

**ASPECTS OF THE PRE- AND POST-SELECTION
CLASSIFICATION PERFORMANCE OF
DISCRIMINANT ANALYSIS AND
LOGISTIC REGRESSION**

NELMARIE LOUW



Dissertation presented for the Degree of Doctor of Philosophy at the University of Stellenbosch

Promoter: Prof. N.J. Le Roux

Co-promoter: Prof. S.J. Steel

Date: November 1997

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature:

Date:

OPSOMMING

Lineêre diskriminantanalise en logistiese regressie is tegnieke wat gebruik kan word vir die klassifikasie van items van onbekende oorsprong in een van 'n aantal groepe. Die agterliggende modelle en aannames vir die gebruik van die twee tegnieke is egter verskillend. In die studie is die twee tegnieke vergelyk ten opsigte van klassifikasie van items.

Eerstens is die twee tegnieke vergelyk in 'n opset waar daar geen data-afhanklike seleksie van veranderlikes plaasvind nie. Verskeie onderliggende verdelings is bestudeer: die normaalverdeling, die dubbeleksponensiaal-verdeling, en die lognormaal verdeling. Die aantal veranderlikes, steekproefgroottes uit die onderskeie groepe en die korrelasiestruktuur tussen die veranderlikes is gevarieer om 'n groot aantal konfigurasies te verkry. Die geval van twee en drie groepe is bestudeer. Die belangrikste gevolgtrekkings wat op grond van die studie gemaak kan word is: vir normaal en dubbeleksponensiaal data vaar lineêre diskriminantanalise beter as logistiese regressie, veral in gevalle waar die verhouding van die aantal veranderlikes tot die totale steekproefgrootte groot is. In die geval van data uit 'n lognormaalverdeling, behoort logistiese regressie die metode van keuse te wees, tensy die verhouding van die aantal veranderlikes tot die totale steekproefgrootte groot is.

Veranderlike seleksie is dikwels die eerste stap in statistiese ontledings. 'n Groot aantal potensieel belangrike veranderlikes word waargeneem, en 'n subversameling wat optimaal is, word gekies om in die verdere ontledings te gebruik. Ten spyte van die feit dat veranderlike seleksie dikwels gebruik word, word die invloed wat 'n seleksie-stap op verdere ontledings van dieselfde data het, dikwels heeltemal geïgnoreer. 'n Belangrike doelwit van die studie was om nuwe seleksietegnieke te ontwikkel wat gebruik kan word in diskriminantanalise en logistiese regressie. Verder is ook aandag gegee aan ontwikkeling van beramers van die foutkoers van 'n diskriminantfunksie wat met geselekteerde veranderlikes gevorm word. 'n Nuwe seleksietegniek, kruis-model validasie (KMV) wat gebruik kan word vir die seleksie van veranderlikes in beide diskriminantanalise en logistiese regressie is ontwikkel. Hierdie tegniek hanteer die seleksie van veranderlikes en die beraming van die na-seleksie foutkoers in een stap, en verskaf 'n metode om die optimale modeldimensie te bepaal, die veranderlikes wat in die model bevat moet word te kies, en ook die na-seleksie foutkoers van die diskriminantfunksie te beraam. 'n Uitgebreide simulasiestudie waarin die voorgestelde KMV-tegniek met ander prosedures in die literatuur vergelyk is, is vir beide diskriminantanalise en logistiese regressie onderneem. In die algemeen het hierdie tegniek beter gevaar as die ander metodes wat beskou is, veral ten opsigte van die akkuraatheid waarmee die na-seleksie foutkoers beraam word.

Ten slotte is daar ook aandag gegee aan voor-toets tipe seleksie. 'n Tegniek is ontwikkel wat gebruik maak van 'n voor-toets beramingsmetode om veranderlikes vir insluiting in 'n lineêre diskriminantfunksie te selekteer. Die tegniek is in 'n simulasiestudie met die KMV-tegniek vergelyk, en vaar baie goed, veral t.o.v. korrekte seleksie. Hierdie tegniek is egter slegs geldig vir ongekorreleerde normaalveranderlikes, wat die gebruik daarvan beperk.

'n Numeries intensiewe benadering is deurgaans in die studie gebruik. Dit is genoodsaak deur die feit dat die probleme wat ondersoek is, nie deur middel van 'n analitiese benadering hanteer kan word nie.

SUMMARY

Discriminant analysis and logistic regression are techniques that can be used to classify entities of unknown origin into one of a number of groups. However, the underlying models and assumptions for application of the two techniques differ. In this study, the two techniques are compared with respect to classification of entities.

Firstly, the two techniques were compared in situations where no data dependent variable selection took place. Several underlying distributions were studied: the normal distribution, the double exponential distribution and the lognormal distribution. The number of variables, sample sizes from the different groups and the correlation structure between the variables were varied to obtain a large number of different configurations. The cases of two and three groups were studied. The most important conclusions are: for normal and double exponential data linear discriminant analysis outperforms logistic regression, especially in cases where the ratio of the number of variables to the total sample size is large. For lognormal data, logistic regression should be preferred, except in cases where the ratio of the number of variables to the total sample size is large.

Variable selection is frequently the first step in statistical analyses. A large number of potentially important variables are observed, and an optimal subset has to be selected for use in further analyses. Despite the fact that variable selection is often used, the influence of a selection step on further analyses of the same data, is often completely ignored. An important aim of this study was to develop new selection techniques for use in discriminant analysis and logistic regression. New estimators of the post-selection error rate were also developed. A new selection technique, cross model validation (CMV) that can be applied both in discriminant analysis and logistic regression, was developed. This technique combines the selection of variables and the estimation of the post-selection error rate. It provides a method to determine the optimal model dimension, to select the variables for the final model and to estimate the post-selection error rate of the discriminant rule. An extensive Monte Carlo simulation study comparing the CMV technique to existing procedures in the literature, was undertaken. In general, this technique outperformed the other methods, especially with respect to the accuracy of estimating the post-selection error rate.

Finally, pre-test type variable selection was considered. A pre-test estimation procedure was adapted for use as selection technique in linear discriminant analysis. In a simulation study, this technique was compared to CMV, and was found to perform well, especially with respect to correct selection. However, this technique is only valid for uncorrelated normal variables, and its applicability is therefore limited.

A numerically intensive approach was used throughout the study, since the problems that were investigated are not amenable to an analytical approach.

To Jacques, Willem and Gerard

ACKNOWLEDGEMENTS

I wish to express my gratitude to:

- **Prof. N. J. Le Roux, my promoter, and Prof. S.J. Steel, my co-promoter, for their invaluable guidance and encouragement throughout this study.**
- **The University of Stellenbosch and the Potchefstroom University for CHE, for the use of their computer facilities.**
- **The Foundation for Research Development, for financial assistance.**
- **My husband, sons and family, for their continuous support.**

CONTENTS

LIST OF CODES USED IN FIGURES	xii
CHAPTER 1 - INTRODUCTION AND SCOPE OF THE THESIS	1
1.1 An overview of classification procedures	1
1.2 Aims and scope of the thesis	2
1.3 The numerically intensive approach	4
1.4 Main contribution	5
CHAPTER 2 - A COMPARISON OF THE CLASSIFICATION PERFORMANCE OF DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION	6
2.1 Introduction : Discriminant analysis and logistic regression	6
2.2 Error rates	11
2.3 Overview of literature comparing discriminant analysis and logistic regression	21
2.4 Monte Carlo simulation study: Two groups	23
2.4.1 The normal case	24
2.4.2 The double exponential case	33
2.4.3 The lognormal case	44
2.5 Monte Carlo simulation study: Three groups	52
2.5.1 The normal case	53
2.5.2 The double exponential case	61
2.5.3 The lognormal case	68
2.6 Comparison of fully polychotomous and individualised binary logistic regression	76
2.7 Conclusions and recommendations	90
CHAPTER 3 - VARIABLE SELECTION AND THE CLASSIFICATION PERFORMANCE OF THE LINEAR DISCRIMINANT FUNCTION	91
3.1 Introduction	91
3.2 Overview of techniques used for variable selection in discriminant analysis	93
3.3 The effect of model dimension on the properties of the resulting classification rule (no selection)	101

3.3.1	The normal case	104
3.3.2	The lognormal case	114
3.4	Comparison of different methods to select a pre-specified number of variables	120
3.4.1	The normal case	121
3.4.2	The lognormal case	124
3.5	The effect of model dimension on the properties of the resulting classification rule (with selection)	126
3.5.1	Comparison of post-selection error rates	126
3.5.1.1	The normal case	127
3.5.1.2	The lognormal case	133
3.5.2	The effect of dimension on post-selection error rate	138
3.6	Conclusions and recommendations	139

CHAPTER 4 - VARIABLE SELECTION AND ERROR RATE ESTIMATION IN DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION BY MEANS OF CROSS MODEL VALIDATION

141

4.1	Introduction	141
4.2	Overview of literature on post-selection error rate estimation	143
4.3	Cross model validation	146
4.3.1	General principles	146
4.3.2	Cross model validation in a regression context	148
4.4	Cross model validation in discriminant analysis	150
4.5	Monte Carlo simulation study for discriminant analysis	154
4.5.1	Inner criterion: forward stepwise selection	155
4.5.1.1	The normal case	155
	- Selection performance	156
	- Expected actual error rate	157
	- Probability of correct selection	157
	- Estimation performance	158
	- Bias	158
	- Unconditional mean squared error	159
4.5.1.2	The double exponential case	168
	- Selection performance	168
	- Expected actual error rate	168
	- Probability of correct selection	169
	- Estimation performance	169
	- Bias	169
	- Unconditional mean squared error	169
4.5.1.3	The lognormal case	175
	- Selection performance	175
	- Expected actual error rate	175
	- Probability of correct selection	176

	- Estimation performance	176
	- Bias	176
	- Unconditional mean squared error	176
4.5.2	Inner criterion : all possible subsets selection based on R^2	182
4.5.2.1	Selection performance	183
	- Expected actual error rate	183
	- Probability of correct selection	183
4.5.2.2	Estimation performance	183
	- Bias	183
	- Unconditional mean squared error	184
4.6	Cross model validation in logistic regression	189
4.7	Monte Carlo simulation study for logistic regression	194
4.7.1	The normal case	195
4.7.1.1	Expected actual error rate	196
4.7.1.2	Bias	196
4.7.1.3	Unconditional mean squared error	196
4.7.2	The double exponential case	200
4.7.2.1	Expected actual error rate	200
4.7.2.2	Bias	200
4.7.2.3	Unconditional mean squared error	200
4.7.3	The lognormal case	204
4.7.3.1	Expected actual error rate	204
4.7.3.2	Bias	204
4.7.3.3	Unconditional mean squared error	204
4.8	Comparison of the performance of cross model validation in discriminant analysis and logistic regression	208
4.8.1	Selection performance	209
4.8.2	Classification performance	210
4.9	Application of cross model validation and other techniques to real life data sets	218
4.9.1	Corporate failure data	218
4.9.2	Swiss bank note data	223
4.10	Conclusions and recommendations	226

CHAPTER 5 - PRE-TEST VARIABLE SELECTION **228**

5.1	Introduction	228
5.2	General aspects of pre-test selection	229
5.3	The PT_q - criterion in discriminant analysis	235
5.4	Error rate estimation	237
5.5	Monte Carlo simulation study	240
5.5.1	Selection performance	241
5.5.1.1	Expected actual error rate	241

5.5.1.2	Probability of correct selection	241
5.5.2	Estimation performance	242
5.5.2.1	Bias	242
5.5.2.2	Unconditional mean squared error	242
CHAPTER 6 - SUMMARY AND DIRECTIONS FOR FURTHER RESEARCH		248
APPENDIX		250
Program 1		250
Program 2		273
Program 3		295
Program 4		318
REFERENCES		332

LIST OF CODES USED IN FIGURES

CHAPTER 2

- Figs. 2.1 - 2.8 and Figs. 2.11 - 2.14: Each of the graphs in these figures is identified by a code of the form DA_x or LR_x with the following interpretation:

DA=discriminant analysis

LR=logistic regression

x=1 : k = 2 feature variables, small sample sizes ($n_0 = n_1 = 25$)

x=2 : k = 2 feature variables, mixed sample sizes ($n_0 = 25, n_1 = 50$)

x=3 : k = 2 feature variables, large sample sizes ($n_0 = n_1 = 100$)

x=4 : k = 10 feature variables, small sample sizes ($n_0 = n_1 = 25$)

x=5 : k = 10 feature variables, mixed sample sizes ($n_0 = 25, n_1 = 50$)

x=6 : k = 10 feature variables, large sample sizes ($n_0 = n_1 = 100$)

- Figs. 2.15 - 2.26: Each of the graphs in these figures is identified by a code of the form DA_x or LR_x with the following interpretation:

DA=discriminant analysis

LR=logistic regression

x=1 : k = 2 feature variables, small sample sizes ($n_0 = n_1 = n_2 = 25$)

x=2 : k = 2 feature variables, large sample sizes ($n_0 = n_1 = n_2 = 100$)

x=3 : k = 10 feature variables, small sample sizes ($n_0 = n_1 = n_2 = 25$)

x=4 : k = 10 feature variables, large sample sizes ($n_0 = n_1 = n_2 = 100$)

- Figs. 2.27 - 2.34: Each of the graphs in these figures is identified by a code of the form FP_x or IR_x with the following interpretation:

FP=fully polychotomous logistic regression

IR=individualised binary logistic regression

x=1 : k = 2 feature variables, small sample sizes ($n_0 = n_1 = n_2 = 25$)

x=2 : k = 2 feature variables, large sample sizes ($n_0 = n_1 = n_2 = 100$)

x=3 : k = 10 feature variables, small sample sizes ($n_0 = n_1 = n_2 = 25$)

x=4 : k = 10 feature variables, large sample sizes ($n_0 = n_1 = n_2 = 100$)

CHAPTER 3

- Figs. 3.1 - 3.18: Each of the graphs in these figures is identified by a code of the form ABxy with the following interpretation:

A=N, D, L : Normal, Double exponential, Lognormal distributions respectively

B=S, L : small ($n_0 = n_1 = 25$) and large ($n_0 = n_1 = 100$) samples respectively

x=1, 2, 3, 4 : equi-correlated feature variables with common correlation

$\rho = -0.1, 0, 0.4,$ and 0.9 respectively

y=1, 2, 3 : number of components with respect to which the two mean vectors differ, viz. $r = 1, 5$ and 10 respectively

CHAPTER 4

- Figs. 4.1 - 4.7 and Figs. 4.16 - 4.19: Each of the graphs in these figures is identified by a code of the form ABxy with the following interpretation:

A=N : Normal distribution

B=S, M, L : small ($n_0 = n_1 = 25$), mixed ($n_0 = 75, n_1 = 25$) and large ($n_0 = n_1 = 100$) samples respectively

x=1, 2, 3, 4 : number of components with respect to which the two mean vectors differ, viz. $r = 1, r = 5, r = 10$ (components of μ_1 given by (4.5.2)) and $r = 10$ (components of μ_1 given by (4.5.3)) respectively

y=1, 2 : uncorrelated feature variables and equicorrelated feature variables ($\rho = 0.9$) respectively

- Figs. 4.8 - 4.15: Each of the graphs in these figures is identified by a code of the form ABx with the following interpretation:

A=D, L : Double exponential and Lognormal distributions respectively

B=S, M, L : small ($n_0 = n_1 = 25$), mixed ($n_0 = 75, n_1 = 25$) and large ($n_0 = n_1 = 100$) samples respectively

x=1, 2, 3, 4 : number of components with respect to which the two mean vectors differ, viz. $r = 1, r = 5, r = 10$ (components of μ_1 given by (4.5.2)) and $r = 10$ (components of μ_1 given by (4.5.3)) respectively

- Figs. 4.20 - 4.35: Each of the graphs in these figures is identified by a code of the form Ax with the following interpretation:

A=N, D, L : Normal, Double exponential and Lognormal distributions respectively

x=1, 2, 3 : number of components with respect to which the two mean vectors differ, viz. $r = 1$, $r = 5$, and $r = 10$ (components of μ_1 given by (4.5.2))

CHAPTER 5

- Figs. 5.1 - 5.4: Each of the graphs in these figures is identified by a code of the form ABxy with the following interpretation:

A=N : Normal distribution

B=S, M, L : small ($n_0 = n_1 = 25$), mixed ($n_0 = 75, n_1 = 25$) and large ($n_0 = n_1 = 100$) samples respectively

x=1, 2, 3, 4 : number of components with respect to which the two mean vectors differ, viz. $r = 1$, $r = 5$, $r = 10$ (components of μ_1 given by (4.5.2)) and $r = 10$ (components of μ_1 given by (4.5.3)) respectively

y=1: uncorrelated feature variables

CHAPTER 1

INTRODUCTION AND SCOPE OF THE THESIS

1.1 AN OVERVIEW OF CLASSIFICATION PROCEDURES

The classification of entities into distinct groups is frequently an issue of theoretical and practical scientific interest. Examples are: in biological taxonomy, using measurements on certain characteristics to classify a new species into one of several genera; in medical diagnosis, using physiological measurements and diagnostic test results to classify a patient into one of a number of prognostic categories; in banking, using financial information to classify a loan applicant as high or low risk; in finance, using accounting information to classify a company into one of a number of categories relating to the risk of the company being declared bankrupt within the next year. In all of these examples, classification is based on measurements of a number of characteristics of the entities under study. These characteristics will be referred to as feature variables.

Classification problems can be grouped into two broad classes (cf. Gnanadesikan et al., 1989). Firstly, problems arise where so-called training data are available, i.e. data consisting of the values of the feature variables for a number of entities, together with the group to which each of these entities belong. This is referred to as *supervised* classification (or supervised pattern recognition). In supervised classification problems the aim is to use the feature data to construct a function(s) of the feature variables that can be used to classify future entities of which the group membership is unknown, into one of the available groups. It should be noted that in the supervised case, the number and nature of the available groups are clearly specified. The second category of classification problems is called *unsupervised*, or unsupervised pattern recognition. In these problems the number and nature of the groups are not specified beforehand, and the group membership of the entities in the sample data is unknown. The aim in unsupervised classification is to use the sample data to group the sample entities into more or less homogeneous groups. Hence, in these cases the group specification is data-dependent.

A number of statistical techniques have been developed for application to classification problems. The techniques that are suitable for the supervised case are often broadly referred to as *discriminant analysis*, while the term cluster analysis is used for a large collection of algorithms that can be applied in the unsupervised case. In its broad sense, the term discriminant analysis includes classical *linear discriminant analysis* and quadratic discriminant analysis, as well as *logistic regression*. The term will, however, not be used in its broad sense in this thesis. In cases in this thesis where the term discriminant analysis is used, it will mostly refer to the analysis of a data set by means of the classical linear discriminant function. Therefore, when discriminant analysis and

logistic regression are compared in Chapter 2 with respect to classification performance, it is discriminant analysis based on the classical linear discriminant function that is under consideration. Since attention in this thesis is restricted to the supervised case, cluster analysis techniques will not be dealt with.

