbf{X}}^* = \mathbf{H}\widehat{\mathbf{X}}\mathbf{A}^{-1}. \end{aligned}$$

Upon substituting $\mathbf{H}\mathbf{X}\mathbf{A}^{-1}$ for \mathbf{K}^* , $\mathbf{H}\widehat{\mathbf{X}}\mathbf{A}^{-1}$ for $\widehat{\mathbf{K}}^*$ and $\mathbf{A}\mathbf{W}^{-1}\mathbf{A}$ for $(\mathbf{W}^*)^{-1}$ in the expression of the n -component vector of within-group sample predictivities corresponding to the CVA biplot constructed from the standardised measurements,

$$\phi^{W^*} = \text{diag} \left(\widehat{\mathbf{K}}^* (\mathbf{W}^*)^{-1} (\widehat{\mathbf{K}}^*)' \right) \left[\text{diag} \left(\mathbf{K}^* (\mathbf{W}^*)^{-1} (\mathbf{K}^*)' \right) \right]^{-1}$$

shows that $\phi^{W^*} = \phi^W$:

$$\begin{aligned} \phi^{W^*} &= \text{diag} \left(\mathbf{H}\widehat{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}\mathbf{A}^{-1}\widehat{\mathbf{X}}'\mathbf{H} \right) \left[\text{diag} \left(\mathbf{H}\mathbf{X}\mathbf{A}^{-1}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{X}' \right) \right]^{-1} \\ &= \text{diag} \left(\mathbf{H}\widehat{\mathbf{X}}\mathbf{W}^{-1}\widehat{\mathbf{X}}'\mathbf{H} \right) \left[\text{diag} \left(\mathbf{H}\mathbf{X}\mathbf{W}^{-1}\mathbf{X}' \right) \right]^{-1} \\ &= \text{diag} \left(\widehat{\mathbf{K}}\mathbf{W}^{-1}\widehat{\mathbf{K}}' \right) \left[\text{diag} \left(\mathbf{K}\mathbf{W}^{-1}\mathbf{K}' \right) \right]^{-1} \\ \longrightarrow \phi^{W^*} &= \phi^W. \end{aligned}$$

Substituting any $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in the above equations shows that the within-group sample predictivity measure is invariant to all non-singular linear transformations of the form $\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x}$.

5.8.4 Within-group sample predictivities of ‘new’ samples

Given a set of m ‘new’ samples that have been interpolated onto an existing CVA biplot, let \mathbf{X}^{new} denote the $m \times p$ matrix with i th row vector giving the measurements of the i th new sample, rescaled to be in the same scales as the measurements in

the matrix \mathbf{X} . That is, the i th row of \mathbf{X}^{new} gives the measurement vector after the overall mean calculated from the set of observed (uncentred) original samples has been subtracted. Let \mathbf{G}^{new} denote the group indicator matrix corresponding to \mathbf{X}^{new} indicating either the true group memberships of the new samples if those are known or the group memberships based on a CVA classification rule (for example the exact CVA classification regions or the approximate classification regions associated with the CVA biplot). The within-group sample predictivities associated with new samples that have been interpolated onto the CVA biplot can be calculated in exactly the same way as for the original samples. Let

$$\mathbf{K}^{\text{new}} = \mathbf{X}^{\text{new}} - \mathbf{G}^{\text{new}} (\mathbf{G}'\mathbf{G})^{-1} \mathbf{G}'\mathbf{X}$$

$$\text{and } \widehat{\mathbf{K}}^{\text{new}} = \mathbf{K}^{\text{new}} \mathbf{M}_r \mathbf{M}_r'$$

where $\mathbf{M} = \mathbf{L}\mathbf{V}$ and \mathbf{V} is the matrix of right singular vectors of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$. Due to the fact that the matrix \mathbf{V}_r in

$$\widehat{\mathbf{K}}^{\text{new}}\mathbf{L} = \mathbf{K}^{\text{new}}\mathbf{L}\mathbf{V}_r\mathbf{V}_r' \quad (5.8.9)$$

is an orthonormal matrix, the decomposition of $\mathbf{K}^{\text{new}}\mathbf{L}$,

$$\mathbf{K}^{\text{new}}\mathbf{L} = \widehat{\mathbf{K}}^{\text{new}}\mathbf{L} + \mathbf{K}^{\text{new}}\mathbf{L} - \widehat{\mathbf{K}}^{\text{new}}\mathbf{L}$$

exhibits Type A orthogonality. The within-group sample predictivity associated with the i th new sample can therefore be defined as:

$$\phi_i^{W(\text{new})} = \frac{\left[\widehat{\mathbf{K}}^{\text{new}} \mathbf{W}^{-1} (\widehat{\mathbf{K}}^{\text{new}})' \right]_{ii}}{\left[\mathbf{K}^{\text{new}} \mathbf{W}^{-1} (\mathbf{K}^{\text{new}})' \right]_{ii}}. \quad (5.8.10)$$

5.9 The overall within-group sample predictivity associated with a group

5.9.1 Definition and properties

The fact that the decomposition of the matrix \mathbf{K} into $\widehat{\mathbf{K}}$ and $\mathbf{K} - \widehat{\mathbf{K}}$ exhibits Type A orthogonality in the metric \mathbf{W} implies that

$$\begin{aligned} \text{tr} \{ \mathbf{K} \mathbf{W}^{-1} \mathbf{K}' \} &= \text{tr} \{ \widehat{\mathbf{K}} \mathbf{W}^{-1} \widehat{\mathbf{K}}' \} + \text{tr} \{ (\mathbf{K} - \widehat{\mathbf{K}}) \mathbf{W}^{-1} (\mathbf{K} - \widehat{\mathbf{K}})' \} \\ \therefore \sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j) &= \sum_{i=1}^{n_j} (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)' \mathbf{W}^{-1} (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j) \\ &\quad + \sum_{i=1}^{n_j} ((\mathbf{x}_i^j - \bar{\mathbf{x}}^j) - (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j))' \mathbf{W}^{-1} ((\mathbf{x}_i^j - \bar{\mathbf{x}}^j) \\ &\quad - (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)) \quad \forall j \in [1 : J] \end{aligned}$$

This implies that the ratio,

$$\begin{aligned} \phi^{jW} &= \frac{\sum_{i=1}^{n_j} (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)' \mathbf{W}^{-1} (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)} \\ \rightarrow \phi^{jW} &= 1 - \frac{\sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j - (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j))' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j - (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j))}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)} \end{aligned}$$

can be used to measure the overall accuracy of the approximations to the deviations of the samples belonging to the j th group from the j th group centroid. The ratio, ϕ^{jW} , will henceforth be referred to as the overall within-group sample predictivity of the j th group, or the j th overall within-group sample predictivity for short. The j th overall within-group sample predictivity is equal to the proportion of the total sample variance within the j th group of canonical observations, with variance being interpreted in terms of deviation from the j th group centroid, that is accounted for in the CVA biplot. The j th overall within-group sample predictivity can therefore be interpreted as a measure which assesses the overall quality of the representation of the spread of the individual samples of the j th group around the centroid of that group.

The overall within-group sample predictivity of the j th group can be expressed as a weighted average of the individual within-group sample predictivities:

$$\phi_i^{jW} = \frac{(\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)' \mathbf{W}^{-1} (\hat{\mathbf{x}}_i^j - \hat{\bar{\mathbf{x}}}^j)}{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)}$$

$$\begin{aligned}
 &\longrightarrow (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j) = \phi_i^{jW} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j) \\
 &\longrightarrow \phi^{jW} = \frac{\sum_{i=1}^{n_j} \phi_i^{jW} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)} \\
 &\longrightarrow \phi^{jW} = \sum_{i=1}^{n_j} \phi_i^{jW} w_i^j \tag{5.9.1} \\
 &\text{where } w_i^j = \frac{(\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \bar{\mathbf{x}}^j)}.
 \end{aligned}$$

Given that ϕ_i^{jW} and w_i^j are non-negative values for all $i \in [1 : n_j]$, it follows that ϕ^{jW} can only take on non-negative values, $j \in [1 : J]$. Since ϕ_i^{jW} is a non-decreasing function of r and w_i^j is a constant function of r , ϕ^{jW} is a non-decreasing function of the dimension of the biplot space, r . The j th overall within-group sample predictivity will equal its maximum value of one if and only if $\phi_i^{jW} = 1 \forall i \in [1 : n_j]$. It is evident that ϕ^{jW} will equal its minimum value of zero if the sample predictivity of each of the n_j the samples belonging to the j th group equal zero, that is:

$$\phi^{jW} = 0 \iff \phi_i^{jW} = 0 \forall i \in [1 : n_j].$$

Consider equation (5.9.1). Since ϕ_i^{jW} and w_i^j are identical for the weighted and two unweighted K -dimensional CVA biplots, the overall within-group sample predictivity of the j th group, ϕ^{jW} , will also be identical for the three K -dimensional CVA biplots.

5.9.2 Scale Invariance

Substituting \mathbf{XF} for \mathbf{X}^* , $\widehat{\mathbf{X}}\mathbf{F}$ for $\widehat{\mathbf{X}}^*$, $\overline{\mathbf{X}}\mathbf{F}$ for $\overline{\mathbf{X}}^*$, $\widehat{\widehat{\mathbf{X}}}\mathbf{F}$ for $\widehat{\widehat{\mathbf{X}}}^*$ and $(\mathbf{F}^{-1})' \mathbf{W}^{-1} \mathbf{F}^{-1}$ for $(\mathbf{W}^*)^{-1}$ in

$$\phi^{jW^*} = \frac{\sum_{i=1}^{n_j} (\widehat{\mathbf{x}}_i^{j*} - \widehat{\widehat{\mathbf{x}}}^{j*})' (\mathbf{W}^*)^{-1} (\widehat{\mathbf{x}}_i^{j*} - \widehat{\widehat{\mathbf{x}}}^{j*})}{\sum_{i=1}^{n_j} (\mathbf{x}_i^{j*} - \overline{\mathbf{x}}^{j*})' (\mathbf{W}^*)^{-1} (\mathbf{x}_i^{j*} - \overline{\mathbf{x}}^{j*})}$$

shows that the overall within-group sample predictivity is invariant to all non-singular linear transformations of the form $\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x}$. This holds in particular also for $\mathbf{F} = \mathbf{A}^{-1}$ which implies that the j th overall within-group sample predictivity associated with the CVA biplot constructed from the standardised measurements, \mathbf{XA}^{-1} , is identical to that associated with the CVA biplot constructed from the unstandardised measurements.

5.10 Mixed contrast predictivities

5.10.1 Definition and Properties

In order for the within-group dispersion of a group to be accurately represented in the CVA biplot the relative magnitudes of the Mahalanobis distances between the centroid and individual samples of that group in the p -dimensional measurement space need to be accurately represented by the relative magnitudes of the Pythagorean distances between the points representing the particular group centroid and individual samples in the CVA biplot. That is, the Pythagorean distances between the point representing the group centroid in the p -dimensional canonical space and the points representing the individual samples belonging to that group need to be accurately approximated in the CVA biplot.

Consider the following orthogonal decomposition of the squared Pythagorean distance between the points representing the i th sample and the j th group centroid in the p -dimensional canonical space:

$$\|e'_i \mathbf{XL} - e'_j \bar{\mathbf{XL}}\|^2 = \|e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}}\|^2 + \left\| \left(e'_i \mathbf{XL} - e'_j \bar{\mathbf{XL}} - \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \right) \right\|^2 \quad (5.10.1)$$

where

$$\begin{aligned} \widehat{\bar{\mathbf{XL}}} &= \bar{\mathbf{X}}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r \\ \text{and } \widehat{\mathbf{XL}} &= \mathbf{X}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r \end{aligned}$$

and \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$. The decomposition in (5.10.1) holds due to the fact that the cross product term,

$$\left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \left(e'_i \mathbf{XL} - e'_j \bar{\mathbf{XL}} - \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \right)'$$

is equal to zero:

$$\begin{aligned} & \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \left(e'_i \mathbf{XL} - e'_j \bar{\mathbf{XL}} - \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \right)' \\ &= \left(e'_i \mathbf{X} - e'_j \bar{\mathbf{X}} \right) \mathbf{L}\mathbf{V}_r\mathbf{V}'_r\mathbf{L}' \left(e'_i \mathbf{X} - e'_j \bar{\mathbf{X}} \right)' - \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right)' \\ &= \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right)' - \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right) \left(e'_i \widehat{\mathbf{XL}} - e'_j \widehat{\bar{\mathbf{XL}}} \right)' \\ &= 0. \end{aligned}$$

The accuracy with which the Pythagorean distance between the i th sample and the j th group centroid in the p -dimensional canonical space is approximated by the corresponding Pythagorean distance in the CVA biplot can therefore (for all three choices of \mathbf{C}) be measured by the ratio,

$$\zeta_i^j = \frac{\|\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}\|^2}{\|\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}\|^2} \quad (5.10.2)$$

$$= 1 - \frac{\|\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L} - (\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L})\|^2}{\|\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}\|^2} \quad (5.10.3)$$

which will henceforth be referred to as the (ij) th mixed contrast predictivity, $i \in [1:n]$, $j \in [J]$.

The mixed contrast predictivity measure has a minimum value of zero which it will attain if and only if

$$\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} = \mathbf{0}'$$

and a maximum value of one which it will attain if and only if

$$\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L} = \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}.$$

The conditions

$$\begin{aligned} \mathbf{e}'_i \mathbf{X}\mathbf{L} &= \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} \\ \text{and } \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L} &= \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \end{aligned}$$

are sufficient but not necessary for ζ_i^j to attain the value of one. This implies that accurate approximations of both the sample $\mathbf{e}'_i \mathbf{X}$ and the j th group centroid are sufficient but not necessary for the Pythagorean distance $\|\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}}\mathbf{L}\|$ to be accurately approximated in the CVA biplot. That is, when both the sample predictivity of $\mathbf{e}'_i \mathbf{X}$ and the j th group predictivity are close to one it can be expected that ζ_i^j will be close to one, however a high value of ζ_i^j does not imply that both or even one of the sample predictivity of $\mathbf{e}'_i \mathbf{X}$ and the j th group predictivity is high.