Discriminant analysis (in its narrow sense) and to a lesser extent also logistic regression, are techniques that depend for their validity on certain parametric assumptions being satisfied. In recent years a number of non-parametric discriminant analysis techniques have been developed that require less restrictive assumptions. Important amongst these are techniques that use various non-parametric estimators of the density functions of the feature variables in the different groups. Kernel density estimators are popular choices in this regard (cf. Silverman, 1986). Another discriminant analysis technique that deserves to be mentioned is classification trees. This technique enjoys growing popularity, and a comprehensive and authoritative reference is Breiman et al. (1984). More recently, Hastie et al. (1994) developed a technique called flexible discriminant analysis based on nonparametric adaptive regression methods. This technique can be applied in cases where the class boundaries are non-linear. Finally, many of the classification problems that can be solved by discriminant analysis techniques, are also amenable to analysis by means of neural networks. The rapidly growing literature on this topic reflects its popularity. Cheng and Titterton (1994) provide a good introduction to and review of the topic, emphasising the close relationship between neural network methodology and a number of statistical techniques.

The above brief survey of classification techniques does not purport to be comprehensive. Nevertheless, it does convey the message that the development of new classification procedures is an area of active research, and that a variety of techniques are available to the researcher who wishes to classify entities.

1.2 AIMS AND SCOPE OF THE THESIS

It is clear from the discussion in the previous paragraph that statistical classification procedures form a wide and diverse field. A large literature on different aspects of such procedures exists, as is evident from the references in Gnanadesikan et al. (1989) and McLachlan (1992). In this section an indication is given of the aspects of statistical classification procedures that are addressed in this thesis.

Attention is restricted throughout the thesis to linear discriminant analysis, based on the well known Anderson classification statistic, and to logistic regression analysis. In Chapter 2, the case of two groups and the case of three groups are discussed, but in the remainder of the thesis attention is restricted to the two group case. Despite the plethora of new classification techniques that are appearing in the literature, linear discriminant analysis and logistic regression remain two of the most frequently used methods in this area. This is confirmed by the wide availability of software for implementing these techniques. Notwithstanding its popularity, there are still a number

of important problems regarding linear discriminant analysis and logistic regression that have not been resolved satisfactorily. Gnanadesikan et al. (1989) provide examples. Important amongst these problems are selecting a subset of the available feature variables for use in a classification function, and estimating the actual error rate of the classification function formed in this way, thereby obtaining a measure of the accuracy with which this function will classify entities of unknown origin. Investigation of variable selection in discriminant analysis and logistic regression, and subsequent estimation of the associated post-selection actual error rate, are therefore two of the main focus points of the thesis.

Before conducting an investigation into these aspects, however, Chapter 2 of the thesis is devoted to a comparison of the classification performance of linear discriminant analysis and logistic regression. The intention in Chapter 2 is to provide at least a partial answer to a question that may easily arise in practice, viz. which of these popular techniques should be used in a specific problem? In general the results of the simulation study described in Chapter 2 seem to indicate that linear discriminant analysis frequently offers more accurate classification than logistic regression, even in cases that are often regarded as non-ideal for linear discriminant analysis, viz. cases where the feature variables are not normally distributed. It also becomes clear that logistic regression suffers from a disadvantage that may not be appreciated sufficiently, viz. non-convergence of the iterative procedure that must be used to estimate the parameters in the logistic regression function in cases where the populations are well separated. In view of the findings in Chapter 2, the main emphasis in the remainder of the thesis is on linear discriminant analysis, although logistic regression is included in the discussion of variable selection in Chapter 4.

A number of aspects related to variable selection in linear discriminant analysis are discussed in Chapter 3. The first aspect that receives attention is the influence of the number of variables in a linear discriminant function on its classification performance, as reflected in its actual error rate. In this part of the study the variables in the linear discriminant function are varied in a pre-specified manner, i.e. no variable selection based on the sample data takes place. An interesting and important fact that comes to light is that a variable with respect to which the two populations do not differ, can significantly improve the classification performance of a linear discriminant function, provided that it is highly correlated with one or more of the variables with respect to which the two populations do differ. It is therefore important in variable selection that variables should not be considered singly (one at a time), but that a multivariate approach should be followed. When selecting variables for inclusion into a linear discriminant function, different selection criteria can be used. These criteria can be divided into two broad categories, viz. *separatory* and *allocatory*. The first category consists of criteria such as the squared multiple correlation coefficient (R^2), Mallows' C_p and Wilks' Λ , while the second consists of criteria based on error rate estimators. The second part of Chapter 3 contains a comparison of two separatory and three allocatory criteria. This comparison is in terms of the error rates of the resulting linear discriminant functions when the criteria are required to select a pre-specified number of variables.

The conclusions emanating from the study in Chapter 3 are applied in Chapter 4 in the development of a new variable selection technique. This technique is based on the concept of *cross model validation*, introduced by Hjorth (1994) in a regression context. An important advantage of cross model validation is that it provides an accurate estimate of the post-selection actual error rate of the classification function based on the selected variables. Chapter 4 therefore also contains the results of an investigation into the problem of assessing the accuracy of a classification function based on selected variables. Cross model validation can also be applied for variable selection and subsequent error rate estimation in a wider context, and in Chapter 4 this is done for logistic regression in addition to linear discriminant analysis. The chapter closes with two examples illustrating application of the cross model validation procedure.

It is probably true that the problem of variable selection has received most attention in a regression analysis context (cf. Miller, 1990), but it is also relevant in many other areas of statistics (cf. Linhart and Zucchini, 1986, for a general discussion of model or variable selection). Many of the variable selection techniques that are developed for use in specific applications can also be applied successfully in other areas. Chapter 5 provides an illustration. In this chapter it is shown how a variable selection technique based on pre-testing can be modified for use in linear discriminant analysis. The thesis closes in Chapter 6 with conclusions and recommendations.

1.3 THE NUMERICALLY INTENSIVE APPROACH

The fairly recent advent of powerful computers has had a marked influence on the theory and practice of statistics. A consequence of the growing availability of computing power is that problems that were formerly considered intractable are nowadays studied, and in many cases solved, by means of computer intensive methods. There are many modern statistical techniques that owe their prominence to the availability of powerful computers. Examples that come to mind are the bootstrap, Markov chain Monte Carlo methods and a variety of simulation methods. Development of new techniques in this area is currently an active field of research.

The post-selection properties of sample statistics, especially in a multivariate setting, is a prime example of a class of problems that is too complicated for an analytical approach, and that has to be addressed numerically. This is mainly the result of the fact that application of a selection criterion is in effect equivalent to a very complex partitioning of the sample space, making the analytical calculation of probabilities and expectations very difficult and in many cases impossible. Analytical contributions to this area have therefore mainly dealt with fairly simple special cases, and have largely been restricted to asymptotic results. In this thesis the focus is on the important cases of small and medium samples. The problems that are addressed are not amenable to analytical arguments, and a numerically intensive approach is therefore essential. Consequently, simulation methods are extensively used in the thesis.

1.4 MAIN CONTRIBUTION

The main contribution of this thesis lies in the practically useful new techniques that are introduced for variable selection and post-selection error rate estimation in discriminant analysis and logistic regression. It is felt that the cross model validation techniques described in Chapter 4 are especially noteworthy in this regard. The thesis contains no theorems establishing optimality properties of the new techniques, since this seems to be impossible owing to the complicated nature of these methods. However, the results of an extensive simulation study reported in the thesis, provide substantial evidence that the proposed techniques perform well.

The programs that are provided in the thesis can also be viewed as a further contribution. These programs were used in the simulation study, but they can easily be adapted for the analysis of a given single data set. It would also be easy to translate such a program into a more readily available language such as S-Plus or SAS. This would then be a valuable facility for the data analyst confronted with the problem of variable selection and post-selection error rate estimation.

CHAPTER 2

A COMPARISON OF THE CLASSIFICATION PERFORMANCE OF LINEAR DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

2.1 INTRODUCTION : DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

Consider the problem of classifying an entity of unknown origin into one of $G + 1$ qualitatively distinct groups, denoted by $\Pi_0, \Pi_1, \dots, \Pi_G$, on the basis of a vector \mathbf{x} of measurements on k feature variables. This is an important problem in many fields, e.g. classification of a patient into one of a number of categories reflecting the severity of a certain disease. In some applications there is an element of prediction involved, e.g. predicting corporate failure based on measurements of financial variables or assessing the likelihood of a student successfully completing a course based on a battery of test scores. There are a number of techniques that can be used in this context, of which discriminant analysis and logistic regression are popular choices that are frequently employed. The aim in this chapter is to evaluate the relative merit of these two techniques when applied for classification purposes under various circumstances.

The Bayesian paradigm provides a convenient framework for constructing classification rules. Introduce a random variable Y that indicates group membership, i.e. $Y = j$ in group Π_j , $j = 0, 1, \dots, G$. Let $\pi_0, \pi_1, \dots, \pi_G$ be the prior probabilities of groups $\Pi_0, \Pi_1, \dots, \Pi_G$ respectively, i.e. $\pi_j = P(Y = j)$, $j = 0, 1, \dots, G$, with $\sum_{i=0}^G \pi_i = 1$.

Denote an entity with observed feature vector \mathbf{x} by $e(\mathbf{x})$. Classification of an entity $e(\mathbf{x})$ of unknown origin into one of the $G + 1$ groups can be done by considering the *posterior probabilities* of group membership, given by

$$\tau_i(\mathbf{x}) = P(Y = i | \mathbf{x}), \quad i = 0, 1, \dots, G, \quad (2.1.1)$$

and allocating $e(\mathbf{x})$ to group j , where $\tau_j(\mathbf{x}) = \max\{\tau_i(\mathbf{x}), i = 0, 1, \dots, G\}$. This leads to the classification rule:

$$C(\mathbf{x}) = j \text{ if } \tau_j(\mathbf{x}) = \max\{\tau_i(\mathbf{x}), i = 0, 1, \dots, G\}. \quad (2.1.2)$$

The rule specified in (2.1.2) is the Bayes classification rule. It maximises the posterior probability of group membership, and is optimal in the sense that it minimises the probability of misclassification.

The Bayes rule can be formulated in terms of the probability density function of the random feature vector \mathbf{X} . Let $f_i(\cdot)$ be the probability density function of \mathbf{X} in group Π_i , with corresponding cumulative distribution function $F_i(\cdot)$, $i = 0, 1, \dots, G$. Then the posterior probability that an entity with feature vector \mathbf{x} belongs to group Π_j is given by

$$\tau_j(\mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\sum_{i=0}^G \pi_i f_i(\mathbf{x})} = \frac{\pi_j f_j(\mathbf{x})}{f(\mathbf{x})}, \quad j = 0, 1, \dots, G, \quad (2.1.3)$$

where $f(\mathbf{x}) = \sum_{i=0}^G \pi_i f_i(\mathbf{x})$ is a mixture of the group conditional probability density functions. Denote the cumulative distribution function corresponding to $f(\cdot)$ by $F(\cdot)$.

Clearly, (2.1.2) can also be formulated as

$$C(\mathbf{x}) = j \text{ if } \pi_j f_j(\mathbf{x}) = \max\{\pi_i f_i(\mathbf{x}), i = 0, 1, \dots, G\}. \quad (2.1.4)$$

Instead of using (2.1.2) or (2.1.4) directly, it is often more convenient to work with the logarithm of the ratios of the posterior probabilities. Without loss of generality, choose Π_0 as reference population, and consider the logarithms of the ratios:

$$\zeta_{i0}(\mathbf{x}) = \log\{\tau_i(\mathbf{x})/\tau_0(\mathbf{x})\} = \log(\pi_i/\pi_0) + \log\{f_i(\mathbf{x})/f_0(\mathbf{x})\}, \quad i = 1, \dots, G.$$

The classification rules (2.1.2) and (2.1.4) have the following equivalent in terms of these log ratios:

$$C(\mathbf{x}) = \begin{cases} 0 & \text{if } \zeta_{i0}(\mathbf{x}) \leq 0 \\ j & \text{if } \zeta_{i0}(\mathbf{x}) \leq \zeta_{j0}(\mathbf{x}), \end{cases} \quad i, j = 1, \dots, G; i \neq j. \quad (2.1.5)$$

If the prior probabilities $\pi_0, \pi_1, \dots, \pi_G$ and the group conditional probability density functions were known, classification of an entity with feature vector \mathbf{x} could be based on the exact values of the $\zeta_{i0}(\mathbf{x})$, $i = 1, \dots, G$. In practice the prior probabilities are often unknown and therefore have to be estimated. The group conditional probability density functions are also often completely unknown, or the functional form may be known, but some parameters may be unknown. In order to estimate these unknown parameters and/or density functions, it is assumed that data are available on entities with known group membership, the so called *training data*. The training data consist of measurements of the k feature variables on each of n entities. Denote the training data set by \mathbf{t} , which is a $n \times (k+1)$ matrix with rows equal to (\mathbf{x}'_j, y_j) , $j = 1, \dots, n$.

Here \mathbf{x}'_j denotes the transpose of the column vector \mathbf{x}_j . A classification rule based on the training data will be denoted by $C(\mathbf{x}; \mathbf{t})$.

The training data are obtained either by sampling from a mixture of the $G + 1$ groups, or by sampling from each group separately. In the case of mixed sampling, yielding a training data set of random size n_i from Π_i , $i = 0, 1, \dots, G$, the prior probabilities are usually estimated by $\hat{\pi}_i = n_i/n$. In the case of separate sampling, yielding samples of fixed sizes from each $F_i(\cdot)$, estimates of the prior probabilities cannot be obtained in this way. To obtain estimates in this case, a random sample of size m from the mixture of the $G + 1$ groups has to be available. If the group membership of the entities in this sample is unknown, the entities are classified using a rule based on the training data and assuming equal prior probabilities. If m_i is the number of entities assigned to group Π_i , then the proportion m_i/m is used as an estimate of the prior probability of group Π_i , $i = 0, 1, \dots, G$. These estimates are biased, and methods exist for bias correction (cf. McLachlan, 1992, p.31). Other methods of estimating the prior probabilities are also discussed by McLachlan (1992).

A number of different approaches to classification using (2.1.5) exist, depending on the degree to which parametric assumptions regarding the group conditional densities are made. Firstly, in a fully parametric approach, it is assumed that the group conditional density functions are known, although some parameters may have to be estimated from the training data. Many assumptions regarding the functional form of the densities are of course possible. The most common assumption is that of a homoscedastic normal model, when the probability density function in each of the groups is given by

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)' \Sigma^{-1}(\mathbf{x} - \mu_i)\right\}, \quad i = 0, 1, \dots, G,$$

where $\mu_0, \mu_1, \dots, \mu_G$ are the group mean vectors in $\Pi_0, \Pi_1, \dots, \Pi_G$ respectively, and Σ is the common covariance matrix. In this case, the log ratios of the posterior probabilities are given by:

$$\begin{aligned} \zeta_{i0}(\mathbf{x}) &= \log\{\tau_i(\mathbf{x})/\tau_0(\mathbf{x})\} \\ &= \log(\pi_i/\pi_0) + \left\{\mathbf{x} - \frac{1}{2}(\mu_i + \mu_0)\right\}' \Sigma^{-1}(\mu_i - \mu_0), \quad i = 1, \dots, G. \end{aligned} \tag{2.1.6}$$

If classification is based on (2.1.6), the *normal linear discriminant rule* is obtained. This rule is seldom of any practical use, since it contains parameters of which the values are unknown. The sample equivalent of this rule is obtained by replacing the parameters $\mu_0, \mu_1, \dots, \mu_G$ and Σ with their customary estimators, the sample means $\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_G$ and the pooled covariance matrix S , obtained from the training data.

For the case of 2 groups ($G=1$), the rule (2.1.5) is given by

$$C(\mathbf{x}) = \begin{cases} 0 & \text{if } \zeta_{10}(\mathbf{x}) \leq 0 \\ 1 & \text{if } \zeta_{10}(\mathbf{x}) > 0 \end{cases}$$

and in the normal case this is equivalent to

$$C(\mathbf{x}) = \begin{cases} 0 & \text{if } \{\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)\}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \leq c \\ 1 & \text{if } \{\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)\}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) > c, \end{cases}$$

where $c = \log(\pi_0/\pi_1)$.

The function $\{\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)\}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ is called the normal linear discriminant function. The sample equivalent of this function is the widely used *Anderson classification statistic* for two group discrimination,

$$W(\mathbf{x}; t) = \{\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_0)\}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0), \quad (2.1.7)$$

cf. Anderson (1951).

For normal populations with common covariance matrix, the normal linear discriminant rule minimises the expected probability of misclassification (cf. Gnanadesikan et al., 1989). The simplicity and general availability of this rule have led to its widespread use when the assumptions of normality and equal covariance matrices are not met, often without proper regard for the robustness of the procedure.

A second approach to classification using (2.1.5) is provided by *logistic regression*. This approach is only partially parametric, as no assumptions regarding the precise functional form of the group conditional probability density functions $f_i(\mathbf{x})$, $i = 0, 1, \dots, G$, are made, but it is assumed that the logarithms of the ratios of the probability density functions are linear functions of \mathbf{x} , i.e.

$$\log\{f_i(\mathbf{x})/f_0(\mathbf{x})\} = \tilde{\beta}_{0i} + \beta_{1i}'\mathbf{x}, \quad i = 1, \dots, G.$$

For this model the log ratios of the posterior probabilities are given by

$$\begin{aligned} \zeta_{i0}(\mathbf{x}) &= \log\{\tau_i(\mathbf{x})/\tau_0(\mathbf{x})\} = \log(\pi_i/\pi_0) + \tilde{\beta}_{0i} + \beta_{1i}'\mathbf{x} \\ &= \beta_{0i} + \beta_{1i}'\mathbf{x}, \quad i = 1, \dots, G. \end{aligned} \quad (2.1.8)$$

The parameters β_{0i} and β_{1i} ($i = 1, \dots, G$) have to be estimated from the training data, usually by means of maximum likelihood estimation. Two other estimation methods



that are seldomly used are noniterative weighted least squares estimation and discriminant function analysis (cf. Hosmer and Lemeshow, 1989, p.18). The former method was proposed by Grizzle et al. (1969), and it consists of one iteration of the iteratively reweighted least squares algorithm that is used to calculate maximum likelihood estimates of the parameters. Estimation of the parameters in (2.1.8) by means of discriminant function analysis is accomplished by assuming that the random feature vector \mathbf{X} is normally distributed, with mean vector μ_j and covariance matrix Σ in Π_j , $j=0,1,\dots,G$. Then the parameters in (2.1.8) can be expressed in terms of $\mu_0, \mu_1, \dots, \mu_G$ and Σ , and by substituting estimates of $\mu_0, \mu_1, \dots, \mu_G$ and Σ into these expressions, estimates are obtained for β_{0i} and β_{1i} , $i=1,\dots,G$.

The investigation in this thesis will be restricted to the case that occurs most commonly in practice, viz. where the parameters in (2.1.8) are estimated by means of maximum likelihood. If β_{0i} and β_{1i} , $i=1,\dots,G$, in (2.1.8) are replaced by their maximum likelihood estimates, the logistic discriminant rule is obtained.

A third approach to the discrimination problem is a fully non-parametric approach, where no assumptions regarding the group conditional distributions are made. This includes methods where non-parametric density estimation is used and tree structured rules such as CART (cf. Breiman et al., 1984). Non-parametric discrimination will not be considered in this thesis. A comprehensive review of this topic is given by McLachlan (1992, Chapter 9).