Note that since

$$\begin{aligned} \left\| \mathbf{e}'_i \widehat{\mathbf{X}} \mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}} \mathbf{L} \right\|^2 &= \frac{1}{n} \left(\mathbf{e}'_i \widehat{\mathbf{X}} - \mathbf{e}'_j \widehat{\mathbf{X}} \right) \widehat{\Sigma}_W^{-1} \left(\mathbf{e}'_i \widehat{\mathbf{X}} - \mathbf{e}'_j \widehat{\mathbf{X}} \right)' \\ \text{and } \left\| \mathbf{e}'_i \mathbf{X} \mathbf{L} - \mathbf{e}'_j \overline{\mathbf{X}} \mathbf{L} \right\|^2 &= \frac{1}{n} \left(\mathbf{e}'_i \mathbf{X} - \mathbf{e}'_j \overline{\mathbf{X}} \right) \widehat{\Sigma}_W^{-1} \left(\mathbf{e}'_i \mathbf{X} - \mathbf{e}'_j \overline{\mathbf{X}} \right)' \end{aligned}$$

ζ_i^j is also equal to the squared approximated Mahalanobis distance between the i th sample and j th group centroid, expressed as a proportion of the squared true Mahalanobis distance between that sample and group centroid. Hence, when all J mixed contrast predictivities associated with the i th sample are close to one, the relative magnitudes of the Mahalanobis distances between that sample and the J group centroids in the measurement space are accurately represented by the relative magnitudes of the corresponding Pythagorean distances in the CVA biplot. This means that if the i th sample is classified as belonging to the j th group by the exact classification regions of the p -dimensional CVA biplot (or equivalently the exact classification regions of the K -dimensional CVA biplot) and ζ_i^j is close to one for all $j \in [1 : J]$, then the i th sample will most likely be classified as belonging to the j th group by the approximate classification regions of the CVA biplot, $i \in [1 : n_j]$, $j \in [1 : J]$. If however a subset of the J mixed contrast predictivities associated with the i th sample are low, it is likely that the classification of the i th sample based on the classification regions of the CVA biplot will differ from that based on the exact (K -dimensional) CVA classification regions. Furthermore, when all the mixed contrast predictivities associated with the j th group are close to one, the within-group dispersion of the j th group is very accurately represented in the CVA biplot. When a substantial proportion of the mixed contrast predictivities associated with the j th group are low, the within-group dispersion of the j th group is poorly represented in the CVA biplot.

5.10.2 Scale invariance

Consider a non-singular linear transformations of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$ where \mathbf{F} is a $p \times p$ non-singular matrix. The (ij) th mixed contrast predictivity associated with the CVA biplot constructed from the transformed measurements $\mathbf{X}\mathbf{F}$ is given by

$$\zeta_i^{j*} = \frac{\left\| \mathbf{e}'_i \widehat{\mathbf{X}}^* \mathbf{L}^* - \mathbf{e}'_j \widehat{\mathbf{X}}^* \mathbf{L}^* \right\|^2}{\left\| \mathbf{e}'_i \mathbf{X}^* \mathbf{L}^* - \mathbf{e}'_j \overline{\mathbf{X}}^* \mathbf{L}^* \right\|^2}$$

where $\mathbf{X}^* = \mathbf{X}\mathbf{F}$, $\widehat{\mathbf{X}}^* = \widehat{\mathbf{X}}\mathbf{F}$, $\overline{\mathbf{X}}^* = \overline{\mathbf{X}}\mathbf{F}$, $\widehat{\mathbf{X}}^* = \widehat{\mathbf{X}}\mathbf{F}$ and $\mathbf{L}^* = \mathbf{F}^{-1}\mathbf{L}$. It is evident that $\zeta_i^{j*} = \zeta_i^j$. That is the mixed contrast predictivity is invariant to all non-singular linear transformations of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$. In particular the (ij) th mixed con-

trast predictivity associated with the CVA biplot constructed from the standardised measurements is identical to that associated with the CVA biplot constructed from the unstandardised measurements.

5.11 Sample predictivities

5.11.1 Definition and properties

Consider again the decomposition of the individual samples in equation (5.8.1). This decomposition shows that approximating the measurements of a sample consists of two parts, namely approximating the corresponding group centroid and approximating the deviation of the sample from that group centroid. It is evident that if both the deviation and the group centroid are accurately predicted, then the sample will most likely be accurately predicted i.e. if both the corresponding within-group sample predictivity and group predictivity are very high then the sample will most likely be accurately predicted. However, it is possible for a sample to be accurately approximated in the CVA biplot even if one or both of the sample's deviation from the corresponding group centroid and the group centroid itself are poorly approximated in the biplot. It is also possible for a sample to be poorly approximated even if both its deviation from the corresponding group centroid and the group centroid itself are relatively accurately approximated.

To investigate the approximation to the measurements of the individual samples, consider the decomposition of the matrix of individual canonical samples, \mathbf{XL} ,

$$\mathbf{XL} = \widehat{\mathbf{XL}} + (\mathbf{XL} - \widehat{\mathbf{XL}}) \quad (5.11.1)$$

where

$$\widehat{\mathbf{XL}} = \mathbf{XLV}_r\mathbf{V}_r' \quad (5.11.2)$$

and \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2}\overline{\mathbf{XL}}$. The decomposition in (5.11.1) exhibits Type A orthogonality due to the fact that the matrix \mathbf{V}_r in equation (5.11.2) is an orthonormal matrix (see Section 3.1). This means that:

$$\begin{aligned} \mathbf{XLL}'\mathbf{X}' &= \widehat{\mathbf{XLL}}'\widehat{\mathbf{X}}' + (\mathbf{XL} - \widehat{\mathbf{XL}})(\mathbf{XL} - \widehat{\mathbf{XL}})' \\ \therefore \mathbf{XW}^{-1}\mathbf{X}' &= \widehat{\mathbf{XW}}^{-1}\widehat{\mathbf{X}}' + (\mathbf{X} - \widehat{\mathbf{X}})\mathbf{W}^{-1}(\mathbf{X} - \widehat{\mathbf{X}})' \end{aligned} \quad (5.11.3)$$

Due to the validity of equation (5.11.3) the overall accuracy of the approximations of the measurements of the i th sample of the j th group, \mathbf{x}_i^j , that are read off from

the predictive CVA biplot axes can be measured by the ratio

$$\phi_i^j = \frac{(\hat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \hat{\mathbf{x}}_i^j}{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \quad (5.11.4)$$

$$= \frac{(\hat{\mathbf{x}}_i^j)' \mathbf{W}^{-1} \hat{\mathbf{x}}_i^j}{(\mathbf{x}_i^j)' \mathbf{W}^{-1} \mathbf{x}_i^j}. \quad (5.11.5)$$

The ratio in (5.11.5) will henceforth be referred to as the sample predictivity of the i th sample of the j th group or simply the sample predictivity of \mathbf{x}_i^j , $i \in [1:n_j]$, $j \in [1:J]$. Note that the definition of the sample predictivity measure is based on the decomposition of a different total sums of squares than the within-group sample predictivity measure. While the within-group sample predictivity measure was based on the decomposition of the total sums of squares associated with the n canonical observations corrected for the group centroids, the sample predictivity measure is based on the decomposition of the total uncorrected sums of squares associated with the n canonical observations.

It is shown below that the sample predictivity of the sample \mathbf{x}_i^j associated with the r -dimensional CVA biplot is equal to the ratio of the squared Pythagorean distance between the origin and the point representing \mathbf{x}_i^j in the r -dimensional CVA biplot, $(\hat{\mathbf{x}}_i^j)' \mathbf{L} = (\mathbf{x}_i^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r'$, to the squared Pythagorean distance between the origin and the point representing \mathbf{x}_i^j in the p -dimensional canonical space, $(\mathbf{x}_i^j)' \mathbf{L}$:

$$\begin{aligned} \phi_i^j &= \frac{(\hat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \hat{\mathbf{x}}_i^j}{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \\ \rightarrow \phi_i^j &= \frac{\|\hat{\mathbf{x}}_i^j \mathbf{L}\|^2}{\|\mathbf{x}_i^j \mathbf{L}\|^2} \\ \rightarrow \phi_i^j &= \frac{\|(\mathbf{x}_i^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r'\|^2}{\|(\mathbf{x}_i^j)' \mathbf{L}\|^2}. \end{aligned} \quad (5.11.6)$$

Equation (5.11.3) implies that ϕ_i^j can be expressed as a decreasing function of the Pythagorean distance between the point representing \mathbf{x}_i^j in the p -dimensional canonical space and the point representing \mathbf{x}_i^j in the r -dimensional CVA biplot space:

$$\phi_i^j = 1 - \frac{\|(\mathbf{x}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_i^j)' \mathbf{L}\|^2}{\|(\mathbf{x}_i^j)' \mathbf{L}\|^2}. \quad (5.11.7)$$

Given that $\|(\mathbf{x}_i^j)'\mathbf{L}\|$ is not necessarily equal to $\|(\mathbf{x}_k^l)'\mathbf{L}\|$, $\phi_i^j > \phi_k^j$ does not necessarily imply that

$$\|(\mathbf{x}_i^j)'\mathbf{L} - (\hat{\mathbf{x}}_i^j)'\mathbf{L}\| < \|(\mathbf{x}_k^l)'\mathbf{L} - (\hat{\mathbf{x}}_k^l)'\mathbf{L}\|.$$

It is shown below that ϕ_i^j can also be expressed as the square of the cosine of the angle between the two vectors $(\mathbf{x}_i^j)'\mathbf{L}$ and $(\mathbf{x}_i^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'$, both emanating from the origin:

$$\begin{aligned}\phi_i^j &= \frac{(\mathbf{x}_i^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'\mathbf{V}_r\mathbf{V}_r'\mathbf{L}'\mathbf{x}_i^j}{\|(\mathbf{x}_i^j)'\mathbf{L}\|^2} \\ \rightarrow \phi_i^j &= \left\{ \frac{(\mathbf{x}_i^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'\mathbf{L}'\mathbf{x}_i^j}{\|(\mathbf{x}_i^j)'\mathbf{L}\| \|(\mathbf{x}_i^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'\|} \right\}^2 \\ \rightarrow \phi_i^j &= \cos^2(\theta_i^j)\end{aligned}\tag{5.11.8}$$

where θ_i^j is the angle between the two vectors $(\mathbf{x}_i^j)'\mathbf{L}$ and $(\mathbf{x}_i^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'$. It is evident from equation (5.11.8) that as θ_i^j increases from zero degrees to 90 degrees, ϕ_i^j decreases from one to zero. Equation (5.11.8) implies that if $\phi_i^j > \phi_k^l$, the angle between the vectors $(\mathbf{x}_i^j)'\mathbf{L}$ and $(\mathbf{x}_i^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'$ is necessarily smaller than the angle between the vectors $(\mathbf{x}_k^l)'\mathbf{L}$ and $(\mathbf{x}_k^l)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'$.

Given that ϕ_i^j is a ratio of sums of squared values, it can only take on non-negative values. From the expressions for ϕ_i^j given in (5.11.6) and (5.11.8), it is evident that ϕ_i^j has a minimum value of zero which it will attain if and only if the vector $(\mathbf{x}_i^j)'\mathbf{L}$ lies orthogonal to the CVA biplot space i.e.

$$\phi_i^j = 0 \leftrightarrow (\mathbf{x}_i^j)'\mathbf{L} \in \mathcal{L}^\perp.$$

The following expression of ϕ_i^j shows that it is a non-decreasing function of the dimension of the CVA biplot space, r :

$$\phi_i^j = \frac{\|(\mathbf{x}_i^j)'\mathbf{L}\mathbf{V}_r\mathbf{V}_r'\|^2}{\|(\mathbf{x}_i^j)'\mathbf{L}\|^2}$$

$$\begin{aligned}
 &= \frac{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r' \mathbf{V}_r \mathbf{V}_r' \mathbf{L}' \mathbf{x}_i^j}{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \\
 &= \frac{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r' \mathbf{L}' \mathbf{x}_i^j}{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{V} \mathbf{V}' \mathbf{L}' \mathbf{x}_i^j} \\
 \longrightarrow \phi_i^j &= \frac{\sum_{k=1}^r \left((\mathbf{x}_i^j)' \mathbf{m}_{(k)} \right)^2}{\sum_{k=1}^p \left((\mathbf{x}_i^j)' \mathbf{m}_{(k)} \right)^2}.
 \end{aligned}$$

It follows that ϕ_i^j will necessarily equal its maximum value of one when $r = p$. This condition is however not necessary for an individual sample predictivity to equal one. Equations (5.11.7) and (5.11.8) imply that the sample \mathbf{x}_i^j will have unit sample predictivity if and only if

$$\begin{aligned}
 \therefore (\hat{\mathbf{x}}_i^j)' \mathbf{L} &= (\mathbf{x}_i^j)' \mathbf{L} \\
 \therefore (\mathbf{x}_i^j)' \mathbf{L} &\in \mathcal{L} = \mathcal{V}(\mathbf{V}_r).
 \end{aligned}$$

Since the first K right singular vectors of $\mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$ do not span the row space of $\mathbf{X} \mathbf{L}$, the n sample predictivities corresponding to the K -dimensional CVA biplot will never all be equal to one. A condition which is necessary and sufficient for all n sample predictivities to equal one is

$$\begin{aligned}
 (\mathbf{x}_i^j)' \mathbf{L} &\in \mathcal{L} = \mathcal{V}(\mathbf{V}_r) \quad \forall i \in [1 : n_j], j \in [1 : J] \\
 \text{i.e.} \quad r &= p.
 \end{aligned}$$

Recall that the K -dimensional CVA biplot space is equal to the row space of $\bar{\mathbf{X}} \mathbf{L}$ irrespective of the \mathbf{C} -matrix used in the construction of the biplot and that consequently

$$\mathbf{V}_K^{\mathbf{I}} (\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\text{Cent}} (\mathbf{V}_K^{\text{Cent}})' = \mathbf{V}_K^{\mathbf{N}} (\mathbf{V}_K^{\mathbf{N}})'.$$

As a result, the sample predictivity of \mathbf{x}_k^l corresponding to the K -dimensional weighted CVA biplot and those corresponding to the two K -dimensional unweighted CVA biplots, will be identical. Like the within-group sample predictivity of the sample \mathbf{x}_k^l , the sample predictivity of sample \mathbf{x}_k^l associated with the $(K + j)$ -dimensional CVA biplot, where $j \in [1 : p - K - 1]$, can only be calculated once it has been decided which of the last $p - K$ right singular vectors of $\mathbf{C}^{1/2} \bar{\mathbf{X}} \mathbf{L}$ will be used to de-

fine the $(K + 1)$ th to $(K + j)$ th dimensions of the $(K + j)$ -dimensional CVA biplot, $k \in [1 : n_l]$, $\ell \in [1 : J]$. What can however be calculated prior to making this decision is the sample predictivity of the sample \mathbf{x}_k^l associated with the orthogonal complement of the K -dimensional CVA biplot space, which is obtained by subtracting the sample predictivity of \mathbf{x}_k^l associated with the K -dimensional CVA biplot space from one. This value can be interpreted as a measure of the amount of information that is lost due to representing \mathbf{x}_k^l in K -dimensional instead of p -dimensional space.

The following expression of the sample predictivity of the sample \mathbf{x}_i^j clearly shows that the overall accuracy of the approximations to the measurements of \mathbf{x}_i^j , as measured by ϕ_i^j , is dependent on the overall accuracy of the approximations to the measurements of $\bar{\mathbf{x}}^j$:

$$\begin{aligned}
 \phi_i^j &= \frac{(\hat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \hat{\mathbf{x}}_i^j}{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \\
 \rightarrow \phi_i^j &= \frac{\left(((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\bar{\mathbf{x}}^j})' \mathbf{L}) + (\hat{\bar{\mathbf{x}}^j})' \mathbf{L} \right) \left(((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\bar{\mathbf{x}}^j})' \mathbf{L}) + (\hat{\bar{\mathbf{x}}^j})' \mathbf{L} \right)'}{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j}. \quad (5.11.9)
 \end{aligned}$$

Equation (5.11.9) shows that approximating the measurements of a sample consists of two parts, namely approximating the corresponding group centroid and approximating the deviation of the sample from that group centroid. It is evident that if both the deviation $\mathbf{x}_i^j - \bar{\mathbf{x}}^j$ and the group centroid $\bar{\mathbf{x}}^j$ are very accurately predicted as measured by the within-group sample predictivity measure and the group predictivity measure, then the sample \mathbf{x}_i^j will most likely be accurately predicted i.e. if ϕ_i^{jW} and ψ^j are very high then ϕ_i^j will most likely be high. However ϕ_i^j can be very high even if one or both of ϕ_i^{jW} and ψ^j are low. If for example the j th group centroid is very accurately approximated in the CVA biplot, samples belonging to the j th group may have quite high sample predictivities even if the deviations of those samples from the j th group centroid are poorly approximated in the biplot. A low sample predictivity also does not imply that both the deviation of the sample from the corresponding group centroid and the group centroid are poorly predicted in the CVA biplot. If for example the j th group centroid is very poorly approximated in the CVA biplot, then samples belonging to the j th group may have quite low sample predictivities even if the deviations of those samples from the j th group centroid are accurately approximated in the biplot.

Some of the concepts discussed above will now be demonstrated at the hand of the fourth simulated data set that was introduced in Section 4.8. Table 5.10 contains the within-group sample predictivities and sample predictivities of the fifth, 22nd, 351st and 366th samples of the fourth simulated data set, corresponding to the two-dimensional weighted CVA biplot. Table 5.10 also contains for each of the four samples, the group predictivity of the corresponding group. Samples five and 22 belong to Group 1, which has a very high group predictivity while samples 351 and 366 belong to Group 4, which has a very low group predictivity.

Table 5.10: *The within-group sample predictivity, sample predictivity and group predictivity of the corresponding group centroid of the fifth, 22nd, 351st and 366th sample of the fourth simulated data set, corresponding to the two-dimensional weighted CVA biplot.*

Sample:	5	22	351	366
Within-group sample predictivity	0.526	0.828	0.922	0.257
Sample predictivity	0.812	0.874	0.472	0.426
Group predictivity	0.998	0.998	0.497	0.497

Note that although both samples five and 22 have high sample predictivities, only sample 22 has a high within-group sample predictivity. Samples 351 and 366 on the other hand both have low sample predictivities while only sample 366 has a low within-group sample predictivity - sample 351 has a very high sample predictivity (0.922). From Table 5.10 there seems to be no association between the within-group sample predictivity of a sample and the group predictivity of the corresponding group. That is, the accuracy with which the CVA biplot predicts the deviation of a sample from the corresponding group centroid, is not associated with the accuracy with which the biplot predicts that group centroid. The within-group sample predictivity measure differs from the sample predictivity measure in this regard.

Recall that when $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$ the j th group predictivity is a decreasing function of the size of the angle between the two vectors $(\bar{\mathbf{x}}^j)' \mathbf{L}$ and $(\bar{\mathbf{x}}^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r'$. Since the canonical centroid of a group is a measure of the central locality of the group of individual canonical observations, the point representing the j th group centroid in the r -dimensional CVA biplot space is a measure of the central locality of the points representing the n_j samples belonging to the j th group in the r -dimensional CVA biplot space, $j \in [1 : J]$. If $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$ and the j th group predictivity is much higher than the k th group predictivity, it can therefore be expected that on average the angles between the vectors in the set, $\{(\mathbf{x}_i^j)' \mathbf{L}\}_{i=1}^{n_j}$, and the corresponding vectors in the set, $\{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r'\}_{i=1}^{n_j}$, will be smaller than the angles between the vectors in the set $\{(\mathbf{x}_l^k)' \mathbf{L}\}_{l=1}^{n_k}$ and the corresponding vectors in the set $\{(\mathbf{x}_l^k)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r'\}_{l=1}^{n_k}$. Since ϕ_i^j is a decreasing function of the size of the angle between $(\mathbf{x}_i^j)' \mathbf{L}$ and $(\mathbf{x}_i^j)' \mathbf{L} \mathbf{V}_r \mathbf{V}_r'$, it follows that if $\mathbf{C} = \mathbf{I}$ or $\mathbf{C} = \mathbf{N}$ and the j th group predictivity is much higher than the k th group predictivity, it can be expected that on average the sample predictivities of the samples of the j th group will be higher than those of the samples of the k th group. The fact that the accuracy of the approximations to the measurements of \mathbf{x}_i^j is dependent on the accuracy of the approximations to the measurements of $\bar{\mathbf{x}}^j$ is also evident from the expression of ϕ_i^j in (5.11.9). Furthermore, since the group predictivities of the larger groups tend to be higher for the weighted CVA biplot than for the unweighted CVA biplot constructed with $\mathbf{C} = \mathbf{I}$ of the same dimension, this implies that it can be expected that on average the sample predictivities of the samples belonging to the larger groups will

be higher for the weighted CVA biplot than for the CVA biplot constructed from $\mathbf{C} = \mathbf{I}$. Similarly it can be expected that the sample predictivities of the samples of the smaller groups will on average be lower for the weighted CVA biplot than for the CVA biplot constructed from $\mathbf{C} = \mathbf{I}$.

5.11.2 Sample predictivities and the accuracy of distances represented in the CVA biplot

Consider the Pythagorean distance between the points representing the i th sample of the j th group and the k th sample of the ℓ th group in the p -dimensional canonical space:

$$\left\{ \left((\mathbf{x}_i^j)' \mathbf{L} - (\mathbf{x}_k^\ell)' \mathbf{L} \right) \left((\mathbf{x}_i^j)' \mathbf{L} - (\mathbf{x}_k^\ell)' \mathbf{L} \right)' \right\}^{1/2} = \frac{1}{\sqrt{n}} \left\{ (\mathbf{x}_i^j - \mathbf{x}_k^\ell)' \widehat{\Sigma}_W^{-1} (\mathbf{x}_i^j - \mathbf{x}_k^\ell) \right\}^{1/2} \quad (5.11.10)$$

$i \in [1 : n_j], k \in [1 : n_\ell], j, \ell \in [1 : J]$. It follows that if $(\hat{\mathbf{x}}_i^j)' \mathbf{L}$ and $(\hat{\mathbf{x}}_k^\ell)' \mathbf{L}$ are accurate approximations to $(\mathbf{x}_i^j)' \mathbf{L}$ and $(\mathbf{x}_k^\ell)' \mathbf{L}$ respectively, then the Pythagorean distance between the two points $(\hat{\mathbf{x}}_i^j)' \mathbf{L}$ and $(\hat{\mathbf{x}}_k^\ell)' \mathbf{L}$,

$$\left\{ \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^\ell)' \mathbf{L} \right) \left((\hat{\mathbf{x}}_i^j)' \mathbf{L} - (\hat{\mathbf{x}}_k^\ell)' \mathbf{L} \right)' \right\}^{1/2}, \quad (5.11.11)$$

will most likely be an accurate approximation of (5.11.10). This means that if both ϕ_i^j and ϕ_k^ℓ are close to one, (5.11.11) will most likely be an accurate approximation of (5.11.10). It follows that when each sample in some set has a very high sample predictivity, then it can be expected that the relative magnitudes of the Pythagorean distances between the points representing those samples in the CVA biplot space will accurately represent the relative magnitudes of the Mahalanobis distances between those samples in the p -dimensional measurement space. That is, if the sample predictivity of each sample in some set of samples is high, then the true intersample relationships (as measured by the Mahalanobis distance metric) will most likely be accurately represented in the CVA biplot. When one or both of ϕ_i^j and ϕ_k^ℓ are low (5.11.11) will likely not be an accurate approximation to (5.11.10). It is however not guaranteed that the approximation will be poor - it is possible that the distance in (5.11.10) will be accurately represented in the CVA biplot even if one or both of the samples are poorly approximated. Hence, given a set of samples, some or all of which have low sample predictivities, the relative magnitudes of the Pythagorean distances between the points representing these samples in the CVA biplot space cannot be trusted to provide an accurate visualisation of the intersample relationships as measured by the Mahalanobis distance metric. Recall from

Chapter 4 that the r -dimensional weighted CVA biplot space is that r -dimensional subspace of the canonical space in which the Pythagorean distances between the samples in the p -dimensional canonical space are optimally represented, that is it is the r -dimensional space in which the true relationships between the samples (as measured by the Mahalanobis distance metric) are optimally represented.

A measure that assesses the accuracy with which the Pythagorean distances between the samples in the CVA biplot approximates the corresponding Pythagorean distances in the p -dimensional canonical space is introduced in Section 5.12.1.

5.11.3 Scale invariance

Substituting $\widehat{\mathbf{X}}\mathbf{A}^{-1}$ for $\widehat{\mathbf{X}}^*$, $\mathbf{X}\mathbf{A}^{-1}$ for \mathbf{X}^* and $\mathbf{A}\mathbf{W}^{-1}\mathbf{A}$ for $(\mathbf{W}^*)^{-1}$ in

$$\phi_i^{j*} = \frac{[\widehat{\mathbf{X}}^* (\mathbf{W}^*)^{-1} (\widehat{\mathbf{X}}^*)']_{ii}}{[\mathbf{X}^* (\mathbf{W}^*)^{-1} (\mathbf{X}^*)']_{ii}}$$

shows that the sample predictivity of the i th sample belonging to the j th group associated with the CVA biplot constructed from the unstandardised measurements is identical to that associated with the CVA biplot constructed from the unstandardised measurements:

$$\begin{aligned} \phi_i^{j*} &= \frac{[\widehat{\mathbf{X}}\mathbf{A}^{-1}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}\mathbf{A}^{-1}\widehat{\mathbf{X}}']_{ii}}{[\mathbf{X}\mathbf{A}^{-1}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{X}']_{ii}} \\ \longrightarrow \phi_i^{j*} &= \phi_i^j. \end{aligned} \quad (5.11.12)$$

Substituting any $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in equation (5.11.12) shows that the sample predictivity measure is invariant to all non-singular linear transformations of the form $\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x}$.

5.11.4 The overall sample predictivity associated with a group

5.11.4.1 Definition and properties

The fact that the decomposition of $\mathbf{X}\mathbf{L}$ into $\widehat{\mathbf{X}}\mathbf{L}$ and $\mathbf{X}\mathbf{L} - \widehat{\mathbf{X}}\mathbf{L}$ exhibits Type A orthogonality implies that

$$\text{tr}\{\mathbf{X}\mathbf{L}\mathbf{L}'\mathbf{X}'\} = \text{tr}\{\widehat{\mathbf{X}}\mathbf{L}\mathbf{L}'\widehat{\mathbf{X}}'\} + \text{tr}\{(\mathbf{X}\mathbf{L} - \widehat{\mathbf{X}}\mathbf{L})(\mathbf{X}\mathbf{L} - \widehat{\mathbf{X}}\mathbf{L})'\}. \quad (5.11.13)$$

The validity of (5.11.13) implies that the ratio

$$\phi^j = \frac{\sum_{i=1}^{n_j} (\hat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \hat{\mathbf{x}}_i^j}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \quad (5.11.14)$$

$$= \frac{\sum_{i=1}^{n_j} (\hat{\mathbf{x}}_i^j)' \mathbf{W}^{-1} \hat{\mathbf{x}}_i^j}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{W}^{-1} \mathbf{x}_i^j} \quad (5.11.15)$$

$$\phi^j = 1 - \frac{\sum_{i=1}^{n_j} (\mathbf{x}_i^j - \hat{\mathbf{x}}_i^j)' \mathbf{W}^{-1} (\mathbf{x}_i^j - \hat{\mathbf{x}}_i^j)}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{W}^{-1} \mathbf{x}_i^j}.$$

can be used to assess the overall accuracy of the approximations to the n_j samples belonging to the j th group, $j \in [1 : J]$. The ratio in (5.11.15) will henceforth be referred to as the overall sample predictivity of the j th group, or simply the j th overall sample predictivity. The j th overall sample predictivity is equal to the proportion of the total sample variance within the j th group of canonical observations, with variance being interpreted in terms of deviation from the overall mean of the n canonical observations (which is $\mathbf{0}$), that is accounted for in the CVA biplot.

Being a ratio of sums of squared values, the j th overall sample predictivity can only take on non-negative values. It is evident that ϕ^j has a minimum value of zero and that

$$\begin{aligned} \phi^j = 0 &\iff (\mathbf{x}_i^j)' \mathbf{L} = \mathbf{0}' \quad \forall i \in [1 : n_j] \\ \therefore \phi^j = 0 &\iff (\mathbf{x}_i^j)' \mathbf{L} \in \mathcal{L}^\perp \quad \forall i \in [1 : n_j] \end{aligned}$$

while it has a maximum value of one and that

$$\begin{aligned} \phi^j = 1 &\iff (\hat{\mathbf{x}}_i^j)' \mathbf{L} = (\mathbf{x}_i^j)' \mathbf{L} \quad \forall i \in [1 : n_j] \\ \therefore \phi^j = 1 &\iff (\mathbf{x}_i^j)' \mathbf{L} \in \mathcal{L} \quad \forall i \in [1 : n_j]. \end{aligned}$$

Since the matrix $\mathbf{X}\mathbf{L}$ is of rank p , this implies that the J groups will all have unit overall sample predictivities if and only if $r = p$. Due to the fact that

$$\mathbf{V}_K^{\mathbf{I}} (\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\text{Cent}} (\mathbf{V}_K^{\text{Cent}})' = \mathbf{V}_K^{\mathbf{N}} (\mathbf{V}_K^{\mathbf{N}})'$$

the overall sample predictivity of the j th group corresponding to the K -dimensional weighted CVA biplot and the two K -dimensional unweighted CVA biplots are iden-

tical.