Finally, it should be mentioned that a Bayesian approach to discriminant analysis can also be employed. McLachlan (1992, p. 29-31) and Geisser (1964, 1966 and 1982) are references in this regard. The Bayesian approach typically entails finding the posterior density function of the parameters given the training data t , based on a prior density function for the parameters. This posterior density is then used as a weighting factor to calculate the predictive density of a feature vector \mathbf{X} within each of the groups. The predictive densities are then used in (2.1.3) to calculate predictive estimates of the posterior probabilities, which can be used in (2.1.2) to classify the entity of unknown origin.

In this chapter a comparative study of the classification performance of the normal linear discriminant rule and the logistic discriminant rule when all the available feature variables are used to construct these rules, will be discussed. In Chapters 3 and 4 the discussion will be extended to include problems surrounding variable selection. The situation where only a subset of the available feature variables are selected for inclusion when forming the classification rule, will be considered.

In Section 2.2, the different error rates that are used to quantify the classification performance of a discriminant function, are defined. A comprehensive overview of error rate estimators is also given. In Section 2.3, the literature in which linear discriminant analysis is compared to logistic regression, is reviewed. The Monte Carlo simulation study in which the performance of these two techniques is compared in the

case of two groups, receives attention in Section 2.4, while the results of a similar study for the three group case, are reported in Section 2.5. In Section 2.6, two approaches for estimating the coefficients of the logistic discriminant function in the case of more than two groups, are compared. The chapter closes in Section 2.7 with a number of conclusions and recommendations.

2.2 ERROR RATES

In order to compare the classification performance of normal linear discriminant analysis and logistic regression, a criterion has to be chosen to assess the probability of misclassifying entities. Various error rates can be defined to quantify the performance of a classification rule, e.g. the *optimal error rate*, the conditional or *actual error rate* and the *unconditional error rate*. In this section the different error rates are defined and error rate estimators are briefly reviewed.

The optimal error rates associated with a classification rule are defined as the probability that a randomly chosen entity from population Π_i is allocated to population Π_j , assuming the relevant parameters of the distributions of the feature vectors to be known:

$$\text{eropt}_{ij}(F) = P(C(\mathbf{X}; F) = j | Y = i), \quad i, j = 0, 1, \dots, G; i \neq j. \quad (2.2.1)$$

In (2.2.1), $F(\mathbf{x}) = \sum_{i=0}^G \pi_i F_i(\mathbf{x})$ is a mixture of the group conditional distribution functions, and $C(\mathbf{X}; F)$ denotes a classification function.

The optimal error rate for group i is given by

$$\text{eropt}_i(F) = \sum_{j \neq i=0}^G \text{eropt}_{ij}(F), \quad i = 0, 1, \dots, G, \quad (2.2.2)$$

and the overall optimal error rate by

$$\text{eropt}(F) = \sum_{i=0}^G \pi_i \text{eropt}_i(F). \quad (2.2.3)$$

To calculate the optimal error rates, the functional form and all the parameters of F have to be known. In the case of multivariate normal populations with means $\mu_0, \mu_1, \dots, \mu_G$ and common covariance matrix Σ , an explicit expression can be obtained for the optimal error rates associated with the normal linear discriminant rule. In the case of two groups ($G = 1$) this expression is given by

$$\text{eropt}_i(c, \Delta) = \Phi\{[(-1)^{i+1}c - \frac{1}{2}\Delta^2]/\Delta\} \quad (2.2.4)$$

where Φ is the standard normal distribution function and Δ^2 is the squared Mahalanobis distance between the two populations, viz.

$$\Delta^2 = (\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0). \quad (2.2.5)$$

The conditional or actual error rates are obtained by calculating the misclassification probabilities conditional on the training data, i.e.

$$\text{eract}_{ij}(F_i; t) = P(C(X; t) = j \mid Y = i, t), \quad i, j = 0, 1, \dots, G; i \neq j. \quad (2.2.6)$$

This is the probability, conditional on the training data, that an entity from group Π_i with random feature vector X , is wrongly classified into group Π_j , $j \neq i$. The actual error rate for group Π_i is given by

$$\text{eract}_i(F_i; t) = \sum_{j \neq i}^G \text{eract}_{ij}(F_i; t), \quad i = 0, 1, \dots, G \quad (2.2.7)$$

and the overall actual error rate by

$$\text{eract}(F; t) = \sum_{i=0}^G \pi_i \text{eract}_i(F_i; t). \quad (2.2.8)$$

In the case of multivariate normal populations with means $\mu_0, \mu_1, \dots, \mu_G$ and common covariance matrix Σ , explicit expressions for the actual error rates associated with the normal linear discriminant rule can once more be obtained. For two groups, this expression is

$$\text{eract}_i(c, \mu_i, \Sigma; t) = \Phi \left[\frac{(-1)^{i+1} [c - \{\mu_i - \frac{1}{2}(\bar{x}_1 + \bar{x}_0)\}]' S^{-1} (\bar{x}_1 - \bar{x}_0)}{[(\bar{x}_1 - \bar{x}_0)' S^{-1} \Sigma S^{-1} (\bar{x}_1 - \bar{x}_0)]^{\frac{1}{2}}} \right], \quad i = 0, 1, \quad (2.2.9)$$

where \bar{x}_0, \bar{x}_1 are the means of the samples taken from Π_0 and Π_1 , respectively, and S is the pooled sample covariance matrix.

In practice the actual error rate is relevant, since it is the error rate corresponding to the classification rule that has been formed from the available training data. In the later comparison of the classification performance of discriminant analysis and logistic regression, actual error rate will be used as criterion of classification performance.

The expected or unconditional error rates are obtained by averaging the conditional error rates over the distribution of the training data. For example,

$$\text{erunc}_{ij}(F_i) = E[\text{eract}_{ij}(F_i; T)], \quad i, j = 0, 1, \dots, G; i \neq j \quad (2.2.10)$$

are the unconditional error rates corresponding to the actual error rates in (2.2.6). Similar expressions define the unconditional error rates corresponding to (2.2.7) and (2.2.8).

The error rates defined above are functions of the unknown distribution parameters and can therefore not be calculated. In practice, these error rates have to be estimated from the sample data. A number of error rate estimators have been defined for the actual error rate and can be broadly grouped into three categories: parametric estimators, nonparametric estimators and smoothed estimators. Some of these error rate estimators will now be discussed briefly.

Firstly, some *parametric error rate estimators*, based on the assumption of a homoscedastic normal model will be discussed. The two group case, with equal prior probabilities, resulting in $c = \log(\pi_0/\pi_1) = 0$, will be considered.

The plug-in principle provides a mechanism for constructing parametric error rate estimators. It entails replacing the unknown parameters in a parametric expression for the error rate by suitable estimators of these parameters. The simplest example is the so-called D-estimator of the actual error rate, originally defined by Fisher (1936). This estimator is obtained by replacing the parameters μ_0, μ_1 and Σ in (2.2.9) with their unbiased estimators \bar{X}_0, \bar{X}_1 and S , obtained from the training data. This yields the estimator $\Phi(-D/2)$, where D^2 is the estimated squared Mahalanobis distance, given by

$$D^2 = (\bar{x}_1 - \bar{x}_0)' S^{-1} (\bar{x}_1 - \bar{x}_0). \quad (2.2.11)$$

As indicated by Lachenbruch and Mickey (1968), this estimator is optimistically biased. Several suggestions have been made for reducing this bias, and some of these are now briefly reviewed.

The shrunken D-estimator (also referred to as the DS-estimator) is obtained in a similar way to the D-estimator, but using $\hat{\Sigma} = (n-2)S/(n-k-3)$ as estimator for Σ instead of S . This estimator is of course only defined for $n > k+3$, and is given by $\Phi(-\frac{1}{2}D\sqrt{(n-k-3)/(n-2)})$. This estimator will always be larger than the D-estimator (since $(n-k-3)/(n-2) < 1$ for any value of k), thus correcting for the optimistic bias of the D-estimator.

Lachenbruch (1968) suggested correcting the above bias by replacing D^2 with the unbiased estimator of Δ^2

$$\hat{\Delta}^2 = \left\{ (n-k-1)/n \right\} D^2 - k(1/n_0 + 1/n_1).$$

A number of asymptotic approaches have also been suggested. McLachlan (1973, 1974, 1975) derived expressions for the asymptotic bias of the plug-in estimator and used these to obtain bias corrected versions of the D-estimator. Lachenbruch and Mickey (1968) used a second order asymptotic expansion of the actual error rate to derive another estimator.

The normal based linear discriminant rule is known to be fairly robust with respect to departures from normality. The same is not true for the error rate estimators based on the normality assumption, and the performance of these estimators deteriorates in non-normal cases (cf. Snapinn and Knoke, 1984 and Konishi and Honda, 1990). Furthermore, the parametric error rate estimators discussed here are estimators of the error rate of the linear discriminant rule, and are therefore not suitable to estimate the error rate of any other discriminant rule, e.g. the logistic discriminant rule.

Nonparametric error rate estimators are not based on any distributional assumptions and can therefore be expected to be more robust than parametric estimators. These estimators can also be used to estimate the error rate of any classification rule, and can therefore be employed for estimation of the error rate of the linear discriminant rule as well as the logistic discriminant rule.

For ease of notation, let $t_i = (\mathbf{x}_{ij}, y_{ij})$, $i = 0, 1, \dots, G$; $j = 1, \dots, n_i$, denote the training data from group Π_i , and let $t = \bigcup_{i=0}^G t_i$ as before denote the entire training data set.

The simplest example of a nonparametric error rate estimator is the apparent error rate (or resubstitution error rate) which was suggested by Smith (1947). It is defined as the proportion of the training data that is misclassified by the classification rule. Consider the classification rule based on the training data set t :

$$C(\mathbf{x}; t) = i \text{ if } \mathbf{e}(\mathbf{x}) \text{ is allocated to group } \Pi_i, i = 0, 1, \dots, G.$$

The apparent error rate of group Π_i is

$$A_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} I[C(\mathbf{x}_{ij}; t) \neq i], \quad i = 0, 1, \dots, G, \quad (2.2.12)$$

where $I[\cdot]$ denotes the indicator function.

The overall apparent error rate is given by

$$A(t) = \frac{1}{n} \sum_{i=0}^G n_i A_i(t). \quad (2.2.13)$$

Because the apparent error rate is calculated by applying the classification rule to the same data from which it was formed, it is optimistically biased (cf. Efron, 1986). The apparent error rate also has a very large variance, which further contributes to its unsuitability as error rate estimator (cf. Glick, 1978).

Several error rate estimators have been developed with the aim of reducing the bias of the apparent error rate. Lachenbruch and Mickey (1968) proposed the leave-one-out estimator. Each case is in turn removed from the training data, and a classification rule based on the remaining data is calculated. This classification rule is then used to classify the 'holdout' observation. The proportion of misclassifications obtained in this way is used to estimate the error rate. To give a formal definition, let $t_{(j)} : [(n-1) \times (k+1)]$ be the training data from which the j -th case, x_j , was deleted. The classification rule based on $t_{(j)}$ is denoted by

$$C_{(j)}(x; t_{(j)}) = i \text{ if } x \text{ is allocated to group } \Pi_i, \quad i = 0, 1, \dots, G.$$

The leave-one-out error rate for group Π_i is given by

$$L_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} I[C_{(j)}(x_{ij}; t_{(j)}) \neq i], \quad i = 0, 1, \dots, G. \quad (2.2.14)$$

The overall leave-one-out error rate is defined as

$$L(t) = \frac{1}{n} \sum_{i=0}^G n_i L_i(t). \quad (2.2.15)$$

Although the leave-one-out error rate has a greatly reduced bias, it has a very large variance, which, according to Glick (1978), 'overwhelms the magnitude of this method's bias reduction'. Based on Monte Carlo simulation studies comparing several error rate estimators, Efron (1983) commented that the leave-one-out method gives a nearly unbiased estimator, 'but often with unacceptably high variability, particularly if n is small'.

McLachlan (1976b) derived the asymptotic bias of the apparent error rate for two multivariate populations with a common covariance matrix, and used this to find a correction term that can be used to reduce the bias.

Efron (1983) applied bootstrap methodology to find an error rate estimator that is less biased than the apparent error rate. The bias of the apparent error rate is estimated by means of resampling methods and the bootstrap estimator is calculated by correcting

the apparent error rate for bias. The bias correction for group Π_i is calculated as follows.

In a separate sampling situation a bootstrap sample $\mathbf{t}_i^* = (\mathbf{x}_{ij}^*, y_{ij}^*)$, $j = 1, \dots, n_i$, of fixed size n_i (where n_i is the size of the training sample obtained from Π_i) is generated by sampling with replacement from \hat{F}_i , the empirical distribution function of \mathbf{x} in \mathbf{t}_i , $i = 0, 1, \dots, G$. The $G+1$ bootstrap samples are then combined to form the bootstrap sample \mathbf{t}^* , i.e. $\mathbf{t}^* = \bigcup_{i=0}^G \mathbf{t}_i^*$. In a mixed sampling situation, a bootstrap sample \mathbf{t}^* of size

n is obtained by sampling with replacement from \hat{F} , the empirical distribution function of \mathbf{x} in \mathbf{t} . In this situation, n_i (the size of the sample $\mathbf{t}_i^* = (\mathbf{x}_{ij}^*, y_{ij}^*)$, $j = 1, \dots, n_i$, obtained in this way) is random.

A classification rule $C^*(\mathbf{x}; \mathbf{t}^*)$ is formed based on the bootstrap sample \mathbf{t}^* , in the same way in which $C(\mathbf{x}; \mathbf{t})$ was formed from \mathbf{t} . The apparent error rate of $C^*(\mathbf{x}; \mathbf{t}^*)$ is then calculated for group Π_i :

$$A_i^*(\mathbf{t}^*) = \frac{1}{n_i} \sum_{j=1}^{n_i} I[C^*(\mathbf{x}_{ij}^*; \mathbf{t}^*) \neq i], \quad i = 0, 1, \dots, G. \quad (2.2.16)$$

The proportion of observations in the training data \mathbf{t}_i misclassified by $C^*(\mathbf{x}; \mathbf{t}^*)$ is also calculated:

$$A_i^{**}(\mathbf{t}) = \frac{1}{n_i} \sum_{j=1}^{n_i} I[C^*(\mathbf{x}_{ij}; \mathbf{t}^*) \neq i], \quad i = 0, 1, \dots, G. \quad (2.2.17)$$

For each group the difference $d_i = A_i^* - A_i^{**}$, $i = 0, 1, \dots, G$, is obtained.

The procedure described above is repeated a large number (say B) of times, giving the differences d_{it} , $i = 0, 1, \dots, G$; $t = 1, \dots, B$. The bootstrap estimator of the bias associated with group Π_i is computed by taking the average of the d_{it} :

$$b_i = \frac{1}{B} \sum_{t=1}^B d_{it}, \quad i = 0, 1, \dots, G.$$

The bootstrap corrected error rate for group Π_i is then obtained by adjusting the apparent error rate for bias:

$$E_i(\mathbf{t}) = A_i(\mathbf{t}) - b_i, \quad 0, 1, \dots, G.$$

The overall bootstrap corrected error rate is given by

$$E(\mathbf{t}) = \frac{1}{n} \sum_{i=0}^G n_i E_i(\mathbf{t}) \quad (2.2.18)$$

Whilst it is true that (2.2.18) has a smaller bias than the apparent error rate, the process of bias correction can easily lead to an unacceptably large increase in the variance of the final estimator (cf. Efron and Tibshirani, 1993, p.138).

Efron (1983) also described some variants of the bootstrap method, such as the randomised bootstrap, the double bootstrap and the 0.632 estimator. To calculate the 0.632 estimator, bootstrap samples are taken in a similar way as described above, but at each step the error rate is estimated by classifying only the cases in the training data which are not part of the bootstrap sample on which the classification rule was based. This estimator is referred to as the e_0 - estimator. The weighted average of the e_0 - estimator and the apparent error rate - the former having a weight of 0.632 and the latter a weight of 0.368 - is calculated to obtain the 0.632 error rate estimator. According to Efron (1983), the 0.632 estimator gave the best performance of the error rate estimators included in his simulation studies (leave-one-out error rate, ordinary bootstrap and other bootstrap variants). Chatterjee and Chatterjee (1983) and Chernick, Murthy and Nealy (1985, 1986a) investigated the use of the e_0 - estimator, but the 0.632 estimator performed better.

Estimators belonging to the final category, viz. *smoothed error rate estimators*, have been developed in an attempt to reduce the variance of the apparent error rate. One way of smoothing the apparent error rate in the case of two groups, is to base an estimator on the estimated posterior probabilities of group membership of the entities in the training data, $\hat{\tau}_i(\mathbf{x}_j; \mathbf{t})$, $j=1, \dots, n$; $i=0,1$. The posterior probability error rate estimator is defined by

$$A_{\text{post}}(\mathbf{t}) = \frac{1}{n} \sum_{j=1}^n \min_i \hat{\tau}_i(\mathbf{x}_j; \mathbf{t}) \quad (2.2.19)$$

Glick (1978) suggested that the large variance of the apparent error rate may be a greater problem than its bias. He therefore proposed a class of smoothed error rate estimators for the univariate case, with the purpose of reducing the variance of the apparent error rate. Snapinn and Knoke (1985) extended these ideas to the multivariate case. For the case of two groups ($G=1$), they suggested a class of normally smoothed error rate estimators, which is defined for cases from group Π_0 by

$$E_0^s(\mathbf{t}) = \frac{1}{n_0} \sum_{j=1}^{n_0} g(\mathbf{x}_{0j}; \mathbf{b}) \quad (2.2.20)$$

with $g(\mathbf{x}; b) = \Phi[\{c - W(\mathbf{x})\} / (bD)]$, where W is the Anderson classification statistic given in (2.1.7) and b is a smoothing constant.

Snapinn and Knoke (1985, 1988) suggested two specific normally smoothed error rate estimators, denoted by NS and NS* respectively, and compared their performance to that of other error rate estimators by means of simulation studies.

For the NS - method the smoothing constant is given by

$$b = \left\{ \frac{[(k+2)(n_0 - 1) + n_1 - 1]}{[n_0(n_0 + n_1 - k - 3)]} \right\}^{\frac{1}{2}} \quad (2.2.21)$$

and for the NS* - method

$$b = \begin{cases} \left\{ \left[\frac{D^2}{c_1 D^2 - c_2} \right] - \left[\frac{(n_0 - 1)}{(n_0)} \right] \right\}^{\frac{1}{2}}, & \text{if } D^2 > c_2 / c_1 \text{ and } n_0 + n_1 > k + 3 \\ \infty, & \text{otherwise,} \end{cases} \quad (2.2.22)$$

where $c_1 = (n_0 + n_1 - k - 3) / (n_0 + n_1 - 2)$ and $c_2 = k(n_0 + n_1) / (n_0 n_1)$.

Details of the derivation of these constants are given in Snapinn and Knoke (1988). For misclassification of a case from Π_1 the estimated error rate $E_1^S(t)$ is defined similarly, and the estimated overall error rate is obtained by calculating the weighted average of the two group specific estimates:

$$E^S(t) = \{n_0 E_0^S(t) + n_1 E_1^S(t)\} / (n_0 + n_1). \quad (2.2.23)$$

Snapinn and Knoke (1988) suggested that the NS*-estimator should be the error rate estimator of choice if the parent distributions are nearly normal. They also mentioned that this estimator is very non-robust, being the worst of all estimators considered in the case of univariate exponential parent distributions.