The overall sample predictivity of the j th group can be expressed as a weighted average of the individual sample predictivities of the n_j samples belonging to the j th group:

$$\begin{aligned}
 \phi_i^j &= \frac{(\hat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \hat{\mathbf{x}}_i^j}{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \\
 \longrightarrow (\hat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \hat{\mathbf{x}}_i^j &= \phi_i^j (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j \\
 \longrightarrow \phi^j &= \frac{\sum_{i=1}^{n_j} \phi_i^j (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \\
 \longrightarrow \phi^j &= \sum_{i=1}^{n_j} \phi_i^j \frac{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j}{\sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j}. \tag{5.11.16}
 \end{aligned}$$

Since ϕ_i^j corresponding to the K -dimensional weighted CVA biplot and the two K -dimensional unweighted CVA biplots are identical for all $i \in [1 : n_j]$, equation (5.11.16) also indicates that the values of ϕ^j corresponding to these three K -dimensional CVA biplots are identical, $j \in [1 : J]$.

Since it can be expected that on average the sample predictivities of the samples belonging to the j th group will be higher than those of the samples belonging to the k th group if the j th group predictivity is much higher than the k th group predictivity, ϕ^j can be expected to be higher than ϕ^k if the j th group predictivity is much higher than the k th group predictivity. Furthermore, since it is expected that on average the samples belonging to the larger groups will have higher sample predictivities for the r -dimensional weighted CVA biplot than for the r -dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot, the overall sample predictivities of the larger groups corresponding to the r -dimensional weighted CVA biplot are expected to be higher than those corresponding to the r -dimensional unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplot.

Some of the concepts regarding group predictivities, overall within-group sample predictivities and overall sample predictivities that have been discussed will now be illustrated at the hand of the fourth and fifth simulated data sets that were introduced in Section 4.8. Table 5.11 contains for each of the four groups of the fourth simulated data set, its overall sample predictivity corresponding to the one-dimensional CVA biplots constructed with $\mathbf{C} = \mathbf{N}$, $\mathbf{C} = \mathbf{I}$ and $\mathbf{C} = (\mathbf{I} - \frac{1}{j} \mathbf{1} \mathbf{1}')$ respectively while Table 5.14 provides the same information for the fifth simulated data set. Table 5.12 provides the group predictivity of each of the four groups of the fourth simulated data set corresponding to the one-dimensional CVA biplots corresponding to $\mathbf{C} = \mathbf{N}$, $\mathbf{C} = \mathbf{I}$ and $\mathbf{C} = (\mathbf{I} - \frac{1}{j} \mathbf{1} \mathbf{1}')$. Table 5.15 provides the same information for the fifth simulated data set. Tables 5.13 and 5.16 provide the within-group predictivities of the four groups associated with the three different one-dimensional CVA biplots for the fourth and fifth simulated data sets respectively.

Table 5.11: *The overall sample predictivities corresponding to the one-dimensional CVA biplots of the fourth simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	150	150	50	50
$\mathbf{C} = \mathbf{N}$	0.246	0.377	0.279	0.176
$\mathbf{C} = \mathbf{I}$	0.150	0.280	0.496	0.259
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.155	0.296	0.506	0.222

Table 5.12: *The group predictivities corresponding to the one-dimensional CVA biplots of the fourth simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	150	150	50	50
$\mathbf{C} = \mathbf{N}$	0.404	0.982	0.432	0.048
$\mathbf{C} = \mathbf{I}$	0.000	0.506	0.895	0.286
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.057	0.792	0.845	0.035

Table 5.13: *The overall within-group sample predictivities of the one-dimensional CVA biplots of the fourth simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	150	150	50	50
$\mathbf{C} = \mathbf{N}$	0.210	0.192	0.155	0.239
$\mathbf{C} = \mathbf{I}$	0.184	0.211	0.172	0.246
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.188	0.206	0.178	0.244

When considering the overall sample predictivities in Tables 5.11 and 5.14 together with the corresponding group predictivities in Tables 5.12 and 5.15 respectively, it is evident that the groups associated with higher group predictivities are also usually associated with higher overall sample predictivities, as expected. On the other hand, when considering the group predictivities together with the overall within-group sample predictivities it seems that there is no association between the accuracy with which the CVA biplot predicts the deviations of the samples belonging to a particular group from that group's centroid and the accuracy with which the CVA biplot predicts the group centroid itself. When comparing the overall sample predictivities corresponding to the three types of CVA biplots for each of the groups, it is evident that the overall sample predictivities of the larger groups tend to be higher for the weighted CVA biplot than for the two unweighted CVA biplots while the overall sample predictivities of the smaller groups tend to be lower for the weighted CVA biplot than for the two unweighted CVA biplots. Upon comparison of the overall within-group sample predictivities associated with the weighted and unweighted CVA biplots of the same dimension, it seems that unlike the group

predictivity and sample predictivity measures, the overall within-group sample predictivity measure is not very sensitive to which \mathbf{C} -matrix is used in the construction of the CVA biplot.

Table 5.14: *The overall sample predictivities corresponding to the one-dimensional CVA biplots of the fifth simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	50	150	50	150
$\mathbf{C} = \mathbf{N}$	0.200	0.391	0.349	0.214
$\mathbf{C} = \mathbf{I}$	0.297	0.298	0.468	0.164
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.268	0.321	0.469	0.167

Table 5.15: *The group predictivities corresponding to the one-dimensional CVA biplots of the fifth simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	50	150	50	150
$\mathbf{C} = \mathbf{N}$	0.145	0.984	0.541	0.358
$\mathbf{C} = \mathbf{I}$	0.451	0.616	0.858	0.007
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.165	0.882	0.774	0.041

Table 5.16: *The overall within-group sample predictivities of the one-dimensional CVA biplots of the fifth simulated data set.*

	Group 1	Group 2	Group 3	Group 4
Group Size	50	150	50	150
$\mathbf{C} = \mathbf{N}$	0.230	0.196	0.226	0.186
$\mathbf{C} = \mathbf{I}$	0.215	0.194	0.220	0.195
$\mathbf{C} = (\mathbf{I} - \frac{1}{j}\mathbf{1}\mathbf{1}')$	0.222	0.195	0.218	0.193

5.11.4.2 Scale Invariance

Substituting $\mathbf{X}\mathbf{A}^{-1}$ for \mathbf{X}^* , $\widehat{\mathbf{X}}\mathbf{A}^{-1}$ for $\widehat{\mathbf{X}}^*$ and $\mathbf{A}\mathbf{L}$ for \mathbf{L}^* in the expression of the j th overall sample predictivity corresponding to the CVA biplot constructed from the standardised measurements,

$$\phi^{j*} = \frac{\sum_{i=1}^{n_j} (\widehat{\mathbf{x}}_i^{j*})' \mathbf{L}^* \mathbf{L}^{*'} \widehat{\mathbf{x}}_i^{j*}}{\sum_{i=1}^{n_j} (\mathbf{x}_i^{j*})' \mathbf{L}^* \mathbf{L}^{*'} \mathbf{x}_i^{j*}}$$

shows that $\phi^{j*} = \phi^j$. Like the sample predictivity measure, the overall sample predictivity measure is invariant to all non-singular linear transformations of the scales of the measurements of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$. This is evident upon substituting an arbitrary $p \times p$ non-singular matrix \mathbf{F} for \mathbf{A}^{-1} in the expressions above.

5.11.5 The total sample predictivity associated with a data set

5.11.5.1 Definition and properties

The overall accuracy of the approximations to the measurements of all the samples, over all the groups can be measured by the following ratio which will henceforth be referred to as the total sample predictivity:

$$\begin{aligned} \phi &= \frac{\text{tr} \{ \widehat{\mathbf{X}} \mathbf{L} \mathbf{L}' \widehat{\mathbf{X}}' \}}{\text{tr} \{ \mathbf{X} \mathbf{L} \mathbf{L}' \mathbf{X}' \}} \\ &= \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (\widehat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \widehat{\mathbf{x}}_i^j}{\sum_{j=1}^J \sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \\ &= \frac{\text{tr} \{ \widehat{\mathbf{X}} \mathbf{L} \mathbf{L}' \widehat{\mathbf{X}}' \}}{\text{tr} \{ \mathbf{X} \mathbf{L} \mathbf{L}' \mathbf{X}' \}} \\ &= \frac{\text{tr} \{ \mathbf{X} \mathbf{L} \mathbf{V}_r \mathbf{V}_r' \mathbf{L}' \mathbf{X}' \}}{\text{tr} \{ \mathbf{X} \mathbf{L} \mathbf{L}' \mathbf{X}' \}} \\ &= \frac{\text{tr} \{ \mathbf{X} \mathbf{L} \mathbf{V}_r \mathbf{V}_r' \mathbf{L}' \mathbf{X}' \}}{\text{tr} \{ \mathbf{X} \mathbf{L} \mathbf{V} \mathbf{V}' \mathbf{L}' \mathbf{X}' \}} \\ \rightarrow \phi &= \frac{\sum_{i=1}^n \sum_{j=1}^r ([\mathbf{X} \mathbf{L} \mathbf{V}_r]_{ij})^2}{\sum_{i=1}^n \sum_{j=1}^p ([\mathbf{X} \mathbf{L} \mathbf{V}]_{ij})^2}. \end{aligned}$$

The total sample predictivity is equal to the proportion of the total sample variance associated with the n canonical observations, with variance being interpreted in terms of deviation from the overall mean of the n canonical observations, that is accounted for in the CVA biplot.

Being defined as the ratio of sums of squared values, the total sample predictivity measure can only take on non-negative values. It is evident that ϕ is a non-decreasing function of the dimension of the biplot space, r , and hence that a sufficient condition for ϕ to equal its maximum value, namely one, is that $r = p$. The total sample predictivity will attain its minimum value of zero if and only if each of the n samples lie in the orthogonal complement of $\mathcal{V}(\mathbf{V}_r)$.

The total sample predictivity associated with a data set can be expressed as the weighted average of the sample predictivities of all the individual samples:

$$(\widehat{\mathbf{x}}_i^j)' \mathbf{L} \mathbf{L}' \widehat{\mathbf{x}}_i^j = \phi_i^j (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j$$

$$\begin{aligned}
 \rightarrow \phi &= \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} \phi_i^j (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j}{\sum_{j=1}^J \sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j} \\
 \rightarrow \phi &= \sum_{j=1}^J \sum_{i=1}^{n_j} \phi_i^j \frac{(\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j}{\sum_{j=1}^J \sum_{i=1}^{n_j} (\mathbf{x}_i^j)' \mathbf{L} \mathbf{L}' \mathbf{x}_i^j}. \quad (5.11.17)
 \end{aligned}$$

The total sample predictivities corresponding to the weighted and two unweighted K -dimensional CVA biplots, will be identical due to the fact that

$$\mathbf{V}_K^{\mathbf{I}} (\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\text{Cent}} (\mathbf{V}_K^{\text{Cent}})' = \mathbf{V}_K^{\mathbf{N}} (\mathbf{V}_K^{\mathbf{N}})'.$$

This is also evident from the expression of ϕ in equation (5.11.17) and the fact that ϕ_i^j is identical for the three K -dimensional CVA biplots for all $i \in [1 : n_j]$ and $j \in [1 : J]$.

5.11.5.2 Scale Invariance

Upon substituting $\mathbf{X}\mathbf{F}$ for \mathbf{X}^* , $\widehat{\mathbf{X}}\mathbf{F}$ for $\widehat{\mathbf{X}}^*$ and $\mathbf{F}^{-1}\mathbf{L}$ for \mathbf{L}^* in the expression of the total sample predictivity corresponding to the CVA biplot constructed from the transformed measurements, $\mathbf{X}\mathbf{F}$,

$$\phi^* = \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (\widehat{\mathbf{x}}_i^{j*})' \mathbf{L}^* \mathbf{L}^{*'} \widehat{\mathbf{x}}_i^{j*}}{\sum_{j=1}^J \sum_{i=1}^{n_j} (\mathbf{x}_i^{j*})' \mathbf{L}^* \mathbf{L}^{*'} \mathbf{x}_i^{j*}}$$

it is evident that the total sample predictivity is invariant to all non-singular linear transformations of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$. This holds in particular for $\mathbf{F} = \mathbf{A}^{-1}$ which implies that the total sample predictivity associated with the CVA biplot constructed from the standardised measurements, $\mathbf{X}\mathbf{A}^{-1}$, is identical to that associated with the CVA biplot constructed from the unstandardised measurements.

5.11.6 Sample predictivities measures of ‘new’ samples

The sample predictivity, overall sample predictivity and total sample predictivity measures associated with ‘new’ samples can be defined in exactly the same way as for the original samples. The approximation to \mathbf{X}^{new} (defined in 5.8.4) which is produced by the r -dimensional CVA biplot constructed from $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ is given by

$$\widehat{\mathbf{X}}^{\text{new}} = \mathbf{X}^{\text{new}} \mathbf{M}_r \mathbf{M}_r' = \mathbf{X}^{\text{new}} \mathbf{L} \mathbf{V}_r \mathbf{V}_r' \mathbf{L}^{-1}. \quad (5.11.18)$$

Since the matrix \mathbf{V}_r in (5.11.18) is an orthonormal matrix, the decomposition of $\mathbf{X}^{\text{new}}\mathbf{L}$,

$$\mathbf{X}^{\text{new}}\mathbf{L} = \widehat{\mathbf{X}}^{\text{new}}\mathbf{L} + (\mathbf{X}^{\text{new}}\mathbf{L} - \widehat{\mathbf{X}}^{\text{new}}\mathbf{L})$$

exhibits Type A orthogonality. This validates the ratios

$$\phi_i^{\text{new}} = \frac{[\widehat{\mathbf{X}}^{\text{new}}\mathbf{L}\mathbf{L}'(\widehat{\mathbf{X}}^{\text{new}})']_{ii}}{[\mathbf{X}^{\text{new}}\mathbf{L}\mathbf{L}'(\mathbf{X}^{\text{new}})']_{ii}} \quad (5.11.19)$$

$$\phi_j^{\text{new}} = \frac{\sum_{i:[G]_{ij}=1} [\widehat{\mathbf{X}}^{\text{new}}\mathbf{L}\mathbf{L}'(\widehat{\mathbf{X}}^{\text{new}})']_{ii}}{\sum_{i:[G]_{ij}=1} [\mathbf{X}^{\text{new}}\mathbf{L}\mathbf{L}'(\mathbf{X}^{\text{new}})']_{ii}} \quad (5.11.20)$$

$$\phi^{\text{new}} = \frac{\text{tr} \{ \widehat{\mathbf{X}}^{\text{new}}\mathbf{L}\mathbf{L}'(\widehat{\mathbf{X}}^{\text{new}})' \}}{\text{tr} \{ \mathbf{X}^{\text{new}}\mathbf{L}\mathbf{L}'(\mathbf{X}^{\text{new}})' \}} \quad (5.11.21)$$

as quality measures. The ratios in (5.11.19), (5.11.20) and (5.11.21) define the sample predictivity of the i th new sample, $\mathbf{x}_i^{\text{new}}$, the overall sample predictivity of the j th group corresponding to the new samples and the total sample predictivity associated with the m new samples respectively.