The normally smoothed estimators described above, were developed in an attempt to reduce the variance of the apparent error rate. In order to achieve bias reduction, Snapinn and Knoke (1988) proposed that the bootstrap and .632 bootstrap methods of Efron (1983) be applied to the NS-estimator, to give the B(NS)-estimator and the B.632(NS)-estimator respectively. In simulation studies of a five-variate normal distribution Snapinn and Knoke (1988) found that these estimators were less biased but had greater variance than the NS-estimator. These estimators do however have a lower unconditional mean squared error than the NS-estimator. The unconditional

mean squared errors of the B(NS)- and B.632(NS)-estimators are also less than that of the estimators calculated by applying the ordinary bootstrap and the .632 bootstrap to the apparent error rate as described by Efron (1983). Snapinn and Knoke (1988) also concluded that the B.632(NS)-estimator generally performed better in their simulation studies than the B(NS)-estimator. For situations where near normality cannot be assumed, they recommended that the NS-estimator should be preferred in the univariate case ($k = 1$). For $k > 1$, the method of choice should be the B.632(NS)-estimator. If $k > 5$, the NS*-estimator may be used if the computational burden of applying the .632 bootstrap method is a concern.

Another method which uses smoothing in conjunction with the bootstrap, was proposed by Sánchez and Cepeda (1989). They suggested smoothing the ordinary as well as Bayesian bootstrap estimators, in an attempt to reduce their variances. To smooth the ordinary bootstrap error rate, a nonparametric kernel estimator of the distribution F was used instead of the empirical distribution used in application of the ordinary bootstrap. Based on a simulation study, they concluded that smoothing improved the performance of the ordinary bootstrap and Bayesian bootstrap error rate estimators, as indicated by a reduction in mean squared error.

A considerable number of papers reviewing and comparing various error rate estimators, have been published. Some of these will be discussed briefly.

Lachenbruch and Mickey (1968) compared parametric estimators to the resubstitution estimator and the "holdout" estimator, and found that the estimators based on the normality assumption outperformed the two nonparametric estimators for normal data. Toussaint (1974) also reported that parametric estimators were superior to nonparametric estimators if normality holds.

McLachlan (1980b) conducted simulation experiments, comparing the bootstrap estimate of the bias of the apparent error rate, to the parametric estimator (cf. McLachlan 1976b) of this bias. Since the means of these estimators were in close agreement for the cases he considered, he defined the efficiency of the bootstrap approach relative to the parametric approach as the ratio of the standard deviations of these estimators. He concluded that the parametric estimator was more efficient for moderately separated bivariate populations ($\Delta = 2$), but for populations that were close together ($\Delta = 1$), the bootstrap estimator was more efficient. The leave-one-out estimator of the bias (defined as the difference between the leave-one-out and apparent error rates) was also included in his study. The standard deviation of this estimator was much larger than that of the other two estimators in all the cases considered, confirming the findings of Glick (1978) and Efron (1983).

Snapinn and Knoke (1984) performed a numerical integration study and a Monte Carlo simulation study to compare the performance of two parametric error rate estimators (viz. the D-estimator and DS-estimator), and two nonparametric error rate estimators (viz. the apparent error rate and the leave-one-out error rate), using the unconditional mean squared errors (UMSE's) of the estimators as criterion. They concluded that

'there is no single error-rate estimator that is best in all situations.' Under the assumption of normality, the parametric estimators performed best when k , the number of feature variables, is small, but are outperformed by the nonparametric estimators for larger values of k and small values of Δ^2 . They also found that the parametric estimators are sensitive to departures from normality. This is confirmed in a study by Konishi and Honda (1990), in which several parametric and nonparametric estimators were compared for a mixture of two multivariate distributions.

Page (1985) evaluated eight parametric error rate estimators (including the D-estimator, the DS-estimator, the L-estimators proposed by Lachenbruch (1967), the M-estimator developed by McLachlan, 1974, and the OS-estimator proposed by Okamoto, 1963) in a Monte Carlo simulation study, considering only the case where the feature variables are normally distributed. For estimation of the actual error rate, the OS-estimator performed best in cases where $k = 4$ and 8 . For $k = 20$, the L-estimator was superior in small sample cases, with the M-estimator best in large sample cases.

Chernick, Murthy and Nealy (1985, 1986a) compared several nonparametric error rate estimators viz. the apparent error rate, the leave-one-out error rate, the ordinary bootstrap, the 0.632 bootstrap, the e_0 -estimator and two other variants of the bootstrap, called the convex bootstrap and the MC-estimator respectively. Their simulation study was done for two and three groups. They studied the case of uncorrelated two and five dimensional normal variables for three different sample sizes and concluded that the 0.632 estimator in general performed best. Chernick, Murthy and Nealy (1986b) investigated the performance of the same error rate estimators for non-normal populations. Data were simulated from Cauchy, uniform and exponential distributions. For the latter two cases, the 0.632 estimator was superior, but for data from the Cauchy distribution, the convex bootstrap and the e_0 -estimator often outperformed the 0.632 estimator.

In contrast to the studies by Chernick, Murthy and Nealy (1985, 1986a), Ganeshanandam and Krzanowski (1990) commented on the 'peculiar' behaviour of the 0.632 estimator. In their simulation study of the multivariate normal case, the 0.632 estimator was found to be the best estimator for small values of Δ , but the worst for large values of Δ . In the case of multivariate binary data, the 0.632 estimator always estimated the error rate in the vicinity of 0.3-0.4, causing it to be a very accurate estimator in some situations, but having large optimistic bias in others. The eleven estimators included in their study were: the apparent error rate, the D-estimator, the OS-estimator proposed by Okamoto (1963), the L-estimator, a bias corrected alternative to the D-estimator suggested by Lachenbruch (1967), the asymptotic unbiased M-estimator derived by McLachlan (1974), the NS-estimator (Snapinn and Knoke, 1985), the leave-one-out estimator (U-estimator) as well as the \bar{U} -estimator, proposed by Lachenbruch and Mickey (1968), the jack-knife (JK) estimator (cf. Efron, 1982; Efron and Gong, 1983) and the 0.632 bootstrap estimator (cf. Efron, 1983). They recommend use of the M, U, \bar{U} , L, JK and OS estimators.

2.3 OVERVIEW OF LITERATURE COMPARING DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

Various authors have compared the efficiency of logistic regression to that of normal discriminant analysis. Aspects with respect to which these comparisons have been done include asymptotic expected error rates, efficiency of estimating the posterior probabilities of group membership, measured by the asymptotic bias and mean squared errors of these estimators, and efficiency of parameter estimation. These comparisons are however mostly for the case of two groups. Very little has been published on the more general situation of more than two groups.

For the two group case, Efron (1975) derived an expression for the asymptotic error rates of the two procedures and also for the relative efficiency of logistic regression compared to normal discrimination in the case of two multivariate normal populations. He concluded that logistic regression is '*between one half and two thirds as effective as normal discrimination for statistically interesting values of the parameters.*'

Press and Wilson (1978) also considered the two group situation and stated that normal discriminant analysis should be the method of choice in the case of multivariate normality, but in cases where one or more of the variables are qualitative - and multivariate normality does not hold - logistic regression should be preferred. Their view was substantiated by means of a theoretical discussion as well as empirical examples in which misclassification rates were considered.

The asymptotic expected error rates of the two procedures in the two group case were compared by McLachlan and Byth (1979) assuming multivariate normality and a common covariance matrix. They derived an expression for the asymptotic expected error rate of logistic regression up to terms of the first order and used the asymptotic expected error rate of normal linear discriminant analysis derived by Okamoto (1963). The values of these asymptotic expected error rates were calculated for different values of Δ^2 , the squared Mahalanobis distance between the two populations, the number of variables and the relative sizes of the two samples. Based on these results, they concluded that the '*performance of the logistic procedure does not fall far short of the normality based method.*' The reason for the apparent contradiction in these findings with Efron's result, is that the first order terms in the asymptotic error rate of logistic regression are approximately two to three times as large as the corresponding terms of the asymptotic error rate for discriminant analysis. For moderate sample sizes, the differences in error rates are very small.

Byth and McLachlan (1980) also compared binary logistic regression to two group normal discrimination for non-normal populations. They considered the asymptotic relative efficiency of logistic regression to normal discriminant analysis on the basis of the asymptotic mean square error of the estimated posterior probability of an observation belonging to a specific group. They studied skewed distributions in which

the degree of skewness was varied, as well as truncated normal distributions, with varying degrees of truncation. In the case of the skewed distributions, they concluded that, when the squared Mahalanobis distance between the populations is small (for example $\Delta^2 = 1$) the logistic regression procedure is more efficient than the normal discriminant procedure, provided that the sample is drawn from a mixture of the two populations in which the more heavily distorted population is at least as prevalent as the less heavily distorted population. When the populations are further apart ($\Delta^2 = 4$ and 9) the efficiency decreases and is in close agreement with the relative efficiency under multivariate normality. In the case of truncated distributions, the logistic regression procedure compared even more favourably to the normal discriminant procedure. The logistic regression procedure is more efficient even in cases where the populations are widely separated ($\Delta^2 = 4$ and 9).

More recently Ruiz-Velasco (1991) calculated the asymptotic efficiency of logistic regression relative to linear discriminant analysis for testing hypotheses about the parameters in the case of two groups and when the explanatory variables are normally distributed explanatory variables. He reported results for the relative efficiency similar to that obtained by Efron (1975) when calculating the asymptotic relative efficiency of the two procedures using misclassification rates.

Bull and Donner (1987) compared two methods that can be used to estimate the parameters of the logistic classification rule in the three group case: maximum likelihood estimation and estimation using discriminant function analysis. They specifically report on the asymptotic relative efficiency of maximum likelihood compared to discriminant function analysis when the feature variables are normally distributed. For the specific cases that they studied, it was found that the asymptotic relative efficiency is significantly affected by factors such as the distance between the populations and correlation between the feature variables.

Rudolpher et al. (1995) describe an extensive simulation study that was undertaken to investigate the classification performance of six techniques in the case of ordinal data from two, three and four groups. The six techniques are: normal discriminant analysis, multinomial logistic regression, ordinal logistic regression, continuation ratio analysis, proportional odds model and the AP classification procedure. In contrast to the findings by Campbell et al. (1991), Rudolpher et al. report that definite benefit is to be gained from using ordinal models when the feature data are ordinal in nature. As far as discriminant analysis and logistic regression are concerned, the differences between their error rates are generally found to be small in the cases considered.

It is clear from the above discussion that contributions in the literature comparing discriminant analysis and logistic regression have focused mainly on the asymptotic performance of these methods and/or the relative efficiency of different methods of estimating the parameters in the logistic classification function. In practice however, the classification performance of the two techniques for small to moderate sample sizes, is frequently the most relevant aspect. Although Rudolpher et al. (1995) investigated the error rates of, amongst others, discriminant analysis and logistic

regression, they restricted attention to cases involving ordinal data. There is therefore a need for a systematic empirical investigation into the error rate performance of the two techniques.

In the remainder of this chapter the two procedures will be compared in the two group situation as well as the three group situation with respect to the expected actual error rates (unconditional error rates). The comparison will be done for normal data, as well as for data from a heavy-tailed symmetrical distribution (the double exponential distribution) and a skewed distribution (the lognormal distribution). The expected actual error rates will be obtained by means of Monte Carlo simulation. An example of the Fortran program used in this regard for the three group case, appears as Program 1 in the Appendix.

2.4 MONTE CARLO SIMULATION STUDY: TWO GROUPS

Consider two groups Π_0 and Π_1 with equal prior probabilities π_0 and π_1 , and an entity $e(\mathbf{x})$ of unknown origin on which k variables x_1, x_2, \dots, x_k have been observed. In linear discriminant analysis the entity $e(\mathbf{x})$ will be classified into group Π_0 if $W(\mathbf{x}; \mathbf{t}) \leq 0$, where $W(\mathbf{x}; \mathbf{t})$ is the Anderson classification statistic given in (2.1.7) with \mathbf{t} the training data set, and into group Π_1 otherwise. If the logistic classification rule is used, the entity is classified into the group with the larger posterior probability (cf. (2.1.8)).

Assuming these classification rules, a Monte Carlo simulation study was done to compare the classification performance of the two techniques. Different underlying distributions for the feature variables x_1, x_2, \dots, x_k were included in the study. Firstly, the case where the feature variables are normally distributed, which satisfies the requirements of normal linear discriminant analysis, was considered. Two further distributions were studied to investigate the effect of non-normality: the double exponential distribution, representing a heavy tailed alternative to the normal distribution, and the lognormal distribution as an example of a skewed distribution. For the purpose of this study, the actual error rates associated with the normal linear discriminant rule and the logistic discriminant rule respectively, were estimated by means of Monte Carlo simulation. To achieve this, a training data set \mathbf{t} from the relevant distribution was generated, and the function $W(\mathbf{x}; \mathbf{t})$ as well as maximum likelihood estimates of the parameters β_{0i} and β_{1i} in (2.1.8) were calculated. The actual error rates conditional on this specific training data set were then estimated by calculating the misclassification rates when both rules were used to classify a large number (5000 per group) of entities generated independently from the same distribution as the training data. This process was repeated 1000 times at each parameter configuration, each time generating a new training data set and estimating its actual error rate in the same way. Finally the unconditional error rate of each of the two techniques at each parameter configuration was obtained by averaging the 1000 actual error rates.

2.4.1 THE NORMAL CASE

In total, twelve cases were investigated. These cases correspond to different specifications of the following factors: the number, k , of feature variables; the covariance structure of these variables; and the sizes of the samples drawn from the two populations. Two values of k were used: $k = 2$ and $k = 10$. With respect to the covariance structure, two choices were made: $\Sigma = \mathbf{I}$, representing independent feature variables with unit variances, and

$$\Sigma = \begin{bmatrix} 1 & \rho & \cdots & \cdots & \rho \\ \rho & \ddots & & & \vdots \\ \vdots & \rho & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix} \quad (2.4.1)$$

in which case the feature variables have unit variances and are equicorrelated. The ρ -values 0.1, 0.5, 0.9 were used, but since the results obtained for these three values are similar, only the results for $\rho = 0.9$ will be reported. Finally, three combinations of sample sizes were used: small sample sizes, viz. $n_0 = n_1 = 25$, mixed samples, viz. $n_0 = 25, n_1 = 50$, and large samples, viz. $n_0 = n_1 = 100$.

For each of the twelve cases identified above, the actual error rates of the two techniques were estimated by simulation at each of the following values of the squared Mahalanobis distance between the two populations: $\Delta^2 = 0, 0.5, 1, 1.5, 2, 3$ and 4. The following parameterisation was used to give these distances: the mean of group Π_0 was chosen as $\mu_0 = \mathbf{0}$, while each of the elements of the mean vector μ_1 was set

equal to $\Delta / \sqrt{\sum_{i=1}^k \sum_{j=1}^k \sigma^{ij}}$, where $\sigma^{ij}, i, j = 1, \dots, k$, are the elements of the inverse of

Σ . The required data were generated by using the IMSL Fortran routine DRNMVN.

For each of the twelve cases identified and at each value of Δ^2 , the simulation output consists of 1000 replicates of the actual error rates of discriminant analysis and logistic regression. Averaging the two sets of 1000 actual error rate values provides estimates of the expected actual error rates (unconditional error rates) of the two techniques. This is the most obvious way of comparing the error rate performance of discriminant analysis and logistic regression. However, investigation of the actual error rate values indicates that a more detailed summary will be informative. It was therefore decided to summarise the simulation output by means of boxplots. These boxplots were constructed for each of the values of Δ^2 that were considered, but only a representative selection of these plots is shown.

A selection of the boxplots for the normal case is given in Figs. 2.1 - 2.4. In addition, the means and standard deviations of the actual error rates are given in Tables 2.1 and 2.2. Each of these figures represents a fixed correlation and Mahalanobis distance. On each graph the following coding is used to denote the actual error rates of discriminant analysis and logistic regression for the different cases: DA_1 and LR_1 are used for small samples ($n_0 = n_1 = 25$) and $k = 2$; DA_2 and LR_2 for mixed samples $n_0 = 25, n_1 = 50$ and $k = 2$; DA_3 and LR_3 for large samples ($n_0 = n_1 = 100$) and $k = 2$; DA_4 and LR_4 for small samples and $k = 10$; DA_5 and LR_5 for mixed samples and $k = 10$ and DA_6 and LR_6 for large samples and $k = 10$. The boxplots were constructed using S-Plus. The notches in the boxes indicate the respective medians of the actual error rates. If the notches do not overlap, it indicates a difference in location at a rough 5% significance level (cf. the S-PLUS Reference Manual, 1991).

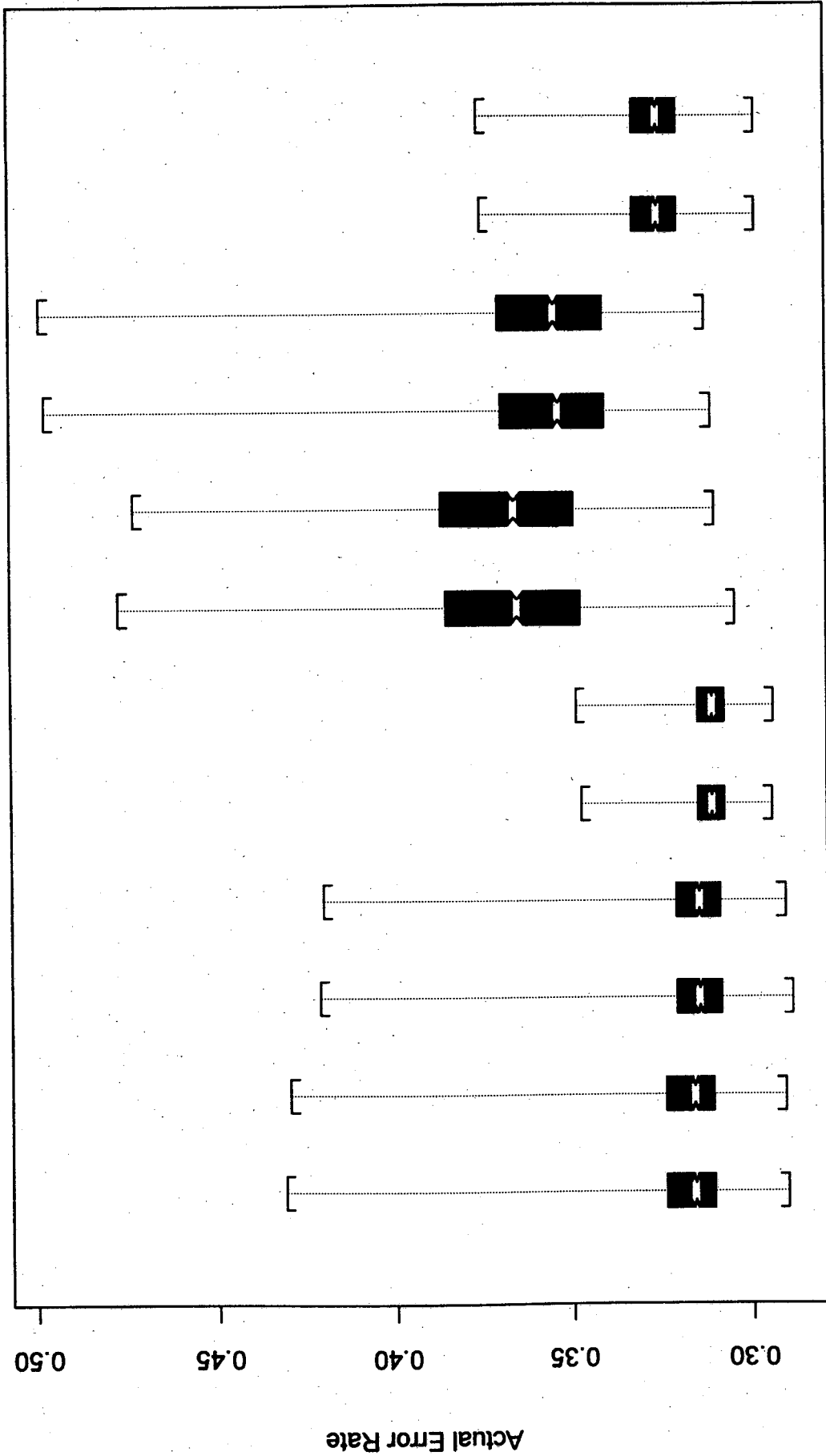
A number of points are clear from perusal of Figs. 2.1 - 2.4 and Tables 2.1 and 2.2.