Recall from Chapter 4 that if the relative group sizes associated with a set of new samples differ substantially from those of the set of original samples, then the approximations of the new samples read off from the predictive biplot axes of one of the r -dimensional unweighted CVA biplots constructed from the set of original samples will most likely be more accurate than the approximations read off from the biplot axes of the r -dimensional weighted CVA biplot constructed from the original samples. This is demonstrated below using the third and fourth simulated data sets introduced in Section 4.8. The samples of the third simulated data set are predicted using the weighted and unweighted (with $\mathbf{C} = \mathbf{I}$) CVA biplots constructed from the fourth simulated data set, respectively. In each case the overall accuracy of the approximations to the samples of the third data set is measured by the total sample predictivity measure. Recall that the third simulated data set contains 50 samples belonging to each of groups one and two and 150 samples belonging to each of groups three and four. On the other hand, the fourth simulated data set contains 150 samples belonging to each of groups one and two and 50 samples belonging to each of groups three and four. Since the relative group sizes of these two data sets differ substantially, it can be expected that the measurements of the samples of the third simulated data set will be predicted more accurately in the unweighted CVA biplot constructed from the fourth simulated data set than in the corresponding weighted CVA biplot.

Table 5.17 contains the total sample predictivity of the third simulated data set corresponding to the weighted and unweighted CVA biplots constructed from the

fourth simulated data set for each of the possible dimensionalities of these biplots. Upon comparison of the total sample predictivities associated with the weighted and unweighted CVA biplots in Table 5.17, the expectation that the samples of the third data set will be more accurately predicted in the unweighted CVA biplot is confirmed.

Table 5.17: *The total sample predictivity of the third simulated data set associated with the unweighted and weighted CVA biplots constructed from the fourth simulated data set.*

	Dim: 1	Dim: 2	Dim: 3	Dim: 4	Dim: 5
Unweighted	0.295	0.547	0.729	0.862	1.000
Weighted	0.216	0.510	0.729	0.863	1.000

Recall that when the CVA biplot is p -dimensional, then irrespective of which \mathbf{C} -matrix is used in the construction of the biplot, all samples will be perfectly represented. This explains why, the closer the dimension of the biplot gets to p , the less the difference between the total sample predictivities of the third simulated data set associated with the unweighted and weighted CVA biplots constructed from the fourth simulated data set become.

5.12 Sample contrast predictivities

5.12.1 Definition and Properties

A measure that assesses the accuracy with which the Pythagorean distances between the individual samples in the CVA biplot approximates the corresponding Pythagorean distances in the p -dimensional canonical space can be defined in a very similar way to the group contrast predictivity measure in Section 5.5 and the mixed contrast predictivity in Section 5.10. Consider the following orthogonal decomposition of the squared Pythagorean distance between the points representing the i th and j th samples in the p -dimensional canonical space:

$$\|\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \mathbf{X}\mathbf{L}\|^2 = \|\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}\|^2 + \|(\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \mathbf{X}\mathbf{L} - (\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}))\|^2 \quad (5.12.1)$$

where

$$\widehat{\mathbf{X}}\mathbf{L} = \mathbf{X}\mathbf{L}\mathbf{V}_r\mathbf{V}'_r$$

and \mathbf{V} is the matrix of right singular vectors of the matrix $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$. The decomposition in (5.12.1) is valid due to the fact that the cross product term

$$(\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})(\mathbf{e}'_i\mathbf{X}\mathbf{L} - \mathbf{e}'_j\mathbf{X}\mathbf{L} - (\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L}))'$$

is equal to zero:

$$\begin{aligned} & (\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})(\mathbf{e}'_i\mathbf{X}\mathbf{L} - \mathbf{e}'_j\mathbf{X}\mathbf{L} - (\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L}))' \\ &= (\mathbf{e}'_i\mathbf{X} - \mathbf{e}'_j\mathbf{X})\mathbf{L}\mathbf{V}_r\mathbf{V}'_r\mathbf{L}'(\mathbf{e}'_i\mathbf{X} - \mathbf{e}'_j\mathbf{X})' - (\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})(\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})' \\ &= (\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})(\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})' - (\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})(\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L})' \\ &= 0. \end{aligned}$$

It follows that the the accuracy with which the Pythagorean distance between the i th and j th samples in the p -dimensional canonical space is approximated by the corresponding Pythagorean distance in the CVA biplot can (for all three choices of \mathbf{C}) be measured by the ratio,

$$\phi_{ij} = \frac{\|\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L}\|^2}{\|\mathbf{e}'_i\mathbf{X}\mathbf{L} - \mathbf{e}'_j\mathbf{X}\mathbf{L}\|^2}. \quad (5.12.2)$$

The ratio in (5.12.2) will henceforth be referred to as the (ij) th sample contrast predictivity, $i, j \in [1 : n]$, $i \neq j$.

Note that since the Pythagorean distance between two points representing two samples in the p -dimensional canonical space is proportional to the Mahalanobis distance between those two samples in the p -dimensional measurement space, ϕ_{ij} is also equal to the squared approximated Mahalanobis distance between the i th and j th samples, expressed as a proportion of the squared true Mahalanobis distance between those two samples.

Like the group contrast predictivity measure, the sample contrast predictivity measure can only take on non-negative values. The (ij) th sample contrast predictivity has a minimum value of zero which it will attain if and only if

$$\mathbf{e}'_i\widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j\widehat{\mathbf{X}}\mathbf{L} = \mathbf{0}'.$$

Equation (5.12.1) implies that ϕ_{ij} can also be expressed as

$$\phi_{ij} = 1 - \frac{\|\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \mathbf{X}\mathbf{L} - (\mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L})\|^2}{\|\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \mathbf{X}\mathbf{L}\|^2}. \quad (5.12.3)$$

It follows that ϕ_{ij} has a maximum value of one which it will attain if and only if

$$\mathbf{e}'_i \mathbf{X}\mathbf{L} - \mathbf{e}'_j \mathbf{X}\mathbf{L} = \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} - \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L}.$$

It is evident that the conditions,

$$\begin{aligned} \mathbf{e}'_i \mathbf{X}\mathbf{L} &= \mathbf{e}'_i \widehat{\mathbf{X}}\mathbf{L} \\ \text{and } \mathbf{e}'_j \mathbf{X}\mathbf{L} &= \mathbf{e}'_j \widehat{\mathbf{X}}\mathbf{L} \end{aligned}$$

are sufficient but not necessary for ϕ_{ij} to attain the value of one. This implies that accurate approximations of both the samples $\mathbf{e}'_i \mathbf{X}$ and $\mathbf{e}'_j \mathbf{X}$ in the CVA biplot suggest that it can be expected that the Pythagorean distances between those two samples in the p -dimensional canonical space is accurately approximated in the CVA biplot. That is, if the sample predictivities of the samples $\mathbf{e}'_i \mathbf{X}$ and $\mathbf{e}'_j \mathbf{X}$ are both close to one, then it can be expected that ϕ_{ij} will be close to one. However, it should be noted that the Pythagorean distance between the points representing the samples $\mathbf{e}'_i \mathbf{X}$ and $\mathbf{e}'_j \mathbf{X}$ in the p -dimensional canonical space can be accurately approximated in the CVA biplot even if one or both of the samples are poorly approximated.

Note that given the values of the group contrast predictivities and mixed contrast predictivities, sample contrast predictivities do not add much (if any) information about the quality of the representation of the group structure in the CVA biplot.

do not add much (if any) information about the quality of the representation of the group structure in the CVA biplot.

5.12.2 Scale invariance

Like the group contrast predictivity measure and the mixed contrast predictivity the sample contrast predictivity measure is invariant to all non-singular linear transformations of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$. This implies that the (ij) th sample contrast predictivity associated with the CVA biplot constructed from the standardised measurements, $\mathbf{X}\mathbf{A}^{-1}$, is identical to that associated with the CVA biplot constructed from the standardised measurements.

5.13 Within-group axis predictivities

5.13.1 Definition and properties

In order to measure the predictive ability of a predictive CVA biplot axis within the J groups, the overall accuracy of the approximations to the deviations of the individual samples' measurements on that variable from the measurements of the corresponding group centroids on that variable needs to be considered. Gardner-Lubbe *et al.* (2008) proposed that the ratio,

$$\pi_k^W = \frac{[\widehat{\mathbf{K}}'\widehat{\mathbf{K}}]_{kk}}{[\mathbf{K}'\mathbf{K}]_{kk}} \quad (5.13.1)$$

which they referred to as the within-group axis predictivity of the k th predictive CVA biplot axis, be used to measure the predictive ability of the k th predictive biplot axis within the groups. Henceforth π_k^W will also be referred to as the k th within-group axis predictivity.

It is shown below that the inner-product matrix $\mathbf{K}'\mathbf{K}$ is equal to the matrix of within-group sums of squares and cross products, \mathbf{W} , while the inner-product matrix $\widehat{\mathbf{K}}'\widehat{\mathbf{K}}$ is equal to $(\mathbf{M}^r)'\mathbf{M}^r$:

$$\begin{aligned} \mathbf{K}'\mathbf{K} &= \mathbf{X}'\mathbf{H}\mathbf{H}\mathbf{X} \\ &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{X} \\ &= \mathbf{X}'\mathbf{X} - \overline{\mathbf{X}}'\mathbf{G}'\mathbf{G}\overline{\mathbf{X}} \\ &= \mathbf{X}'\mathbf{X} - \overline{\mathbf{X}}'\mathbf{N}\overline{\mathbf{X}} \\ \longrightarrow \mathbf{K}'\mathbf{K} &= \mathbf{W} \\ \widehat{\mathbf{K}}'\widehat{\mathbf{K}} &= (\mathbf{M}^r)'\mathbf{M}_r'\mathbf{K}'\mathbf{K}\mathbf{M}_r\mathbf{M}^r \\ &= (\mathbf{M}^r)'\mathbf{M}_r'\mathbf{W}\mathbf{M}_r\mathbf{M}^r \\ &= (\mathbf{M}^r)'\mathbf{M}_r'(\mathbf{M}^r)'\mathbf{M}^r \\ &= (\mathbf{M}^r)'\mathbf{I}_r\mathbf{M}^r \\ \longrightarrow \widehat{\mathbf{K}}'\widehat{\mathbf{K}} &= (\mathbf{M}^r)'\mathbf{M}^r. \end{aligned}$$

The k th within-group axis predictivity can therefore be expressed as:

$$\pi_k^W = \frac{[(\mathbf{M}^r)'\mathbf{M}^r]_{kk}}{[(\mathbf{M}^{-1})'\mathbf{M}^{-1}]_{kk}}$$

$$\therefore \pi_k^W = \frac{\sum_{i=1}^r ([\mathbf{M}^{-1}]_{ik})^2}{\sum_{i=1}^p ([\mathbf{M}^{-1}]_{ik})^2}. \quad (5.13.2)$$

The p -component vector of within-group axis predictivities with k th element equal to π_k^W follows as:

$$\begin{aligned} \boldsymbol{\pi}^W &= \text{diag}(\widehat{\mathbf{K}}'\widehat{\mathbf{K}}) [\text{diag}(\mathbf{K}'\mathbf{K})]^{-1} \\ &= \text{diag}((\mathbf{M}^r)'\mathbf{M}^r) [\text{diag}(\mathbf{W})]^{-1} \\ \longrightarrow \boldsymbol{\pi}^W &= \text{diag}((\mathbf{M}^r)'\mathbf{M}^r) \left[\text{diag}((\mathbf{M}^{-1})'\mathbf{M}^{-1}) \right]^{-1}. \end{aligned}$$

It is shown below that the cross products matrix $\widehat{\mathbf{K}}'(\mathbf{K} - \widehat{\mathbf{K}})$ is equal to the null matrix and hence that the decomposition,

$$\mathbf{K} = \widehat{\mathbf{K}} + (\mathbf{K} - \widehat{\mathbf{K}}) \quad (5.13.3)$$

exhibits Type B orthogonality:

$$\begin{aligned} \widehat{\mathbf{K}}'(\mathbf{K} - \widehat{\mathbf{K}}) &= \widehat{\mathbf{K}}'\mathbf{K} - \widehat{\mathbf{K}}'\widehat{\mathbf{K}} \\ &= (\mathbf{M}^r)'\mathbf{M}'_r\mathbf{K}'\mathbf{K} - (\mathbf{M}^r)'\mathbf{M}'_r\mathbf{K}'\mathbf{K}\mathbf{M}_r\mathbf{M}^r \\ &= (\mathbf{M}^r)'\mathbf{M}'_r\mathbf{W} - (\mathbf{M}^r)'\mathbf{M}'_r\mathbf{W}\mathbf{M}_r\mathbf{M}^r \\ &= (\mathbf{M}^r)'\mathbf{M}'_r\mathbf{W} - (\mathbf{M}^r)'\mathbf{M}^r \text{ since } \mathbf{M}'_r\mathbf{W}\mathbf{M}_r = \mathbf{I}_r \\ &= (\mathbf{M}^r)'\mathbf{M}'_r\mathbf{W}\mathbf{M}^{-1}\mathbf{M} - (\mathbf{M}^r)'\mathbf{M}^r \\ &= (\mathbf{M}^r)'\mathbf{I}_r - (\mathbf{M}^r)'\mathbf{M}^r \\ &= (\mathbf{M}^r)'\mathbf{M}^r - (\mathbf{M}^r)'\mathbf{M}^r \\ \longrightarrow \widehat{\mathbf{K}}'(\mathbf{K} - \widehat{\mathbf{K}}) &= \mathbf{0}. \end{aligned}$$

The Type B orthogonality property of the decomposition in (5.13.3) validates π_k^W as a quality measure. For details regarding the definition of within-group axis predictivities for biplot axes that have been interpolated onto an existing CVA biplot, see Gower *et al.* (2011).

Being defined as a ratio of sums of squares, the within-group axis predictivity measure can only take on non-negative values. It is evident from the expression of π_k^W in (5.13.2) that π_k^W is a non-decreasing function of the dimension of the CVA biplot space, r , and that it has a maximum value of one which will be necessarily attained when $r = p$. The derivation below shows that the k th within-group axis predictivity is equal to the ratio of the squared length of the projection of the vector

$\mathbf{e}'_k (\mathbf{L}^{-1})'$ onto the r -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_r)$, to the squared length of the vector $\mathbf{e}'_k (\mathbf{L}^{-1})'$:

$$\begin{aligned} \pi_k^W &= \frac{[\widehat{\mathbf{K}}'\widehat{\mathbf{K}}]_{kk}}{[\mathbf{K}'\mathbf{K}]_{kk}} \\ &= \frac{[(\mathbf{M}^r)'\mathbf{M}^r]_{kk}}{[(\mathbf{M}^{-1})'\mathbf{M}^{-1}]_{kk}} \\ &= \frac{\|\mathbf{e}_k (\mathbf{L}^{-1})' \mathbf{V}_r\|^2}{\|\mathbf{e}'_k (\mathbf{L}^{-1})' \mathbf{V}\|^2} \\ \longrightarrow \pi_k^W &= \frac{\|\mathbf{e}_k (\mathbf{L}^{-1})' \mathbf{V}_r \mathbf{V}'_r\|^2}{\|\mathbf{e}'_k (\mathbf{L}^{-1})'\|^2}. \end{aligned} \quad (5.13.4)$$

It is evident that the k th within-group axis predictivity is a decreasing function of the orthogonal distance between the point $\mathbf{e}'_k (\mathbf{L}^{-1})'$ and its orthogonal projection onto the r -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_r)$. Hence, the k th within-group axis predictivity has a maximum value of one which will be attained if and only if the point $\mathbf{e}'_k (\mathbf{L}^{-1})'$ lies in the r -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_r)$, and a minimum value of zero which will be attained if and only if the point $\mathbf{e}'_k (\mathbf{L}^{-1})'$ lies in the orthogonal complement of the CVA biplot space. All p the biplot axes will therefore have unit within-group axis predictivities if and only if

$$\mathbf{e}'_k (\mathbf{L}^{-1})' \in \mathcal{V}(\mathbf{V}_r) \quad \forall k \in [1 : p]$$

that is if and only if $r = p$. All p biplot axes will never have zero within-group axis predictivities since this is possible if and only if

$$\mathbf{e}'_k (\mathbf{L}^{-1})' \in \mathcal{V}^\perp(\mathbf{V}_r)$$

that is if and only if $r = 0$. Note that since the matrices $\mathbf{V}_K \mathbf{V}'_K$ and $\mathbf{V} \mathbf{V} = \mathbf{I}_p$ are identical for the matrix \mathbf{V} obtained from the svd of $\overline{\mathbf{X}} \mathbf{L}$, the svd of $\mathbf{N}^{1/2} \overline{\mathbf{X}} \mathbf{L}$ and the svd of $(\mathbf{I} - \frac{1}{j} \mathbf{1} \mathbf{1}') \overline{\mathbf{X}} \mathbf{L}$, each of the p within-group axis predictivities will be identical for the weighted and the two unweighted K -dimensional CVA biplots. Since π_k^W is a non-decreasing function of the dimension of the CVA biplot space, r , equation (5.13.4) implies that the orthogonal distance between the point $\mathbf{e}'_k (\mathbf{L}^{-1})'$ and its orthogonal projection onto the CVA biplot space is a non-decreasing function of the dimension of the CVA biplot space. Note however that since the lengths of the different row vectors of $(\mathbf{L}^{-1})'$ are not necessarily the same, $\pi_k^W > \pi_j^W$ does not

necessarily imply that the point $\mathbf{e}'_k (\mathbf{L}^{-1})'$ is closer to the CVA biplot space than the point $\mathbf{e}'_j (\mathbf{L}^{-1})'$.

Note that π_k^W associated with the K -dimensional weighted CVA biplot and that associated with each of two the K -dimensional unweighted CVA biplots will be identical due to the fact that

$$\mathbf{V}_K^{\mathbf{I}} (\mathbf{V}_K^{\mathbf{I}})' = \mathbf{V}_K^{\text{Cent}} (\mathbf{V}_K^{\text{Cent}})' = \mathbf{V}_K^{\mathbf{N}} (\mathbf{V}_K^{\mathbf{N}})' .$$

Recall that the square of the cosine of the angle between the j th and k th predictive biplot axes of the p -dimensional CVA biplot is equal to the correlation coefficient between the j th and k th measured variables, that is

$$\begin{aligned} \cos(\theta_{jk}) &= \frac{[\widehat{\boldsymbol{\Sigma}}_W]_{jk}}{\sqrt{[\widehat{\boldsymbol{\Sigma}}_W]_{jj}} \sqrt{[\widehat{\boldsymbol{\Sigma}}_W]_{kk}}} \\ \rightarrow \cos(\theta_{jk}) &= \frac{\mathbf{e}'_j (\mathbf{M}^{-1})' \mathbf{M}^{-1} \mathbf{e}_k}{\sqrt{\mathbf{e}'_j (\mathbf{M}^{-1})' \mathbf{M}^{-1} \mathbf{e}_j} \sqrt{\mathbf{e}'_k (\mathbf{M}^{-1})' \mathbf{M}^{-1} \mathbf{e}_k}} . \end{aligned}$$

The square of the cosine of the angle between the j th and k th predictive biplot axes of the r -dimensional CVA biplot, that is

$$\frac{\mathbf{e}'_j (\mathbf{M}^r)' \mathbf{M}^r \mathbf{e}_k}{\sqrt{\mathbf{e}'_j (\mathbf{M}^r)' \mathbf{M}^r \mathbf{e}_j} \sqrt{\mathbf{e}'_k (\mathbf{M}^r)' \mathbf{M}^r \mathbf{e}_k}} .$$

therefore approximates that correlation coefficient between the j th and k th measured variables, although it is not evident what the associated minimisation criteria is.

Note that if the points $\mathbf{e}'_i (\mathbf{L}^{-1})'$ and $\mathbf{e}'_j (\mathbf{L}^{-1})'$, lie close to the r -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_r)$, then the angle between the vectors, $\mathbf{e}'_i \mathbf{L} \mathbf{V}_r \mathbf{V}'_r$ and $\mathbf{e}'_j (\mathbf{L}^{-1})' \mathbf{V}_r \mathbf{V}'_r$, will be an accurate approximation of the angle between the vectors, $\mathbf{e}'_i (\mathbf{L}^{-1})'$ and $\mathbf{e}'_j (\mathbf{L}^{-1})'$, or equivalently the angle between the vectors,

$$\mathbf{e}'_i (\mathbf{L}^{-1})' \mathbf{V}_r = \mathbf{e}'_i (\mathbf{M}^r)' \quad \text{and} \quad \mathbf{e}'_j (\mathbf{L}^{-1})' \mathbf{V}_r = \mathbf{e}'_j (\mathbf{M}^r)'$$

will be an accurate approximation of the angle between the vectors,

$$\mathbf{e}'_i (\mathbf{L}^{-1})' \mathbf{V} = \mathbf{e}'_i (\mathbf{M}^{-1})' \quad \text{and} \quad \mathbf{e}'_j (\mathbf{L}^{-1})' \mathbf{V} = \mathbf{e}'_j (\mathbf{M}^{-1})' .$$

Since the i th and j th within-group axis predictivities will both be close to one if and only if the points, $\mathbf{e}_i'(\mathbf{L}^{-1})'$ and $\mathbf{e}_j'(\mathbf{L}^{-1})'$, lie close to r -dimensional CVA biplot space, $\mathcal{V}(\mathbf{V}_r)$, it follows that if both the i th and j th within-group axis predictivities are close to one, then the square of the cosine of the angle between the i th and j th predictive biplot axes of the r -dimensional CVA biplot will be an accurate approximation of the correlation coefficient between the i th and j th measured variables.

The above mentioned concepts will now be illustrated at the hand of the *race* data set described in Gower *et al.* (2011). The *race* data set provides information on 799 individuals from four different race groups, namely Black, Coloured, Indian and White. Each of the 799 individuals was measured on six variables namely (1) a total score on a Literacy Assessment Module (*TOTScore*), (2) the number of years of education completed by the respondent (*edyrs*), (3) the age of the respondent (*age*), (4) the decile of the monthly expenditure per household member (*pcexpdec*), (5) the number of years of education completed by the mother of the respondent (*mEdY*) and (6) the race group of the respondent. Amongst the 799 individuals in the data set, 605 are Black, 90 are Coloured, 19 are Indian and 85 are White. The two-dimensional predictive weighted CVA biplot of the *race* data set that only shows the predictive biplot axes is provided in Figure 5.2 and the sample within-group correlation matrix associated with the data set is provided in Table 5.18.

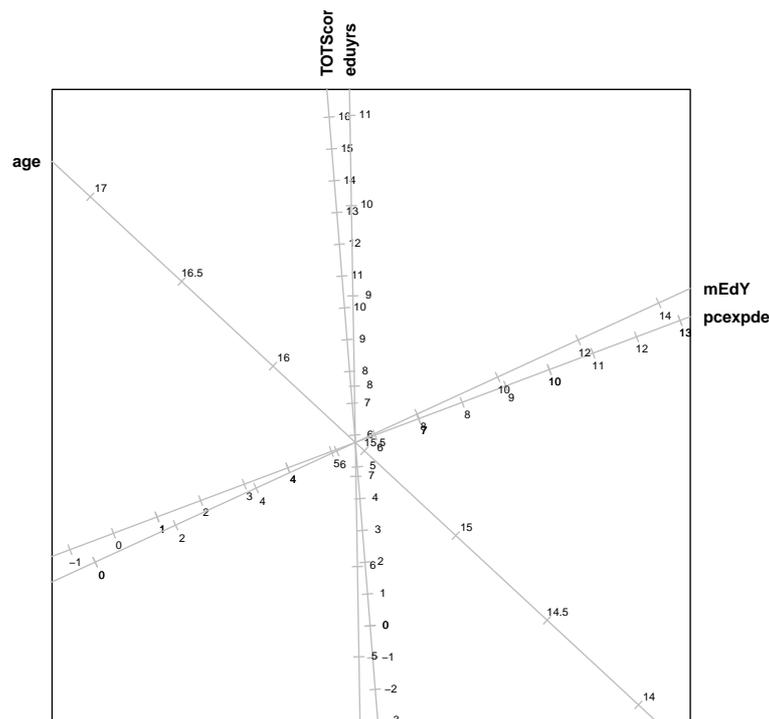


Figure 5.2: *Two-dimensional weighted predictive CVA biplot of the race data set.*

Note that the biplot axes representing these two variables have very high within-group axis predictivities. It can therefore be expected that the size of the angle

between the biplot axes representing the variables $pcexpdec$ and $TOTScore$ will be indicative of the magnitude of the correlation coefficient between the two variables. The angle between the two biplot axes seem to indicate that the variables $pcexpdec$ and $TOTScore$ are weakly positively correlated. This is confirmed by the small positive correlation coefficient between these two variables (see Table 5.18). The small angle between the biplot axes representing the variables $mEdY$ and $pcexpdec$ seem to indicate a high correlation between the two variables. It is however evident from the sample within-group correlation matrix provided in Table 5.18 that the two variables are only weakly positively correlated.

Table 5.18: *The sample within-group correlation matrix associated with the race data set.*

	TOTScore	eduyrs	age	pcexpdec	mEdY
TOTScore	1.00	0.39	0.14	0.25	0.21
eduyrs	0.39	1.00	0.47	0.25	0.24
age	0.14	0.47	1.00	-0.02	-0.1
pcexpdec	0.25	0.25	-0.02	1.00	0.38
mEdY	0.21	0.24	-0.1	0.38	1.00

Table 5.19: *The within-group axis predictivities associated with the two-dimensional weighted CVA biplot of the race data set.*

TOTScore	eduyrs	age	pcexpdec	mEdY
0.9952	0.1968	0.0516	0.9279	0.3419

Consider the within-group axis predictivities corresponding to the two-dimensional weighted predictive CVA biplot of the *race* data set provided in Table 5.19. Note the very low within-group axis predictivity of the biplot axis that represents the variable $mEdY$. This low within-group axis predictivity explains why the angle between the two biplot axes representing the variables $mEdY$ and $pcexpdec$ in the biplot in Figure 5.2 is misleading.

5.13.2 Scale invariance

The k th within-group axis predictivity corresponding to the CVA biplot constructed from the standardised measurements,

$$\pi_k^{W^*} = \frac{[(\mathbf{M}^{*r})' (\mathbf{M}^*)^r]_{kk}}{[(\mathbf{M}^*)^{-1r} (\mathbf{M}^*)^{-1}]_{kk}},$$

is identical to the k th within-group axis predictivity corresponding to the CVA biplot constructed from the unstandardised measurements,

$$\pi_k^W = \frac{[(\mathbf{M}^r)' \mathbf{M}^r]_{kk}}{[(\mathbf{M}^{-1})' \mathbf{M}^{-1}]_{kk}}.$$

This is evident upon substituting $\mathbf{A}\mathbf{M}$ for \mathbf{M}^* in the expression for π_k^W :

$$\begin{aligned} \pi_k^{W^*} &= \frac{[(\mathbf{M}^*)^{r'} (\mathbf{M}^*)^r]_{kk}}{[(\mathbf{M}^*)^{-1'} (\mathbf{M}^*)^{-1}]_{kk}} \\ &= \frac{[\mathbf{A}^{-1} (\mathbf{M}^r)' \mathbf{M}^r \mathbf{A}^{-1}]_{kk}}{[\mathbf{A}^{-1} (\mathbf{M}^{-1})' \mathbf{M}^{-1} \mathbf{A}^{-1}]_{kk}} \\ &= \frac{\frac{1}{a_{kk}} [(\mathbf{M}^r)' \mathbf{M}^r]_{kk} \frac{1}{a_{kk}}}{\frac{1}{a_{kk}} [(\mathbf{M}^{-1})' \mathbf{M}^{-1}]_{kk} \frac{1}{a_{kk}}} \\ &= \frac{[(\mathbf{M}^r)' \mathbf{M}^r]_{kk}}{[(\mathbf{M}^{-1})' \mathbf{M}^{-1}]_{kk}} \\ \longrightarrow \pi_k^{W^*} &= \pi_k^W. \end{aligned}$$

Like the adequacy measure the within-group axis predictivity measure is only invariant to non-singular linear transformations of the form $\mathbf{x} \longrightarrow \mathbf{F}'\mathbf{x}$ when the matrix \mathbf{F} is a $p \times p$ non-singular diagonal matrix.