1. At a fixed configuration, the median actual error rates of discriminant analysis and logistic regression differ only slightly, except for the small and mixed sample cases with $k = 10$ at moderate to large values of Δ^2 (see Figs. 2.2 and 2.4). In these cases the median actual error rate of discriminant analysis is significantly lower than that of logistic regression. The same trends are evident when considering the means of the actual error rates in Tables 2.1 and 2.2. These results are in line with the asymptotic findings of McLachlan and Byth (1979) that the differences in the error rates of discriminant analysis and logistic regression are generally very small in the case of normal data. Nevertheless, in view of the fact that discriminant analysis never performs worse than logistic regression and outperforms logistic regression appreciably in some practically important cases, the use of discriminant analysis is recommended for normal feature data.
2. Comparing corresponding graphs in Figs. 2.1 - 2.2 (representing uncorrelated cases) and Figs. 2.3 - 2.4 (representing correlated cases), it is clear that the presence of dependence between the feature variables has little or no effect on the actual error rates. The same conclusion is reached by comparing corresponding entries in Tables 2.1 and 2.2. It should be borne in mind that the actual error rates displayed in e.g. Fig. 2.3 correspond to the same Mahalanobis distance between the groups as in Fig. 2.1, i.e. the influence of a non-diagonal covariance matrix was taken into account when specifying the elements of the mean vector μ_1 (see the explanation of the parameterisation given above). Naturally, if the mean vectors are kept fixed, a decrease in error rate is expected if the introduction of correlation between the feature variables leads to an increase in the value of Δ^2 (cf. Mardia et al., 1988, p. 324).
3. For a fixed number of variables, an increase in the total sample size ($n = n_0 + n_1$) leads to a reduction in the actual error rates of both techniques. This reduction is larger in the cases where $k = 10$ than in the cases where $k = 2$. For the case $k = 10$,

the superiority of discriminant analysis to logistic regression at certain values of Δ^2 , seems to depend on sample size. The difference is large in the small sample case, smaller in the mixed sample case and largely disappears when the sample sizes are large (see Figs. 2.1 - 2.2). The variation of the error rates as displayed by the ranges in the boxplots and the standard deviations in Tables 2.1 and 2.2, is also much larger for small and mixed sample cases than for large sample cases. For small and mixed samples the error rates are highly variable, especially in the case where $k = 10$. These findings are valid for the cases of uncorrelated and correlated feature variables.

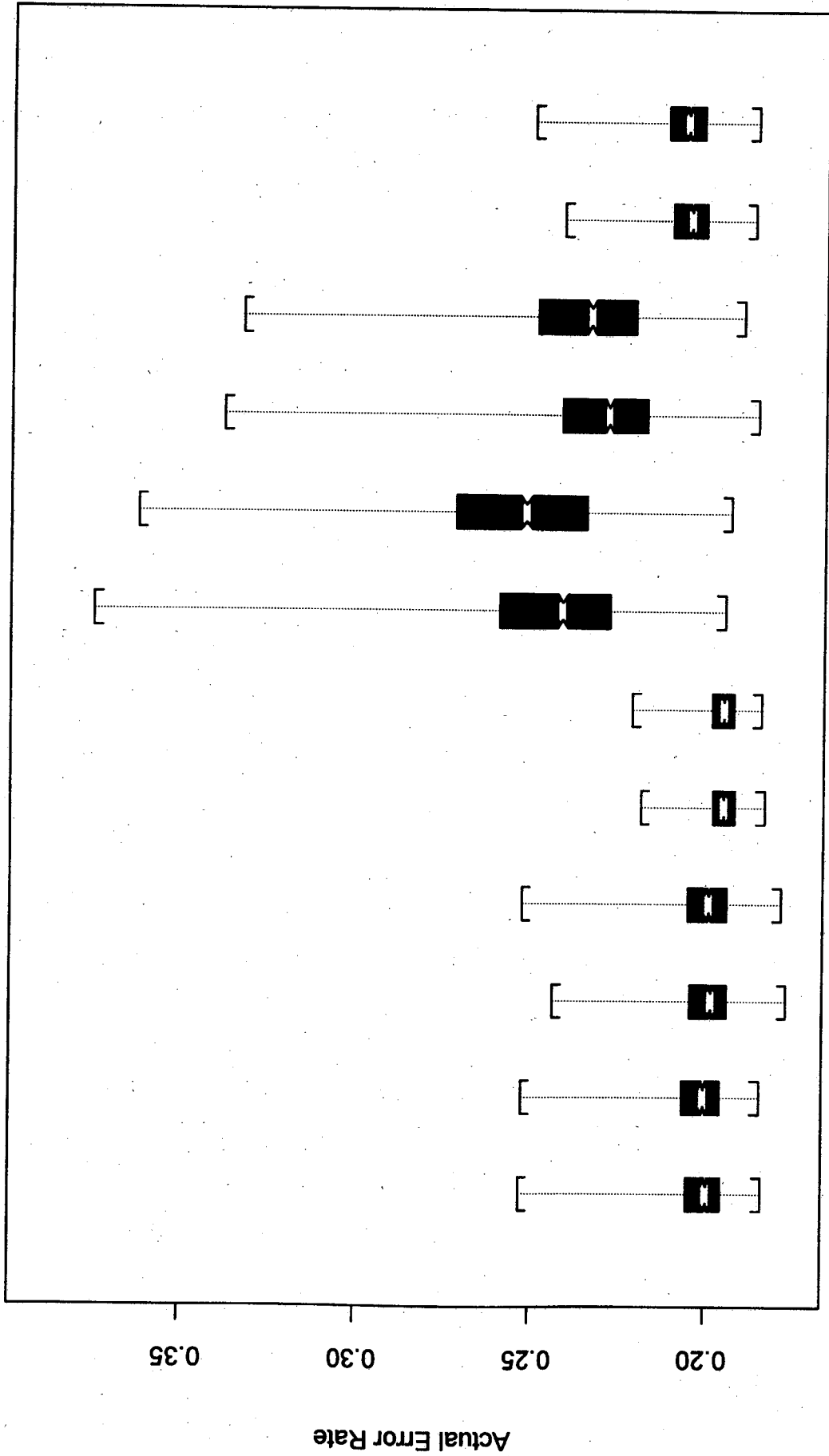
4. For a fixed sample size, the error rates are smaller for the cases where $k = 2$ than for the cases where $k = 10$. The difference seems to decrease with an increase in the total sample size. For fixed sample size, the variation in the error rates is larger when $k = 10$ than when $k = 2$.

From remarks 3 and 4 it is clear that the total sample size relative to the number of variables has an influence on the magnitude of the error rates. In fact, for the cases under consideration, the actual error rate is a monotone decreasing function of the ratio of the total sample size to the number of feature variables.



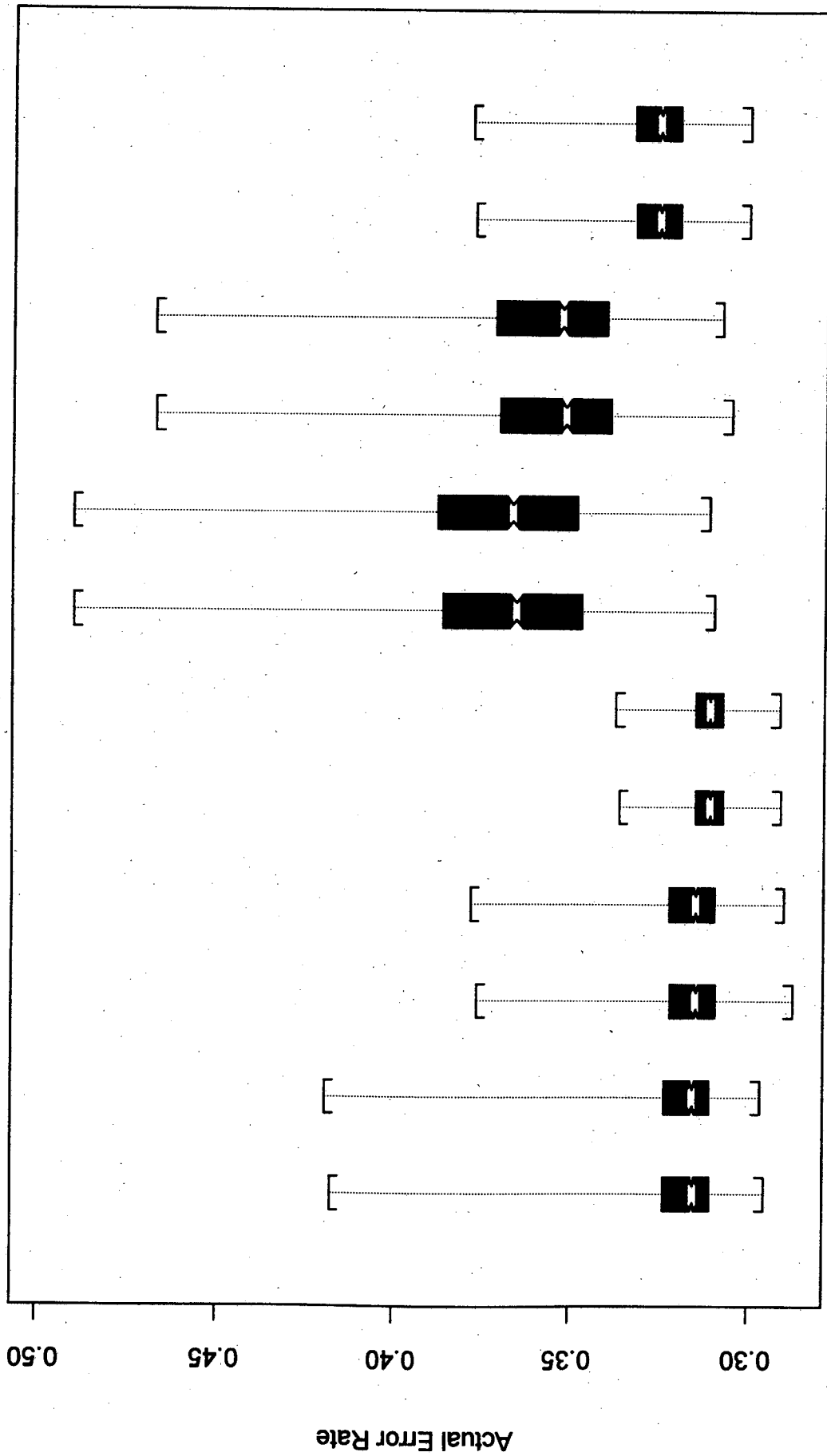
DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6

FIG. 2.1: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, NORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1



DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6

FIG. 2.2: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, NORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 3



DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6

FIG. 2.3: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, NORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE=1

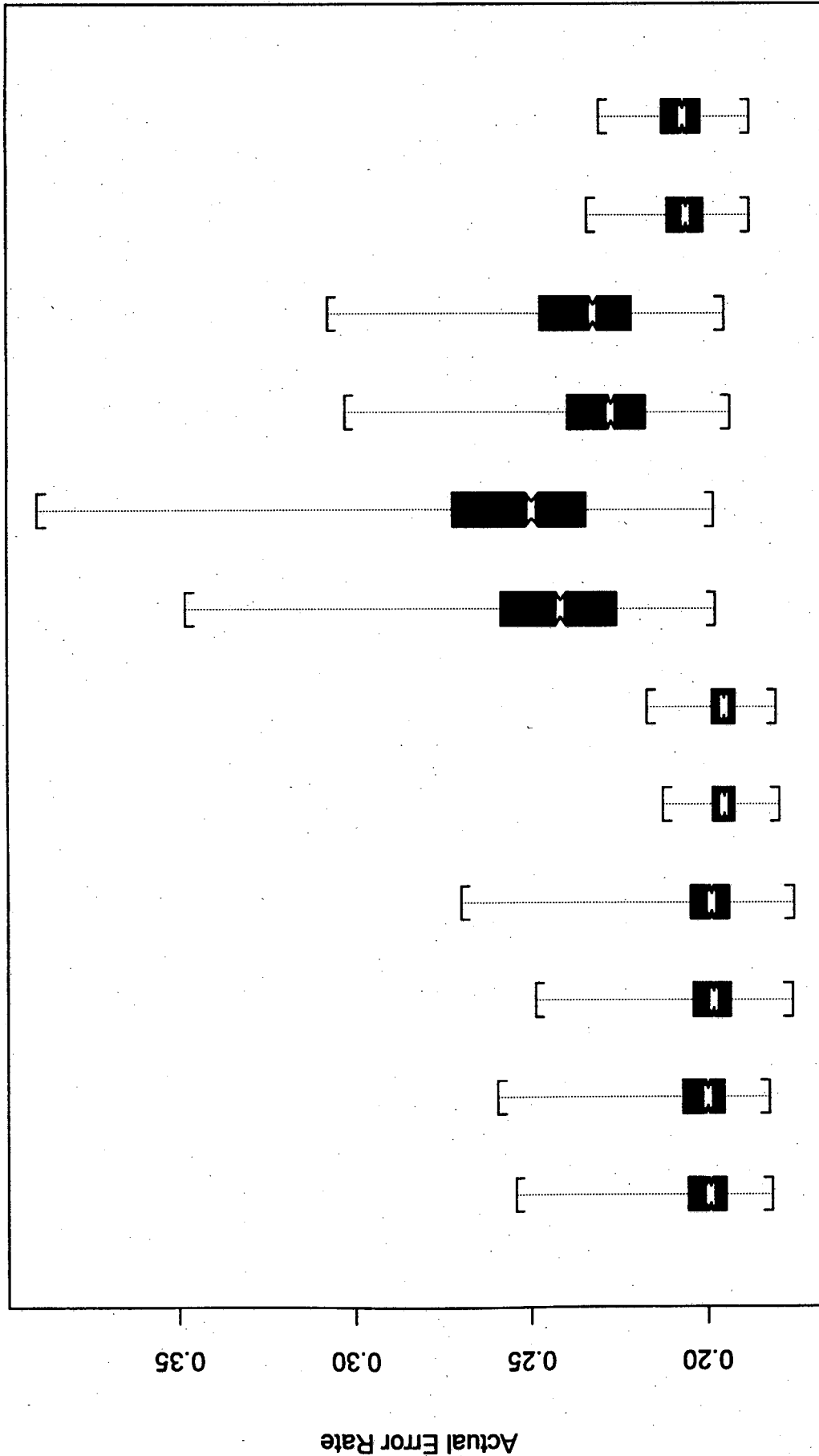


FIG. 2.4: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, NORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 3

TABLE 2.1 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES TWO GROUPS, NORMAL DATA ($\rho = 0$)

k=2	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.50016 (.00489)	.50018 (.00492)	.50059 (.00685)	.50059 (.00687)	.49990 (.00518)	.49989 (.00519)
1	.31982 (.01499)	.31990 (.01495)	.31624 (.01202)	.31649 (.01212)	.31129 (.00553)	.31331 (.00557)
2	.24819 (.01042)	.24860 (.01069)	.24623 (.00983)	.24658 (.00992)	.24189 (.00494)	.24202 (.00503)
3	.20143 (.00915)	.20227 (.00963)	.19957 (.00838)	.20028 (.00903)	.19540 (.00448)	.19561 (.00456)
4	.16615 (.00943)	.16740 (.01403)	.16401 (.00774)	.16520 (.00909)	.16037 (.00410)	.16066 (.00423)

k=10	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.50000 (.00498)	.49997 (.00502)	.49988 (.00706)	.49999 (.00713)	.49986 (.00475)	.49987 (.00474)
1	.36896 (.02863)	.37030 (.02859)	.35739 (.02379)	.35835 (.02410)	.32660 (.00965)	.32672 (.00968)
2	.29385 (.02571)	.29860 (.02747)	.28062 (.02008)	.28400 (.02170)	.25433 (.00796)	.25482 (.00827)
3	.24506 (.02442)	.25444 (.02880)	.23062 (.00188)	.23666 (.02146)	.20624 (.00727)	.20722 (.00754)
4	.20617 (.02282)	.22076 (.03289)	.19186 (.01618)	.20222 (.02031)	.17040 (.00679)	.17202 (.00736)

TABLE 2.2 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES TWO GROUPS, NORMAL DATA ($\rho = 0.9$)

k=2	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.50017 (.00476)	.50019 (.00473)	.49999 (.00683)	.49987 (.00686)	.50022 (.00476)	.50022 (.00477)
1	.31928 (.01496)	.31936 (.01498)	.31607 (.01093)	.31633 (.01105)	.31110 (.00592)	.31115 (.00595)
2	.24829 (.01115)	.24861 (.01165)	.24608 (.00872)	.24653 (.00898)	.24190 (.00514)	.24199 (.00510)
3	.20136 (.00967)	.20216 (.01031)	.19920 (.00869)	.19998 (.00927)	.19534 (.00467)	.19555 (.00471)
4	.16603 (.00894)	.16774 (.01056)	.16412 (.00764)	.16528 (.00871)	.16054 (.00427)	.16082 (.00441)

k=10	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.50004 (.00502)	.50001 (.00498)	.49963 (.00701)	.49957 (.00697)	.50008 (.00485)	.50007 (.00485)
1	.36917 (.02994)	.37060 (.03029)	.35675 (.02439)	.35790 (.02464)	.32702 (.01025)	.32718 (.01028)
2	.29716 (.02768)	.30180 (.03036)	.28256 (.02205)	.28566 (.02316)	.25468 (.00825)	.25511 (.00843)
3	.24450 (.02497)	.25433 (.02836)	.22936 (.01748)	.23531 (.01994)	.20633 (.00739)	.20731 (.00768)
4	.20757 (.02348)	.22388 (.03160)	.19272 (.01766)	.20188 (.02298)	.16997 (.00644)	.17168 (.00719)

2.4.2 THE DOUBLE EXPONENTIAL CASE

The double exponential distribution was included in the simulation study as an example of a heavy-tailed symmetrical distribution. Exactly the same cases as described in paragraph 2.4.1 for the normal case, were investigated. The required data were generated as follows.

The probability density function (p.d.f.) of the univariate double exponential distribution with mean μ and variance σ^2 , is given by

$$f(x) = \exp\{-\sqrt{2}|x - \mu|/\sigma\} / \sqrt{2}\sigma, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0. \quad (2.4.2)$$

An observation from this distribution can be generated as follows. Let U_1 and U_2 be i.i.d. (independent and identically distributed) uniform(0,1) random variables. Then $Y = -\log(U_1)$ is a standard exponential random variable and

$$Z = YI(U_2 < 0.5) - YI(U_2 \geq 0.5)$$

has p.d.f. (2.4.2) with $\mu = 0$ and $\sigma = \sqrt{2}$. Hence, $X = \sigma Z / \sqrt{2} + \mu$ will have p.d.f. (2.4.2). For $\Sigma = I$, this procedure was independently repeated k times, taking $\sigma = 1$, thereby obtaining values of the k feature variables. The required uniform(0,1) values were generated by using the IMSL Fortran routine DRNUN. The same values of Δ^2 as in the normal case were used, and the same parameterisation of the two mean vectors as in the normal case was used to give these Mahalanobis distances.

For Σ as in (2.4.1), the problem is to generate values of random variables X_1, \dots, X_k that have marginal p.d.f.'s as in (2.4.2) and that have the required covariance structure. A procedure that approximately accomplishes this can be based on the following argument.

Consider a random vector $Z = [Z_1, \dots, Z_k]'$ that is multivariate normally distributed with $E(Z) = \mathbf{0}$ and with covariance matrix as in (2.4.1). Then $U_j = \Phi(Z_j)$, $j = 1, \dots, k$, are uniform(0,1) random variables. Now suppose G is some given cumulative distribution function. Then $Y_j = G^{-1}(U_j)$, $j = 1, \dots, k$, will be random variables, each with marginal distribution function G . The question now arises: given that Z has covariance matrix Σ , and that the Y_j , $j = 1, \dots, k$, are obtained from Z as described above, what can be said about the covariance matrix of $Y = [Y_1, \dots, Y_k]'$? This seems to be a difficult question to answer in general. For the case corresponding to (2.4.2) with $\mu = 0$ and $\sigma^2 = 1$,

$$G(t) = \begin{cases} \frac{1}{2}e^{\sqrt{2}t} & , \text{ if } t \leq 0 \\ 1 - \frac{1}{2}e^{-\sqrt{2}t} & , \text{ if } t > 0, \end{cases} \quad (2.4.3)$$

and therefore

$$G^{-1}(s) = \begin{cases} \frac{1}{\sqrt{2}}\log(2s) & , \text{ if } 0 < s \leq 0.5 \\ -\frac{1}{\sqrt{2}}\log(2(1-s)) & , \text{ if } 0.5 < s < 1. \end{cases} \quad (2.4.4)$$

Simulation experiments were conducted with this G and G^{-1} , using different values for ρ in (2.4.1). These experiments indicated that by taking $\rho = 0.905$ in the covariance matrix (2.4.1) of Z , a covariance matrix is obtained for the random vector Y that is very nearly equal to (2.4.1) with $\rho = 0.9$. Based on these findings, values of X_1, \dots, X_k with marginal p.d.f.'s given by (2.4.2) with $\mu = 0$ and $\sigma = 1$, and with covariance matrix approximately given by (2.4.1) with $\rho = 0.9$, were generated by taking

$$X_j = G^{-1}[\Phi(Z_j)], \quad j = 1, \dots, k,$$

where Z_1, \dots, Z_k satisfy the multivariate normal requirements stated above, and with G^{-1} as in (2.4.4). This was done for both of the groups in the study, and the required Mahalanobis distances were thereafter obtained by adding the appropriate μ_{ij} -values to the observations generated for group Π_1 .

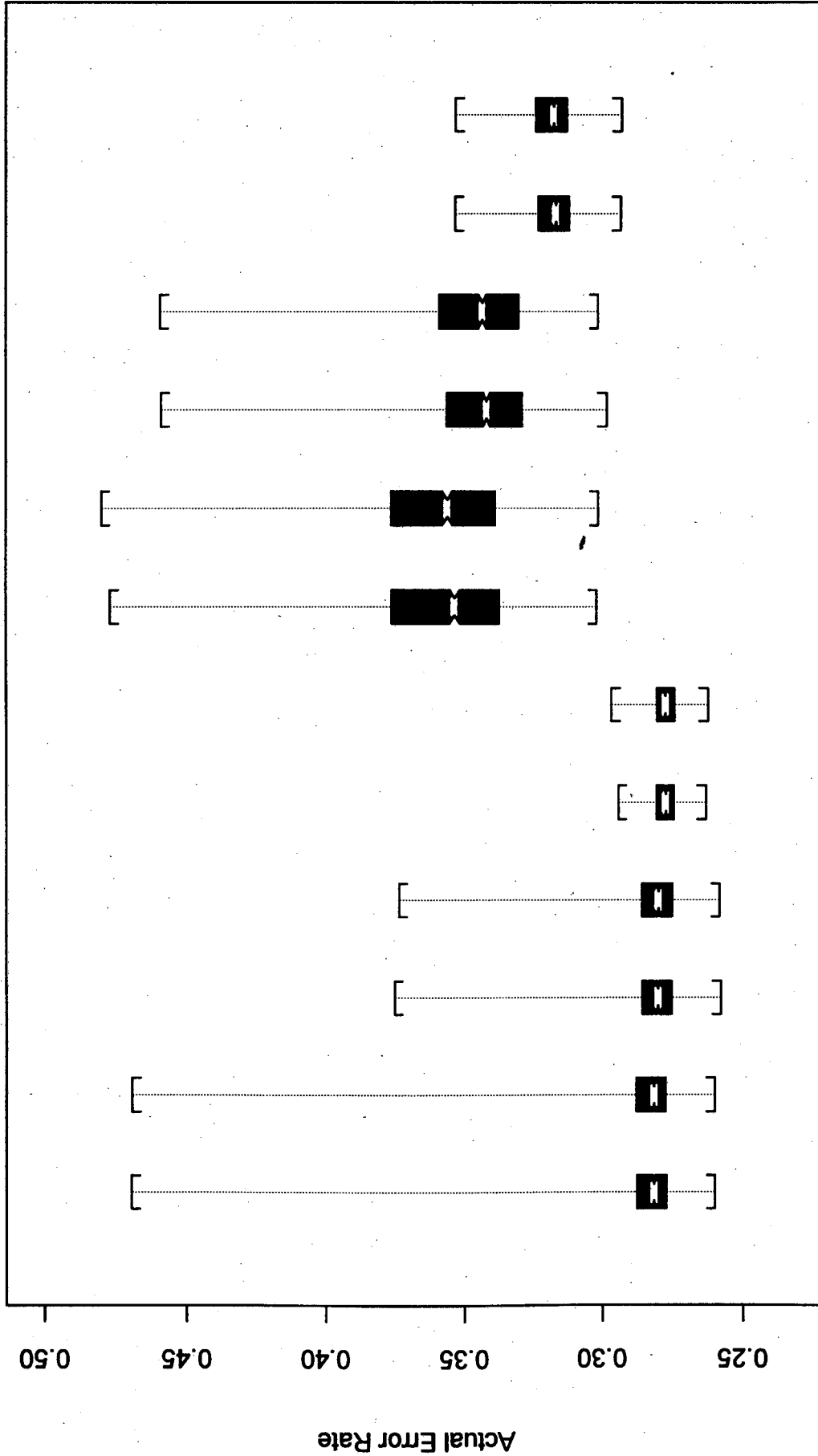
The simulation output is summarised in boxplots, of which a selection appears in Figs. 2.5 - 2.8. Tables 2.3 and 2.4 contain the means and standard deviations of the actual error rates. The same coding as in the normal case is used to denote the different cases. Perusal of these graphs and tables leads to the following remarks.

1. The differences in the actual error rates of discriminant analysis and logistic regression are once again very small, with the same exception as in the normal case, viz. the small and mixed sample cases with $k = 10$. For these cases, discriminant analysis performed significantly better than logistic regression when $\Delta^2 = 3, 4$ (see Figs. 2.6 and 2.8 for cases where $\Delta^2 = 3$). Discriminant analysis is therefore once more the method of choice.
2. The effect of sample size and the number of feature variables on the actual error rates seems to be the same as in the normal case.
3. The introduction of correlation between the feature variables affected the error rates of both discriminant analysis and logistic regression. When comparing the error rates of the uncorrelated cases (Figs. 2.5 and 2.6 and Table 2.3) to the error rates of the corresponding equicorrelated cases (Figs. 2.7 and 2.8 and Table 2.4) at the same

values of Δ^2 , it is evident that the error rates are lower in the equicorrelated case. This is however accompanied by slightly larger variation.

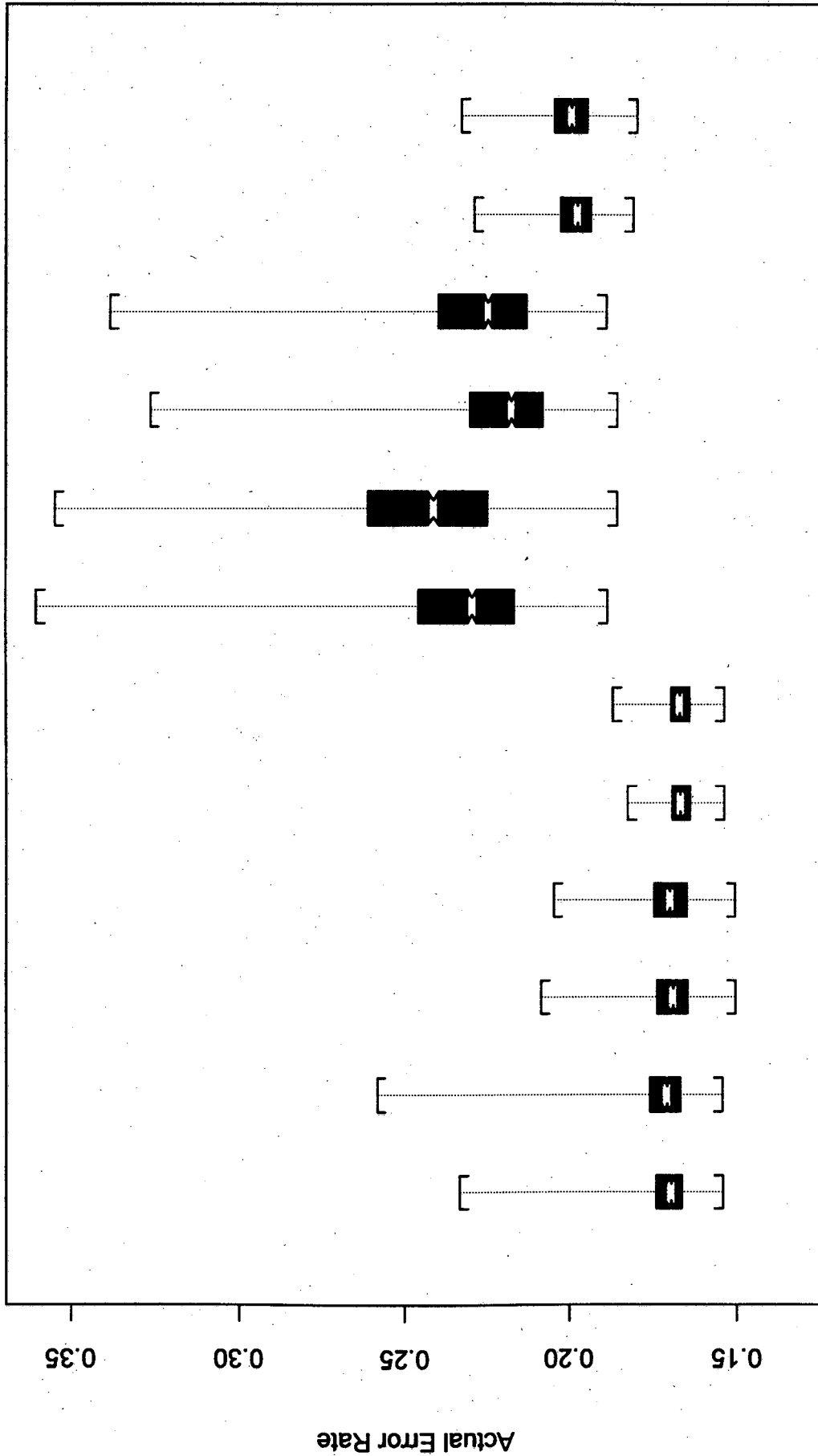
4. Especially at small values of Δ^2 ($\Delta^2 = 1, 2$) the ranges of the actual error rates of both techniques are very large in the small sample case with $k = 2$ (see Figs. 2.5 and 2.7 for cases where $\Delta^2 = 1$).

5. Comparing the actual error rates of the two techniques for the double exponential case with the error rates of corresponding configurations of the normal case at the same values of Δ^2 , it is clear that the error rates are much smaller for the double exponential case than for the normal case. The difference between the corresponding error rates is larger for $k = 2$ than for $k = 10$. This seems intuitively surprising, in view of the fact that the double exponential distribution is heavy tailed and discrimination could therefore be expected to be more difficult than in the normal case. A closer examination of data from the two distributions suggested an explanation for this error rate behaviour. Two random samples of 1000 observations each were generated from two 2-dimensional normal populations with respective mean vectors $\mathbf{0}$ and $\mu_1 = (\Delta/\sqrt{2}, \Delta/\sqrt{2})'$ and common covariance matrix $\Sigma = \mathbf{I}$. The same was done for the double exponential distribution. The two normal samples were then represented on a single scatterplot (see Fig. 2.9), with a similar graph being constructed for the double exponential samples (see Fig. 2.10). Inspection of these graphs shows the following: although it is evident that the double exponential samples contain a larger number of extreme observations that will clearly be misclassified by a classification rule, larger proportions of the double exponential samples from the different groups are concentrated some distance apart than in the case of the normal samples. These observations will almost certainly be correctly classified by any reasonable rule. Although the normal samples contain fewer extreme observations that will definitely be misclassified, in total the overlap between the two normal samples is larger than in the comparable double exponential case, leading to the larger actual error rate in the normal case.



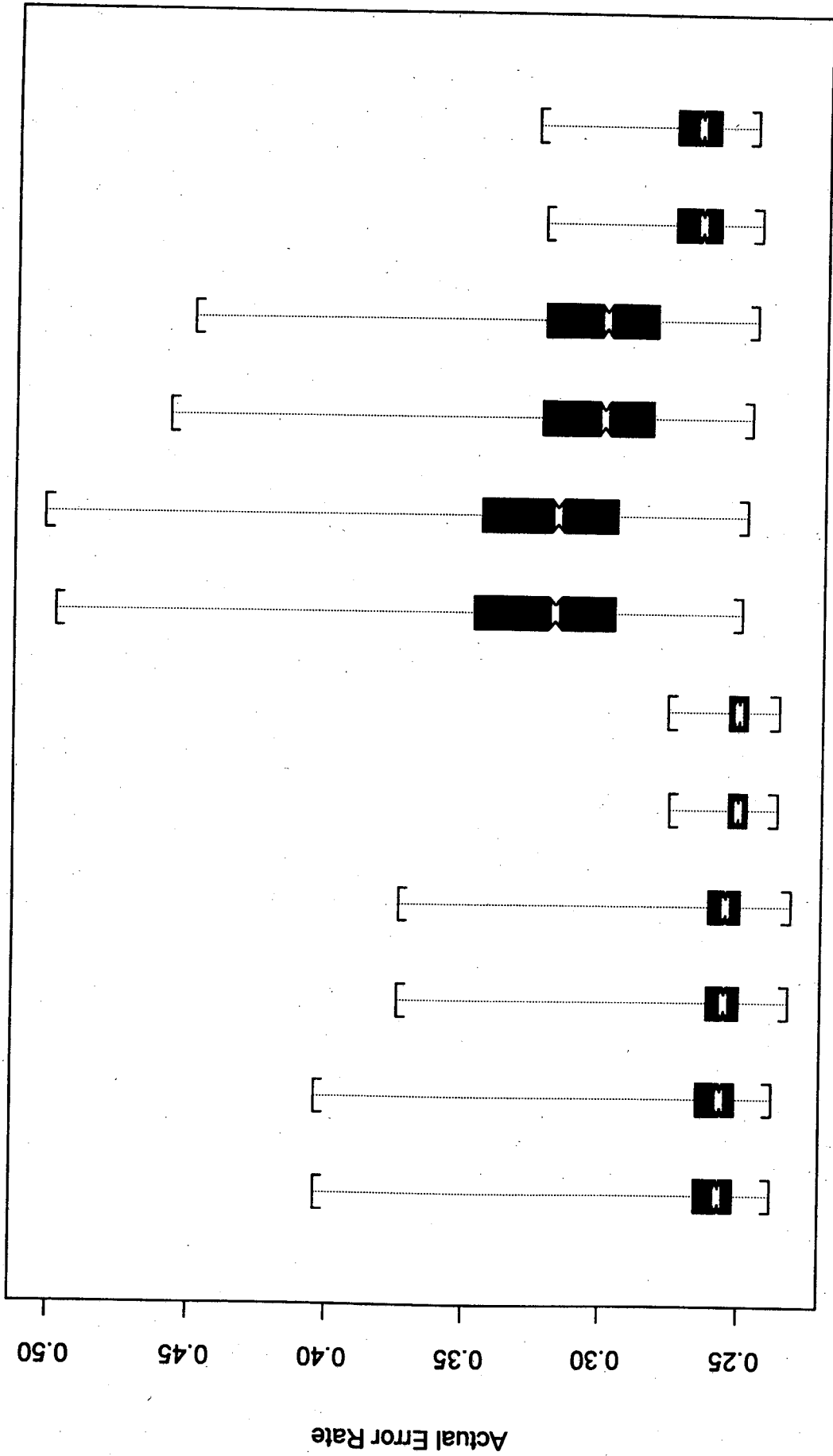
DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6

FIG. 2.5: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, DOUBLE EXPONENTIAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1



DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6

FIG. 2.6: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, DOUBLE EXPONENTIAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 3



DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6

FIG. 2.7: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, DOUBLE EXPONENTIAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 1