5.13.3 The relationship between axis predictivities and within-group axis predictivities

Recall that when the dimension of the CVA biplot is equal to $K = \min(J - 1, p)$, all p the axis predictivities are equal to one, the reason being that $\mathbf{C}^{1/2} \widehat{\mathbf{X}} \mathbf{M}_K \mathbf{M}^K = \mathbf{C}^{1/2} \overline{\mathbf{X}}$. However, since it is not true that $\widehat{\mathbf{K}} \mathbf{M}_K \mathbf{M}^K = \mathbf{K}$, it follows that the p within-group axis predictivities associated with the K -dimensional CVA biplot will not all equal one. That is, each of the within-group axis predictivities associated with the K -dimensional CVA biplot will be at most as large as the corresponding axis predictivity. It can be shown that this relationship is true in general. That is, it can be shown that the k th axis predictivity associated with the r -dimensional CVA biplot is at least as large as the k th within-group axis predictivity, $r \in [1 : p]$, $k \in [1 : p]$. The derivation of this result is very similar to the derivation showing that the k adequacy associated with the r -dimensional PCA biplot is at most as large as the k th axis predictivity associated with that biplot. Consider the expressions of

the k th within-group axis predictivity, π_k^W , and the k th axis predictivity, π_k :

$$\pi_k^W = \frac{\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2}{\sum_{j=1}^p ([\mathbf{M}^{-1}]_{jk})^2}$$

$$\pi_k = \frac{\sum_{j=1}^r \lambda_j ([\mathbf{M}^{-1}]_{jk})^2}{\sum_{j=1}^p \lambda_j ([\mathbf{M}^{-1}]_{jk})^2}.$$

It is evident that in order to show that the k th axis predictivity is at least as large as the k th within-group axis predictivity, it needs to be shown that

$$\frac{\sum_{j=1}^r \lambda_j ([\mathbf{M}^{-1}]_{jk})^2}{\sum_{j=1}^p \lambda_j ([\mathbf{M}^{-1}]_{jk})^2} \geq \frac{\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2}{\sum_{j=1}^p ([\mathbf{M}^{-1}]_{jk})^2}.$$

That is, it needs to be shown that

$$\rightarrow \frac{\sum_{j=1}^r \lambda_j ([\mathbf{M}^{-1}]_{jk})^2}{\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2} \geq \frac{\sum_{j=1}^p \lambda_j ([\mathbf{M}^{-1}]_{jk})^2}{\sum_{j=1}^p ([\mathbf{M}^{-1}]_{jk})^2}.$$

Let S_r denote the summation,

$$S_r = \frac{\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j}{\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2}, \quad r \in [1 : p].$$

It is evident that

$$\pi_k \geq \pi_k^W \iff S_r \geq S_p.$$

It will now be shown that $S_r \geq S_{r+1} \forall r \in [1 : (p-1)]$, and hence that $S_r \geq S_p \forall r \in [1 : p]$.

$$S_r - S_{r+1} = \frac{\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j}{\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2} - \frac{\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j}{\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2}$$

$$\begin{aligned}
 &= \frac{\left(\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2\right) \sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j - \left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j}{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \left(\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2\right)} \\
 &= \frac{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j + ([\mathbf{M}^{-1}]_{(r+1)k})^2 \sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j}{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \left(\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2\right)} \\
 &= \frac{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j + \left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) ([\mathbf{M}^{-1}]_{(r+1)k})^2 \lambda_{r+1}}{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \left(\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2\right)} \\
 &= \frac{([\mathbf{M}^{-1}]_{(r+1)k})^2 \sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j - \left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) ([\mathbf{M}^{-1}]_{(r+1)k})^2 \lambda_{r+1}}{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \left(\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2\right)} \\
 &= \frac{([\mathbf{M}^{-1}]_{(r+1)k})^2 \left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_j - \sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2 \lambda_{r+1}\right)}{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \left(\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2\right)} \\
 \longrightarrow S_r - S_{r+1} &= \frac{([\mathbf{M}^{-1}]_{(r+1)k})^2 \left(\sum_{j=1}^r (\lambda_j - \lambda_{r+1}) ([\mathbf{M}^{-1}]_{jk})^2\right)}{\left(\sum_{j=1}^r ([\mathbf{M}^{-1}]_{jk})^2\right) \left(\sum_{j=1}^{r+1} ([\mathbf{M}^{-1}]_{jk})^2\right)}.
 \end{aligned}$$

It is evident that the sign of $S_r - S_{r+1}$ only depends on the sign of $\lambda_j - \lambda_{r+1}$ for $j \in [1 : r]$. Since $\lambda_j \geq \lambda_{r+1}$ when $j \in [1 : r]$, it follows that $\lambda_j - \lambda_{r+1} \geq 0 \forall j \in [1 : r]$ and hence that

$$\begin{aligned}
 S_r - S_{r+1} &\geq 0 \quad \forall r \in [1 : (p-1)] \\
 \longrightarrow S_r &\geq S_{r+1} \quad \forall r \in [1 : (p-1)].
 \end{aligned}$$

It follows that

$$S_1 \geq S_2 \geq \dots \geq S_p$$

and hence that

$$S_r \geq S_p \quad \forall r \in [1 : (p-1)].$$

■

5.14 Changing the CVA biplot scaffolding axes

Up to now the r -dimensional CVA biplot has only been constructed from the first r right singular vectors of the matrix $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ (or equivalently, the first r canonical variables as defined in Section 4.2). Any r of the p right singular vectors of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ can however be used to construct an r -dimensional CVA biplot. It is however important to remember that only the first K right singular vectors of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ contribute to the separation of the groups and that any r -dimensional CVA biplot that is constructed from a set of r right singular vectors other than the set of the first r right singular vectors, will separate the groups to a lesser extent than does the CVA biplot constructed from the first r right singular vectors.

When investigating a data set, additional r -dimensional CVA biplots to the one constructed from the first r right singular vector of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ can be used to obtain additional information. If for example the investigator is particularly interested in the accurate representation of one (or more) of the measured variables, he/she can construct a CVA biplot from those right singular vectors of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ which will yield a CVA biplot with the highest possible axis predictivities and within-group axis predictivities for those variables. If on the other hand if it is of particular interest to the investigator that the spread of the samples around their corresponding group centroids is accurately represented for a particular group (or groups), a CVA biplot in which the overall within-group sample predictivities associated with this group (or groups) can be constructed.

In the case where an r -dimensional CVA biplot that is constructed from a set of r right singular vectors other than the set of the first r right singular vectors, let $\mathbf{V}_r^\#$ denote the $p \times r$ matrix containing the right singular vectors from which the biplot was constructed as column vectors. The quality measures associated with the r -dimensional CVA biplot that is constructed from the column vectors of $\mathbf{V}_r^\#$ can be defined in exactly the same way as before but substituting the matrix $\mathbf{V}_r^\#$ for \mathbf{V}_r in the definitions of the quality measures given throughout this chapter.

Recall that the j th largest eigenvalue of the two-sided eigenvalue problem,

$$\bar{\mathbf{X}}' \mathbf{C} \bar{\mathbf{X}} \mathbf{m} = \lambda \mathbf{W} \mathbf{m} \quad (5.14.1)$$

measures the extent to which the j th canonical variable as defined in Section 4.2 separates the groups i.e. it measures the contribution of the j th right singular vector of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ to the separation of the groups. It follows that the summation of the eigenvalues of the two-sided eigenvalue problem in (5.14.1) corresponding to the eigenvectors that are given by the column vectors of $\mathbf{L}\mathbf{V}_r^\#$ can be used to measure the extent to which the groups are separated in the r -dimensional CVA biplot constructed from the r column vectors of $\mathbf{L}\mathbf{V}_r^\#$. Note that the overall quality with respect to the canonical variables associated with the r -dimensional CVA biplot constructed from the column vectors of $\mathbf{V}_r^\#$, is given by the ratio of the summation of the corresponding eigenvalues of the two-sided eigenvalue problem in (5.14.1) to the summation of the first K eigenvalues of that eigenvalue problem. It follows that

the overall quality with respect to the canonical variables associated with the CVA biplot constructed from the column vectors of $\mathbf{V}_r^\#$ can be used to measure the extent to which the groups are separated in that CVA biplot. This value can be compared to the overall quality with respect to the canonical variables associated with the CVA biplot constructed from the first r right singular vectors of $\mathbf{C}^{1/2}\bar{\mathbf{X}}\mathbf{L}$ to gain insight as to how much worse the separation of the group are in the CVA biplot constructed from the column vectors of $\mathbf{V}_r^\#$ than in the CVA biplot constructed from the column vectors of \mathbf{V}_r .

5.15 Summary

As is the case with a PCA biplot, the relationships and predictions suggested by a CVA biplot are meaningless if the biplot is not an accurate representation of reality. The overall quality of a CVA biplot can be measured with respect to the canonical variables or with respect to the original measured variables. The overall quality with respect to the canonical variables measures the overall accuracy of the approximations of the elements of the matrix of canonical means, $\bar{\mathbf{X}}\mathbf{L}$ and provides information regarding the extent to which the groups are separated in the CVA biplot compared to in the p -dimensional canonical space. The overall quality with respect to the original variables on the other hand measures the overall accuracy of the approximations to the elements of the matrix observed group means, $\bar{\mathbf{X}}$. The extent to which the groups are separated in the CVA biplot provides useful information to the investigator regarding the quality of the representation of the group structure underlying the data set under investigation. The investigator is however typically not interested in the accuracy of the approximations of the measurements on the canonical variables, but rather in the accuracy of the approximations of the measurements on the original measured variables. The investigator will therefore typically consider the overall quality with respect to the original variables when assessing the overall predictive ability of the CVA biplot. The overall quality of the CVA biplot with respect to the original variables is however a very crude measure of the quality of the CVA biplot which does not necessarily suggest accurate information about the quality of the individual biplot axes or the quality of the representation of the individual group centroids and individual samples. Measures of the quality of the different individual aspects of the CVA biplot are required so that those aspects that are accurately represented in the biplot can be identified - conclusions regarding these aspects alone should be made from the visual inspection of the biplot.

The adequacy of a CVA biplot axis is equal to the ratio of the squared length representing one unit of the corresponding variable in the CVA biplot to the squared length representing one unit of that variable in the p -dimensional canonical space. The axis predictivity of a CVA biplot axis measures that axis' predictive ability with respect to the measurements of the group centroids whereas the within-group axis predictivity of that biplot axis measures that axis' predictive ability with respect to the deviations of the individual samples from their corresponding group centroids. The axis predictivity of a CVA biplot axis is at least as large as the within-group axis predictivity of that biplot axis. Group predictivity is the analogue of the sample

predictivity measure of the PCA biplot in the context of the CVA biplot. The group predictivity associated with a group measures the overall accuracy with which that group's mean measurements are predicted in the CVA biplot. All J the group predictivities and all p the axis predictivities associated with the K -dimensional CVA biplot, where $K = \min(J - 1, p)$, are equal to their maximum values.

Sample predictivities and within-group sample predictivities respectively measure the accuracy with which the individual samples' measurements are predicted and the accuracy with which the deviations of the individual samples from their respective group centroids are predicted. Neither all n the sample predictivities nor all n the within-group sample predictivities will be equal to their maximum values in the K -dimensional CVA biplot - only in the p -dimensional CVA biplot will all these measures be equal to their maximum values. The overall accuracy of the approximations to the samples belonging to a particular group can be measured by the overall sample predictivity associated with that group, which is defined as a weighted average of the sample predictivities of the samples belonging to that group. Similarly, the overall accuracy of the approximations to the deviations of the samples belonging to a particular group from the centroid of that group can be measured by the overall within-group sample predictivity associated with that group. The overall within-group sample predictivity associated with a particular group is defined as a weighted average of the within-group sample predictivities of the samples belonging to that group. The overall accuracy of the approximations of all the samples in the data set can be measured by the total sample predictivity, which is defined as a weighted average of the sample predictivities of all the samples in the data set.

All the quality measures discussed in this Chapter except for the overall quality with respect to the original variables take on identical values for the CVA biplot constructed from the unstandardised measurements and that constructed from the standardised measurements. Only a subset of these quality measures are however invariant to all non-singular transformations of the form $\mathbf{x} \rightarrow \mathbf{F}'\mathbf{x}$, where \mathbf{F} denotes an arbitrary $p \times p$ non-singular matrix.

Three quality measures were proposed to assess the accuracy with which the Pythagorean distances in the CVA biplot approximates the corresponding Pythagorean distances in the canonical space. Group contrast predictivities measure the accuracy of the Pythagorean distances between the group centroids in the CVA biplot, sample contrast predictivities measure the accuracy of the Pythagorean distances between the individual samples in the CVA biplot and mixed contrast predictivities measure the accuracy of the Pythagorean distances between the individual samples and the group centroids in the CVA biplot.

Of these three quality measures, it is probably the mixed contrast predictivity measure that is most informative. The set of mixed contrast predictivities associated with a particular group provides information about the accuracy of the representation of the within-group dispersion of that group in the CVA biplot. The set of mixed contrast predictivities associated with a particular sample can suggest whether or not it is likely that the classification of the sample based on the classification regions of the CVA biplot will be the same as that based on the exact (K -dimensional) CVA classification regions. The values of group contrast predictivities suggest informa-

tion about the extent of the separation of the groups in the CVA biplot compared to in the K -dimensional subspace of the p -dimensional canonical space that perfectly contains the J canonical means i.e. \mathbb{C}^K . It also suggests information about the quality of the representation of the exact (K -dimensional) CVA classification regions in the CVA biplot. Given the values of the mixed contrast predictivities and group contrast predictivities, the sample predictivities do not really provide any additional information about the quality of the representation of the group structure in the CVA biplot.

When it is the accurate representation of the group structure underlying the data set which is of primary interest to the investigator, the quality measures which are most informative are the overall quality with respect to the canonical variables, group predictivities, within-group sample predictivities, overall within-group sample predictivities, group contrast predictivities and mixed contrast predictivities. When it is of particular importance to the investigator to know with respect to which measured variables the groups differ (or are similar), then the axis predictivities and within-group axis predictivities should be considered - if these measures of a biplot axis are not high, then the overlap of the groups with respect to that axis is meaningless.

If the investigator is interested in the accuracy of the approximations of the group centroids and individual samples, then the relevant quality measures to consider are the group predictivities, sample predictivities, overall sample predictivities and the total sample predictivities.