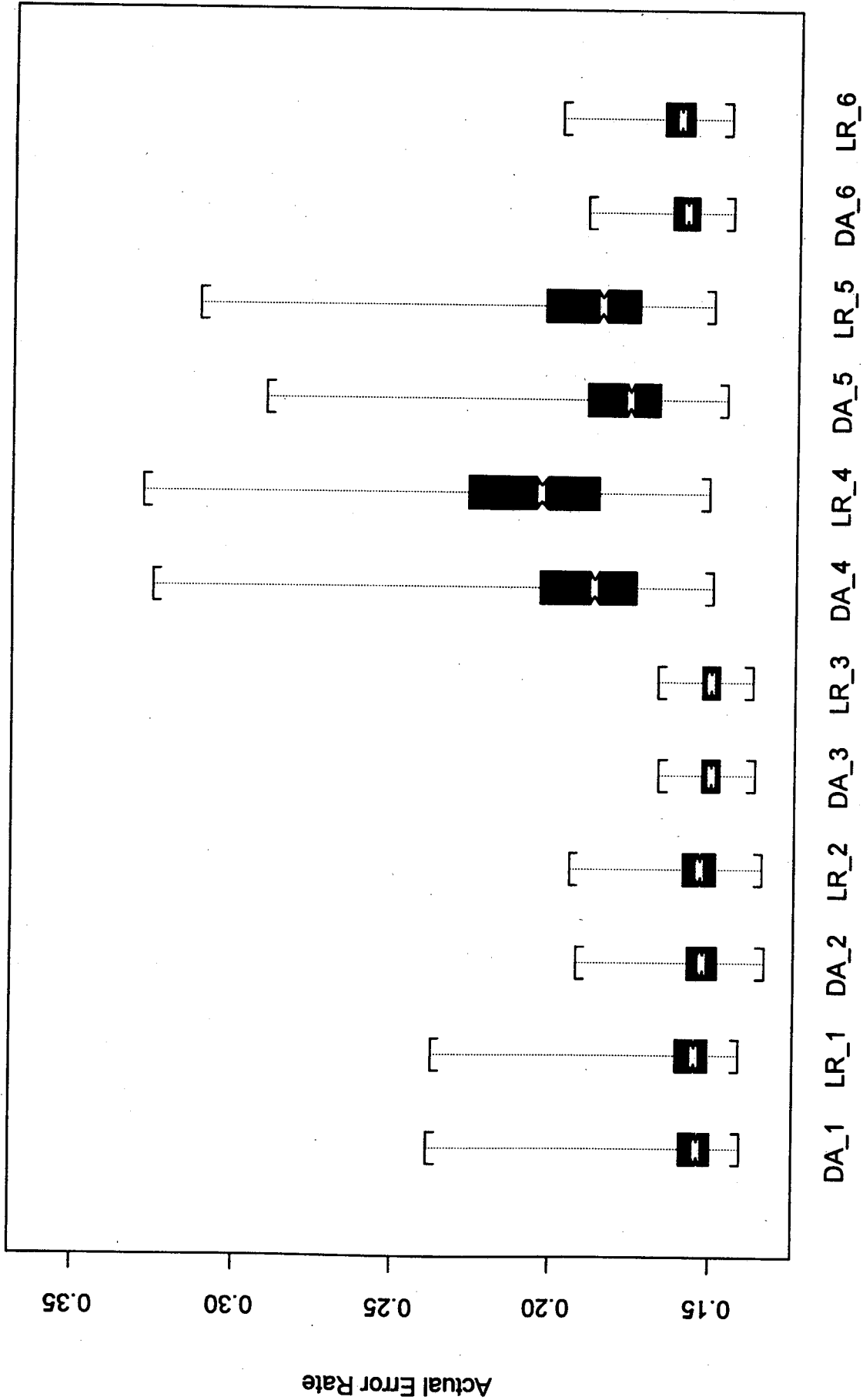


FIG. 2.8: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, DOUBLE EXPONENTIAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 3

TABLE 2.4 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES TWO GROUPS, DOUBLE EXPONENTIAL DATA ($\rho = 0.9$)

k=2	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.49999 (.00491)	.49998 (.00488)	.50014 (.00704)	.50012 (.00702)	.50000 (.00507)	.49999 (.00506)
1	.26318 (.02029)	.26261 (.01962)	.25849 (.01540)	.25827 (.01499)	.25160 (.00489)	.25150 (.00486)
2	.19535 (.01200)	.19590 (.01253)	.19230 (.00876)	.19267 (.00915)	.18828 (.00453)	.18836 (.00456)
3	.15562 (.00866)	.15711 (.01018)	.15322 (.00737)	.15452 (.00828)	.15036 (.00410)	.15062 (.00429)
4	.12941 (.00822)	.13156 (.00954)	.12740 (.00658)	.12889 (.00790)	.12438 (.00356)	.12484 (.00387)

k=10	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.50018 (.00495)	.50018 (.00490)	.50040 (.00699)	.50025 (.00708)	.49993 (.00505)	.49995 (.00502)
1	.32426 (.03867)	.32301 (.03760)	.30656 (.03190)	.30576 (.03127)	.26933 (.01302)	.26981 (.01260)
2	.24014 (.03187)	.24733 (.03283)	.22599 (.02504)	.23065 (.02509)	.19951 (.00778)	.20085 (.00786)
3	.19225 (.02537)	.20890 (.03082)	.18057 (.01913)	.19086 (.02352)	.15990 (.00625)	.16224 (.00686)
4	.16075 (.02245)	.18142 (.03121)	.14860 (.01414)	.16403 (.02364)	.13303 (.00545)	.13600 (.00651)

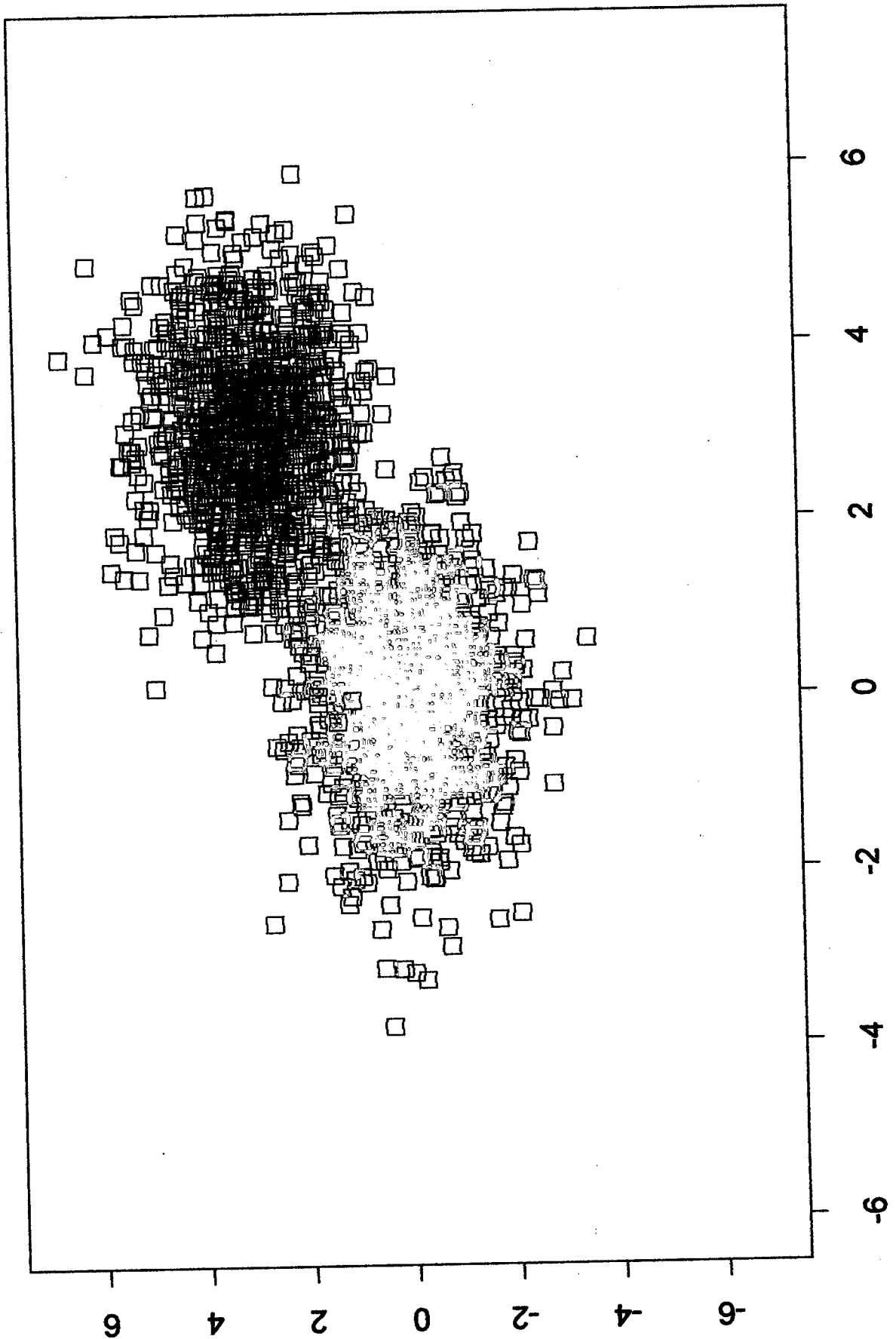


FIG. 2.9: SCATTERPLOT OF DATA FROM TWO NORMAL GROUPS, MAHALANOBIS DISTANCE = 3

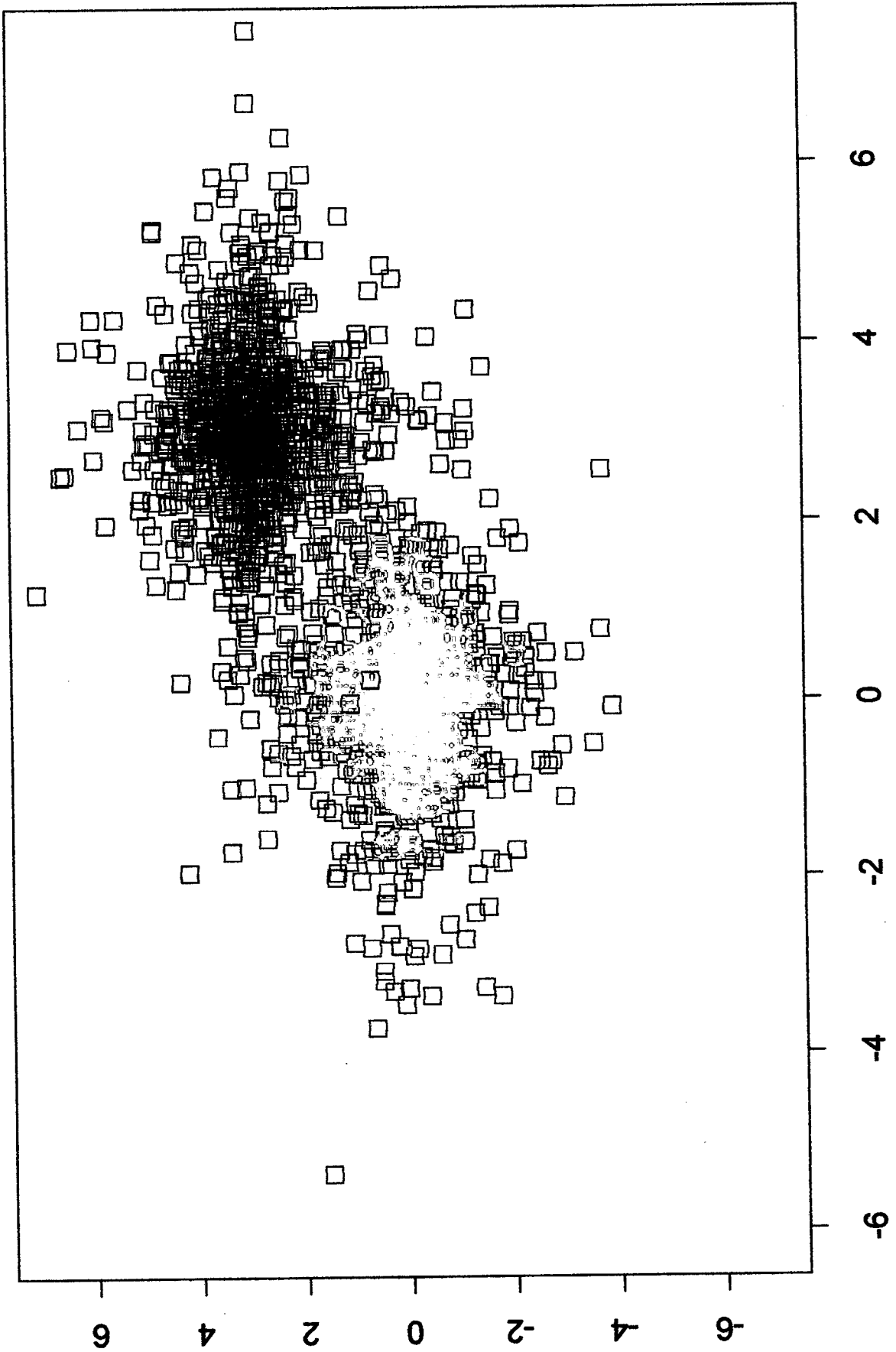


FIG. 2.10: SCATTERPLOT OF DATA FROM TWO DOUBLE EXPONENTIAL GROUPS,
MAHALANOBIS DISTANCE = 3

2.4.3 THE LOGNORMAL CASE

To study the classification performance of the normal linear discriminant rule and the logistic discriminant rule in the case of a skewed distribution, data were generated from the multivariate lognormal distribution. The same twelve cases described in paragraph 2.4.1 for the normal case, were included in this investigation. The Johnson translation system (Johnson, 1986) was used to generate the data. A k -dimensional variable \mathbf{Z} was generated from the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ , using the IMSL routine DRNMVN. The components of \mathbf{Z} were then transformed as follows to yield lognormal variables:

$$X_{ij} = \lambda_{ij} \exp(Z_{ij}) + \xi_{ij}, \quad i = 0, 1; j = 1, \dots, k.$$

For the uncorrelated case, $\Sigma = \mathbf{I}$ was used. For the correlated case, simulation experiments similar to those described above for the double exponential distribution were conducted. From these experiments it was concluded that using $\rho = 0.935$ in (2.4.1) for the multivariate normal distribution results in a covariance matrix as in (2.4.1) for the lognormal variables with ρ approximately equal to 0.9. The shape of the resulting lognormal distribution is determined by the means and variances of the original normal variables. The parameters λ_{ij} and ξ_{ij} do not affect the shape of the distribution, but control the scale and location of the X_{ij} .

For each of the twelve cases studied, the actual error rates of the two techniques were estimated by simulation at each of the following values of the squared Mahalanobis distance between the two populations: $\Delta^2 = 0, 0.5, 1, 1.5, 2, 3$ and 4. To obtain these distances, the following choices were made for the values of λ_{ij} and ξ_{ij} :

$$\lambda_{ij} = 1/\sqrt{e^2 - e}, \quad i = 0, 1; j = 1, \dots, k$$

$$\xi_{0j} = -1/\sqrt{e-1} \quad \text{and} \quad \xi_{1j} = \Delta / \sqrt{\sum_{i=1}^k \sum_{h=1}^k \sigma^{ih}} - 1/\sqrt{e-1}, \quad j = 1, \dots, k$$

where σ^{ih} , $i, h = 1, \dots, k$, are the elements of the inverse of the covariance matrix. For the uncorrelated case where $\Sigma = \mathbf{I}$, the term $\sqrt{\sum_{i=1}^k \sum_{h=1}^k \sigma^{ih}}$ is equal to k , the number of variables.

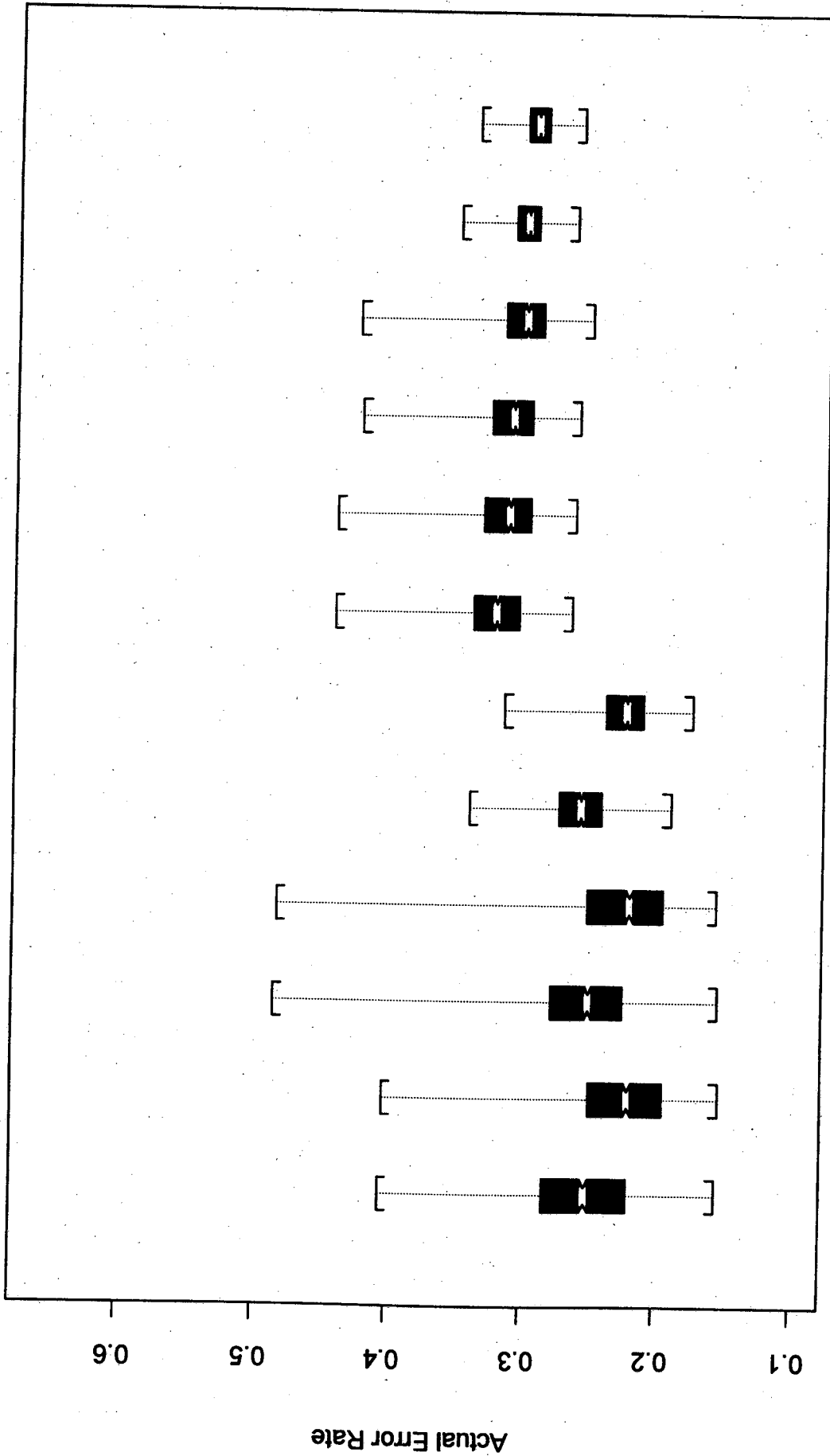
These choices of λ_{ij} and ξ_{ij} yield lognormal variables with

$$\mu_{0j} = 0 ; \mu_{1j} = \Delta / \sqrt{\sum_{i=1}^k \sum_{h=1}^k \sigma_{ij}^{2h}} \text{ and } \sigma_{ij}^2 = 1, \quad j = 1, \dots, k.$$