All the quality measures associated with the CVA biplot that were discussed in this Chapter are defined as ratios of sums of squared values. The fact that the decomposition of $\mathbf{C}^{1/2}\overline{\mathbf{X}}\mathbf{L}$ in equation (5.1.2) exhibits Type A and Type B orthogonality validates all the quality measures with respect to the canonical variables while the fact that the decomposition of $\mathbf{C}^{1/2}\overline{\mathbf{X}}$ in equation (5.1.4) exhibits Type A orthogonality in the metric \mathbf{W} as well as Type B orthogonality, validates all the quality measures regarding the measured variables except for the three quality measures that assess the accuracy of the Pythagorean distances in the CVA biplot. The latter quality measures are validated by the orthogonal decompositions of the relevant squared true distances into the squared approximated distances in the CVA biplot and the squared residual distances in the orthogonal complement of the CVA biplot.

Chapter 6 - Conclusion

6.1 What has been achieved in this thesis?

In this thesis the main topics of discussion were the PCA and CVA biplots, with the primary focus falling on the quality measures associated with these biplots. However, since a sound understanding of the statistical analyses that underlie these biplots aids in the interpretation of the biplots, a detailed study of PCA and CVA was provided prior to the investigation of the PCA and CVA biplots respectively. For both PCA and CVA different perspectives from which the analyses can be viewed (as well as the connection between them) were studied as the different perspectives highlight different aspects of the analyses and so contribute to an improved understanding of the biplots and their interpretation. For PCA detailed studies of the routes followed by Pearson and Hotelling were provided. CVA was studied from the perspectives of LDA, CCA and a two-step approach. The close relationship between CVA and MANOVA was also highlighted in an outline of important hypotheses to be tested prior to performing CVA.

Following the study of PCA, an explanation of the construction of the PCA biplot was provided. This explanation commenced with an outline of the construction of the traditional PCA biplot proposed by Gabriel (1971) after which the adjustments proposed by Gower and Hand (1996) were discussed. The study of the PCA biplot was followed by an in depth investigation of the quality measures of the PCA biplot and the relationships between these quality measures. The main contribution of this study to work done on PCA biplots is the investigation of the effect of standardisation of the PCA biplot and the PCA biplot quality measures. Furthermore, a method whereby the sample predictivities associated with the last few principal components are used to detect observations that substantially deviate from the correlation structure of the bulk of the data set was proposed in Chapter 3.

Subsequent to the study of the different perspectives on CVA an explanation of the construction of the CVA biplot was provided. Three types of CVA biplots were discussed namely the weighted CVA biplot and two different unweighted CVA biplots. The main contribution of this study to work that has been done on CVA biplots and their associated quality measures is the investigation that was performed on the effect of accounting for the (possibly) different group sizes in the construction of the CVA biplot on (1) the representation of the group structure underlying a data set and (2) the quality measures of the CVA biplot. This investigation was performed using a few moderately sized simulated data sets. A much larger simulation study is required to confirm the meaningfulness of the preliminary results obtained in Chapters 4 and 5.

The three quality measures relating to the accuracy of the representation of the Pythagorean distances in the p -dimensional canonical space in the CVA biplot which were defined in Chapter 5, namely group contrast predictivities, sample contrast predictivities and mixed contrast predictivities form the second largest contribution of this study to research that has been done on the quality measures associated with CVA biplots. The mixed contrast predictivity measure is arguably the most informative of the three measures as this measure provides information about (1) the quality of the representation of the within-group dispersion in the CVA biplot and (2) information about the classification of the individual samples based on the (approximate) classification regions in the CVA biplot compared to that based on the exact K -dimensional CVA classification regions. The group contrast predictivities suggest information about the extent of the separation of the groups in the CVA biplot compared to in the K -dimensional subspace of the p -dimensional canonical space that perfectly contains the J group centroids i.e. \mathbb{C}^K . It also suggests information about the quality of the representation of the exact (K -dimensional) CVA classification regions in the CVA biplot. Sample contrast predictivity measure provides information about the accuracy with which the Pythagorean distances between the samples in the p -dimensional canonical space are approximated in the CVA biplot space. Given the values of the mixed contrast predictivities and group contrast predictivities, the sample predictivities do not really provide any additional information about the quality of the representation of the group structure in the CVA biplot.

Quality measures that assess the accuracy with which the Pythagorean distances in the PCA biplot approximate the corresponding Pythagorean distances in the measurement space can be defined in a similar way to the three contrast predictivity measures defined for the CVA biplot. It is however the desire for information on the quality of the representation of the group structure underlying the data set represented in a CVA biplot that prompted the need for information regarding the accuracy of the representation of the true Pythagorean distances in the CVA biplot.

A quality measure that is still lacking for the CVA biplot is a single measure that assesses the quality of the representation of the exact K -dimensional CVA classification regions in the CVA biplot. For reasons already mentioned, the mixed contrast predictivities, group predictivities or group contrast predictivities could perhaps form the basis of such a measure.

6.2 The way forward

There are some interesting and important questions that still remain unanswered and could form the basis for new research. Three such questions are: (1) how can PCA biplot quality measures can be used to identify outliers? (2) how should the quality measures corresponding to a robust PCA (RPCA) biplot be defined and (3) how can CVA biplot quality measures aid in a variable selection process?

6.2.1 The robust PCA biplot

In Chapter 3 a simple method whereby which observations that substantially deviate from the correlation structure of the bulk of the data set can be detected using sample predictivities was suggested. Discarding these samples in the construction of the PCA biplot will likely yield a more robust PCA biplot. Instead of discarding influential outliers to obtain a more robust PCA biplot, the construction of the biplot can be based on Robust PCA. The resulting biplot is called a robust PCA biplot (Gardner (2001); Wedlake (2008)). In robust PCA biplots the influence of influential observations on the biplot is restricted. Over the last number of years substantial work has been done on the quality measures of the PCA biplot (Gardner-Lubbe *et al.* (2008); Gower *et al.* (2011)). This work needs to be integrated with work that has been done on robust PCA and the robust PCA biplot.

6.2.2 Variable Selection

When the number of variables on which the samples of a data set is measured, is large, it is often possible to discard some of the variables without much, if any, loss of information. Reducing the number of variables accounted for in a classifier will increase the bias, but reduce the variability of the predictions it produces. Often, especially when the number of variables accounted for in the classifier is large, it is possible to discard some of the variables without increasing the bias of its predictions much, if at all, while at the same time reducing the variability of its predictions to such an extent that the expected prediction error (EPE), or expected misclassification rate, of the classifier is reduced. According to the principal of parsimony, a model (classifier) constructed from a subset of the measured variables is preferred to the model (classifier) constructed from the complete set of measured variables when the model (classifier) based on the subset does not perform statistically significantly worse than the model (classifier) based on the complete set. Hence, when the classifier based on a subset of the measured variables has an EPE smaller or equal to that of the classifier based on the complete set of measured variables, the classifier based on the subset should be used.

Estimates of a classifier's EPE can be calculated using resampling methods like cross validation and bootstrap (Efron and Tibshirani, 1993). When the number of measured variables is small, an estimate of EPE can be calculated for each possible subset of variables and the CVA biplot then based on the subset which produced the smallest estimate EPE. In the case where the smallest estimate of EPE is produced by more than one subset and the subsets are of differing sizes, the principal of parsimony dictates that the smallest subset which produce this minimum estimate of EPE should be used. When more than one subset of the same size yield the smallest estimate of EPE, one of the subsets can be chosen at random. When the number of measured variables is large, such a process might be too time consuming and alternative methods need to be used to identify the subset of variables to work with. Examples of less time consuming variable selection methods are forward and backward stepwise variable selection. It is important to remember that stepwise variable selection procedures will not necessary yield the best possible subset of variables. In addition to this - when the number of variables is very large, even

stepwise selection methods can become too time consuming. It is evident that a need exists for even faster methods of variable selection. It is of particular interest whether or not the quality measures of the CVA biplot can aid in the variable selection process. That is, whether the quality measures of the CVA biplot constructed from the complete set of variables can be used to identify a subset of variables associated with a smaller estimate of EPE (or expected misclassification rate). Once the variables that together yield the greatest separation of the groups in two / three dimensions have been identified a CVA biplot can be constructed from the the subset of the original data set that corresponds to these variables.

Some research on this topic has already been done. Gardner (2001) showed that the adequacies of the CVA biplot axes cannot differentiate between a variable that individually separates the groups substantially but does not contribute much to the separation of the groups when considered together with one or more other variables, i.e. a redundant variable, and a variable that both separates the groups on its own and substantially adds to the separation of the groups when considered together with one or more other variables Guyon and Elisseff (2003) demonstrated that it is possible that two or more irrelevant variables, that is variables that individually cannot separate the groups, can separate the groups very well if they are used together.

It should be investigated whether or not one (or more) of the predictivity measures associated with the CVA biplot can aid in the variable selection process.

6.3 To conclude...

In the words of Chambers *et al.* (1983) “there is no single statistical tool that is as powerful as a well-chosen graph”.

SOLI DEO GLORIA

Bibliography

- Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley: New York.
- Anton, H. and Rorres, C. (2000). *Elementary Linear Algebra, Applications Version*. Eighth edition edn. John Wiley & Sons, Inc: New York.
- Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. Third edition edn. Chichester: Wiley.
- Bartlett, M. (1938). Further aspects of the theory of multiple regression. In: *Proceedings of the Cambridge Philosophical Society*, vol. 34, pp. 33–40.
- Box, G. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, vol. 36, pp. 317–346.
- Box, G. (1953). Non-normality and tests on variances. *Biometrika*, vol. 40, pp. 318–335.
- Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, vol. 1 (2), pp. 245–276.
- Chambers, J., Cleveland, W., Kleiner, B. and Tukey, P. (1983). *Graphical Methods for Data Analysis*. Wadsworth International Group: Belmont, California.
- Cox, T. and Cox, M. (2001). *Multidimensional Scaling*. Second edition edn. Chapman & Hall/CRC: Boca Raton, FL.
- Critchley, F. (1985). Influence in principal component analysis. *Biometrika*, vol. 7, pp. 627–636.
- Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, vol. 1 (3), pp. 211–218.
- Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap (Monographs on Statistics and Applied Probability, No. 57)*. Chapman and Hall, London.
- Everitt, B. (1994). Exploring multivariate data graphically: A brief review with examples. *Journal of Applied Statistics*, vol. 21 (3), pp. 63–93.
- Fisher, A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, vol. 7, pp. 179–188.

- Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer Verlag: New York.
- Gabriel, K. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, vol. 58, pp. 453–467.
- Gabriel, K. (1972). Analysis of meteorological data by means of canonical decomposition of biplots. *Journal of Applied Meteorology*, vol. 11, pp. 1071–1077.
- Gardner, S. (2001). *Extensions of Biplot Methodology to Discriminant Analysis with Applications of Non-parametric Principal Components*. Ph.D. thesis, University of Stellenbosch.
- Gardner, S., Gower, J.C. and Le Roux, N.J. (2006). A synthesis of canonical variate analysis, generalised canonical correlation and procrustes analysis. *Computational Statistics and Data Analysis*, vol. 50, pp. 107 – 134.
- Gardner, S. and Le Roux, N.J. (2003). Graphics and visualisation in practice: Biplots for exploring multidimensional reality. In: *Bulletin of the International Statistical Institute, 54th proceedings*, pp. 270–273. Berlin, Germany.
- Gardner, S. and Le Roux, N.J. (2004). Modified Biplots for Enhancing Two-class Discriminant Analysis. In: *Classification, clustering and data mining applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago*, pp. 233–240. Springer-Verlag, New York.
- Gardner, S., Le Roux, N.J., Rypstra, T. and Swart, J. (2005). Extending a Scatterplot for Displaying Group Structure in Multivariate Data: A Case Study. *ORiON*, vol. 21, pp. 111–124.
- Gardner-Lubbe, S., Le Roux, N.J. and Gower, J.C. (2008). Measures of Fit in Principal Component and Canonical Variate Analyses. *Journal of Applied Statistics*, vol. 35 (9), pp. 947–965.
- Gittins, R. (1985). *Canonical Analysis: A Review with Applications in Ecology*. Springer-Verlag: Berlin.
- Gnanadesikan, R. and Kettinger, J. (1972). Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics*, vol. 28, pp. 81–124.
- Gower, J. and Hand, D. (1996). *Biplots*. Chapman and Hall: London.
- Gower, J., Lubbe, S. and Le Roux, N. (2011). *Understanding Biplots*.
- Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press: London.
- Guyon, I. and Elisseeff, A. (2003). Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182.

- Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. (1986). *Robust Statistics: An Approach Based on Influence Functions*. Wiley: New York.
- Harville, D. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag: New York.
- Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible Discriminant Analysis By Optimal Scoring. *Journal of the American Statistical Association*, vol. 89, pp. 1255–1270.
- Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall: London.
- Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520.
- Hotelling, H. (1935). The Most Predictable Criterion. *Journal of Educational Psychology*, vol. 26, pp. 139–192.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, vol. 28, pp. 321–377.
- Hubert, M., Rousseeuw, P. and Branden, K.V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, vol. 47, pp. 64–79.
- Johnson, R. and Wichern, D. (2002). *Applied Multivariate Statistical Analysis*. 5th edn. Prentice-Hall, Inc. Upper Saddle River: New Jersey.
- Jolliffe, I. (2002). *Principal Component Analysis*. Second edition edn. Springer-Verlag: New York.
- Kshirsagar, A. (1972). *Multivariate Analysis*. Marcel Dekker, Inc: New York.
- Lehmann, E. (1988). *Encyclopedia of Statistical Sciences. Statistics: An overview*, vol. 8. Wiley: New York.
- Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate Analysis*. 5th edn. Academic Press: London.
- Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, vol. 2, pp. 559 – 572.
- R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at: <http://www.R-project.org/>
- Rao, C. (1948). The Utilization of Multiple Measurements in Problems of Biological classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 9, pp. 89–324.
- Rao, C. (1952). *Advanced Statistical Methods in Biometric Research*. John Wiley & Sons, Inc: New York.

Rao, C. (1965). *Linear Statistical Inference and Its Applications*. Wiley: New York.

Stewart, J. (2003). *Calculus: International Student Edition*. 5th edn.

Tiku, M. and Balakrishnan, N. (1985). Testing the Equality of Variance-Covariance Matrices the Robust Way. *Communications in Statistics - Theory and Methods*, vol. 14, pp. 3033 – 3051.

Wedlake, R. (2008). *Robust Principal Component Analysis Biplots*. Master's thesis, University of Stellenbosch.