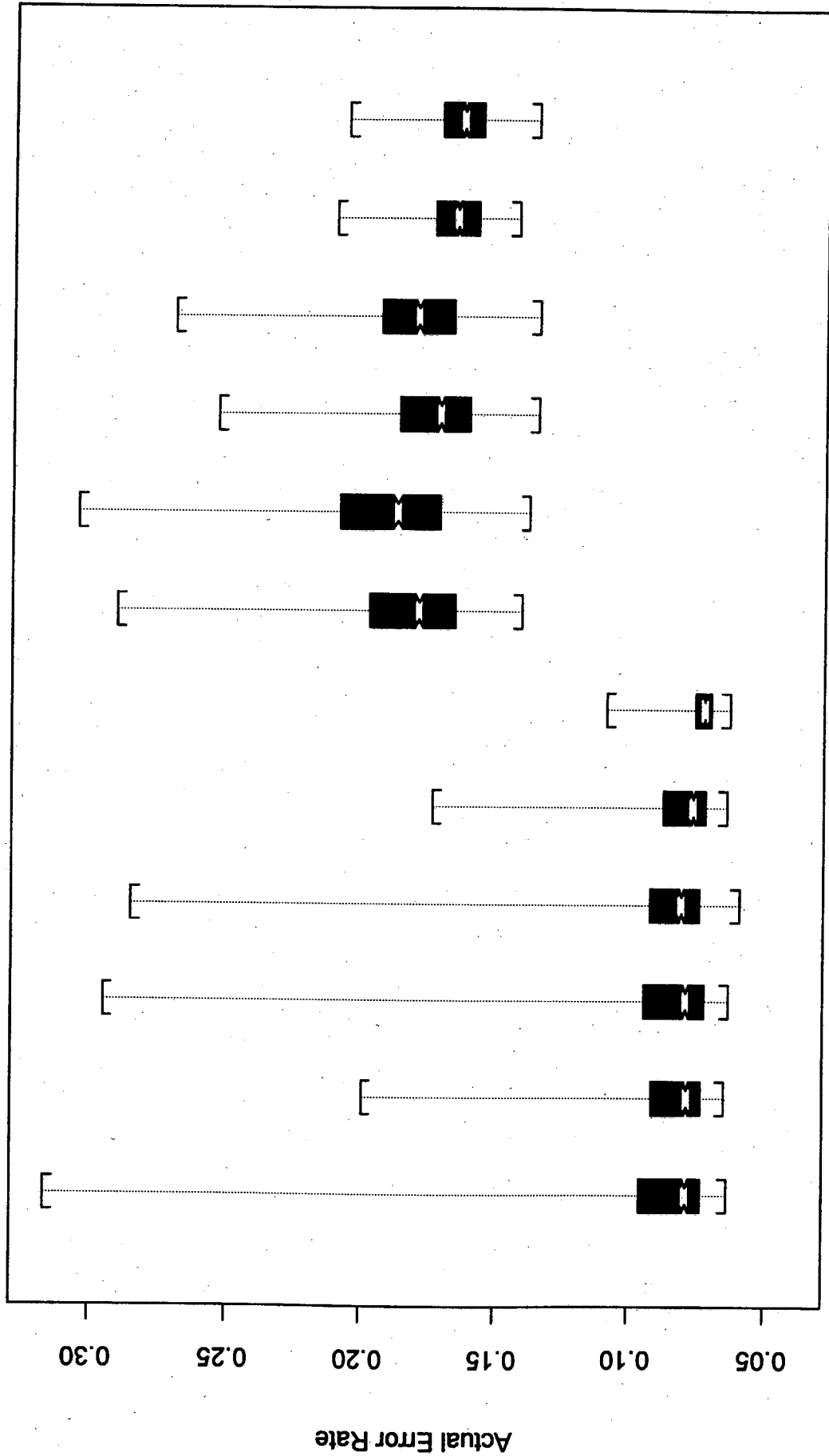
A selection of boxplots of the simulation output of the lognormal cases are given in Figs. 2.11 - 2.14. Tables 2.5 and 2.6 contain the means and standard deviations of the actual error rates. The following points can be made:

1. The median actual error rate of logistic regression is significantly lower than that of discriminant analysis at small values of Δ^2 ($\Delta^2 = 1, 2$) (see Figs. 2.11 and 2.13 for cases where $\Delta^2 = 1$). At larger values of Δ^2 ($\Delta^2 = 3, 4$), the differences in the actual error rates are smaller and neither of the techniques consistently outperforms the other (see Figs. 2.12 and 2.14 for cases where $\Delta^2 = 3$). In the case of independent feature variables at a value of $\Delta^2 = 3$, the median actual error rate of logistic regression is significantly smaller than the median actual error rate of discriminant analysis for the large sample case with $k = 2$, while the opposite is true for the small and mixed sample cases with $k = 10$ (see Fig. 2.12). Logistic regression should therefore be the method of choice for lognormal data, although in cases where the ratio of the total sample size to the number of variables is small, discriminant analysis may be preferred.
2. The effect of total sample size and the number of feature variables is the same as in the normal and double exponential cases.
3. The presence of correlation between the lognormal feature variables leads to a large reduction in the error rates of discriminant analysis and logistic regression when compared to similar configurations for the uncorrelated case, especially for the cases where $k = 10$.
4. When comparing the error rates of the lognormal case to that of corresponding normal and double exponential cases at the same values of Δ^2 , it is evident that the error rates are smallest in the lognormal case. This is to be expected, since the skewed shape of the lognormal distribution results in less overlap between the two groups at a given value of Δ^2 than in both the normal and double exponential cases.

Finally, it should be mentioned that logistic regression suffers from the disadvantage that the maximum likelihood estimates of β_{0i} and β_{1i} do not always exist. This occurs in cases of complete separation of the two groups (cf. Albert and Anderson, 1984, and Lesaffre and Albert, 1989). Such cases were excluded from the simulation study, and additional cases were generated to ensure a total of 1000 valid repetitions. For the normal and double exponential distributions, this problem occurred only at very large separations ($\Delta^2 = 9$, a case which was not included in the final simulation study). In the case of the lognormal distribution however, it occurred at smaller values of Δ^2 ($\Delta^2 = 3, 4$). The problem was aggravated by an increase in the ratio of the number of variables to the total sample size.



DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6
FIG. 2.11: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, LOGNORMAL DATA,
CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1



DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6
 FIG. 2.12: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, LOGNORMAL DATA,
 CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 3

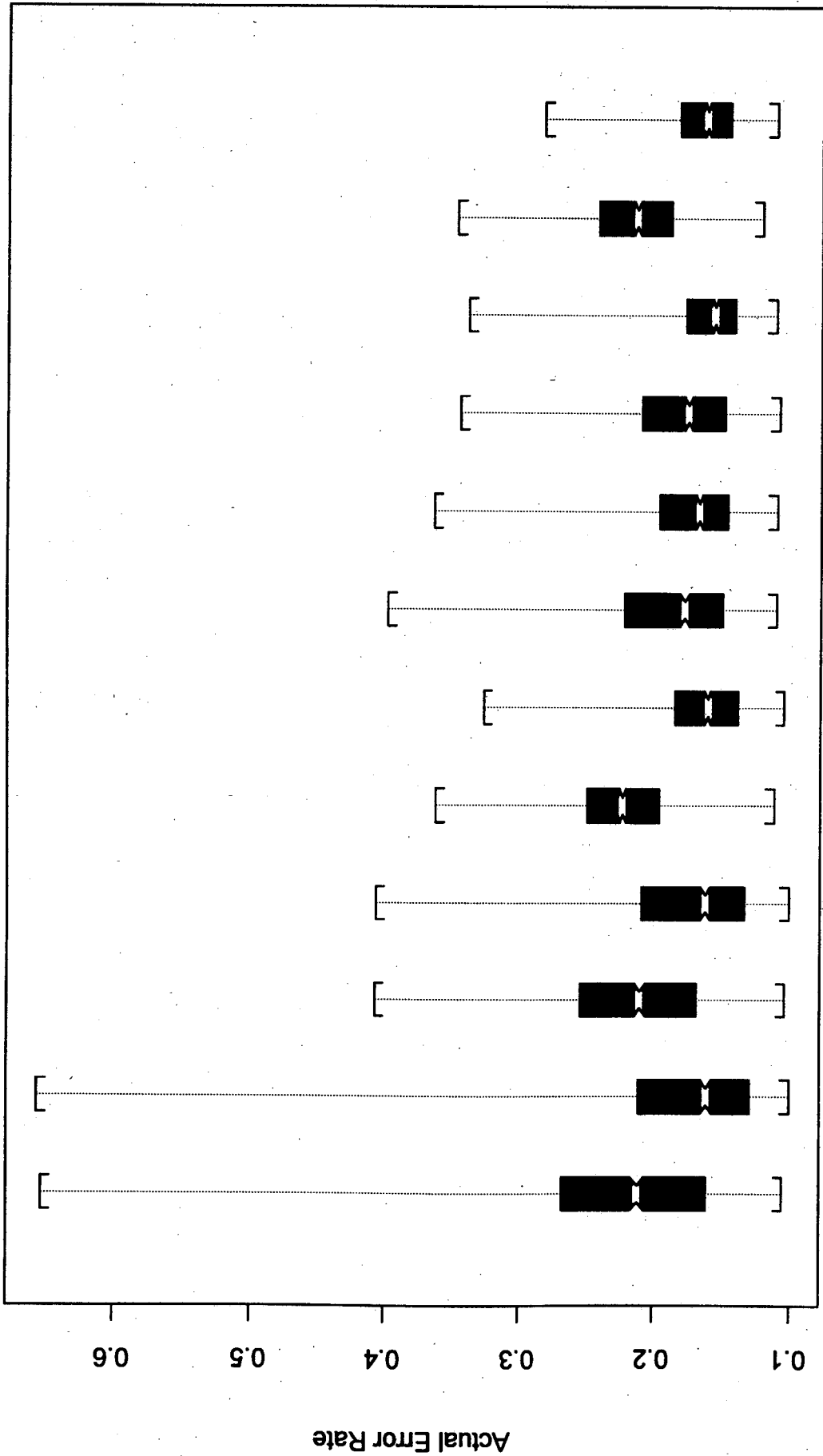
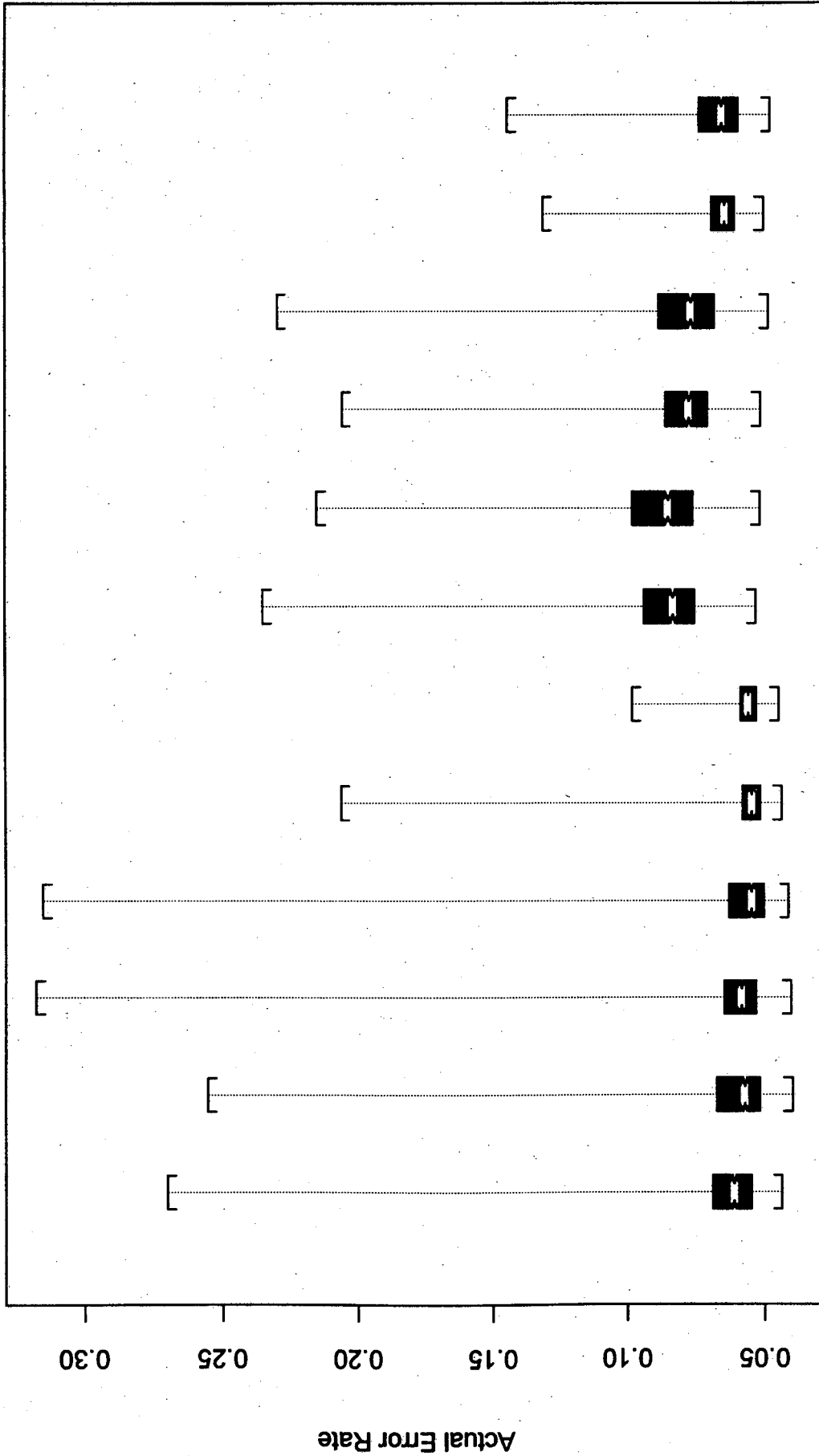


FIG. 2.13: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, LOGNORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 1



DA_1 LR_1 DA_2 LR_2 DA_3 LR_3 DA_4 LR_4 DA_5 LR_5 DA_6 LR_6

FIG. 2.14: ACTUAL ERROR RATES OF DA AND LR, 2 GROUPS, LOGNORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 3

TABLE 2.5 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES TWO GROUPS, LOGNORMAL DATA ($\rho = 0$)

k=2	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.49962 (.00456)	.49963 (.00460)	.49986 (.00681)	.49987 (.00682)	.49963 (.00467)	.49967 (.00464)
1	.25309 (.04508)	.22508 (.04129)	.25297 (.03930)	.22733 (.04354)	.25767 (.02321)	.22581 (.02174)
2	.14439 (.04280)	.12829 (.02928)	.14650 (.03950)	.13061 (.03179)	.14744 (.02433)	.12056 (.01463)
3	.09043 (.02881)	.08531 (.01859)	.088535 (.02670)	.086354 (.02040)	.08159 (.01457)	.07389 (.00612)
4	.06320 (.01371)	.06413 (.01419)	.06158 (.01391)	.06274 (.01298)	.05650 (.00562)	.05559 (.00355)

k=10	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.50012 (.00515)	.50013 (.00506)	.50016 (.00681)	.50023 (.00678)	.50017 (.00495)	.50020 (.00491)
1	.32334 (.02648)	.31731 (.02692)	.31335 (.02327)	.30515 (.02203)	.30272 (.01274)	.29631 (.01219)
2	.23982 (.02609)	.239992 (.02779)	.22831 (.02065)	.22582 (.02039)	.22077 (.01284)	.21472 (.01170)
3	.18535 (.02412)	.19199 (.02742)	.17459 (.01879)	.18160 (.02403)	.16679 (.01185)	.16443 (.01112)
4	.14596 (.02149)	.15966 (.02899)	.13542 (.01603)	.15111 (.02171)	.12798 (.00988)	.12953 (.01000)

TABLE 2.6 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES TWO GROUPS, LOGNORMAL DATA ($\rho = 0.9$)

k=2	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.49994 (.00479)	.49996 (.00485)	.49990 (.00683)	.49985 (.00693)	.50016 (.00463)	.50014 (.00465)
1	.21703 (.07048)	.17798 (.06210)	.21424 (.05873)	.18024 (.06095)	.22334 (.03937)	.16440 (.03378)
2	.10415 (.05258)	.08968 (.03402)	.09816 (.04323)	.08772 (.03264)	.08641 (.02816)	.07211 (.00739)
3	.06833 (.02845)	.06404 (.02122)	.06282 (.02307)	.05962 (.01962)	.05599 (.00959)	.05654 (.00550)
4	.05370 (.01938)	.05108 (.01559)	.05083 (.01094)	.04698 (.01306)	.04772 (.00543)	.04649 (.00470)

k=10	SMALL SAMPLES		MIXED SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR	DA	LR
0	.50012 (.00465)	.50008 (.00490)	.49979 (.00672)	.49978 (.00682)	.50004 (.00499)	.50004 (.00498)
1	.19015 (.05148)	.17508 (.03823)	.18257 (.04244)	.16225 (.02941)	.21547 (.03766)	.16637 (.02871)
2	.11262 (.02752)	.11658 (.02872)	.10259 (.02137)	.10736 (.02253)	.08890 (.02035)	.08753 (.01318)
3	.08746 (.01881)	.09088 (.02233)	.07990 (.01408)	.07995 (.01976)	.06574 (.00772)	.06805 (.01176)
4	.07440 (.01362)	.07673 (.01813)	.06754 (.01039)	.06336 (.01545)	.05588 (.00593)	.05554 (.01024)

2.5 MONTE CARLO SIMULATION STUDY: THREE GROUPS

Consider three groups Π_0, Π_1 and Π_2 with equal prior probabilities π_0, π_1 and π_2 respectively. An entity $e(\mathbf{x})$ of unknown origin can be classified into one of the three groups using the classification rule (2.1.5), which is formulated in terms of the logarithms of the ratios of the posterior probabilities of the groups.

If the normal linear discriminant rule is used, the log ratios of the posterior probabilities are given by (2.1.6), which has the following sample equivalent for the case of three groups with equal prior probabilities:

$$\hat{c}_{i0}(\mathbf{x}) = \{\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_0)\}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0), \quad i = 1, 2. \quad (2.5.1)$$

The classification rule (2.1.5) with the log ratios estimated by (2.5.1) is equivalent to the rule

$$C(\mathbf{x}) = j \quad \text{if} \quad D_j^2 = \min\{D_i^2, i = 0, 1, 2\},$$

where $D_i^2 = (\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)$, $i = 0, 1, 2$ is the squared sample Mahalanobis distance between \mathbf{x} and the mean vector of the training sample from population Π_i . This is the form in which the classification rule was used in the simulation study.

For the logistic discriminant rule, the log ratios of the posterior probabilities are given by (2.1.8). In a fully polychotomous analysis the parameters β_{0i} and β_{1i} , $i = 1, 2$ are estimated from the training data by means of maximum likelihood.

Many of the readily available statistical software packages do however not offer the facility of a fully polychotomous logistic regression. An alternative strategy that is recommended by Begg and Gray (1984), is to perform a number of individualised binary logistic regression analyses. In the case of three groups, this is done by choosing one of the groups, say Π_0 , as reference group, and performing two separate binary logistic regression analyses involving groups Π_0 and Π_1 , and Π_0 and Π_2 respectively. The parameter estimates obtained in this way in general differ from the estimates obtained when a fully polychotomous analysis is performed. Begg and Gray (1984) studied the asymptotic relative efficiency of the estimates obtained from the individualised approach. They found that these efficiencies are generally high in the case of parameter estimation, but that "occasionally a predicted (posterior) probability will be estimated with a more substantial loss of efficiency". It is therefore not unreasonable to expect these two approaches to yield classification rules that differ with respect to error rates. In the Monte Carlo simulation study, both approaches were investigated. In this section the error rates obtained via the fully polychotomous approach will be used in the comparison of the classification performance of logistic

regression with that of discriminant analysis. In Section 2.6 the error rates obtained by the two approaches to logistic regression will be compared.

As in the two group case, 1000 training data sets were generated at each parameter configuration. For each of these training data sets the actual error rates were estimated by calculating the misclassification rates based on 5000 entities generated from each of the three groups.

2.5.1 THE NORMAL CASE

In the Monte Carlo study for three groups, eight cases were investigated. These cases were obtained by varying the number of feature variables ($k = 2$ and 10), the covariance structure of the variables (using $\Sigma = I$ and Σ given by (2.4.1) with $\rho = 0.9$) and the training sample sizes ($n_0 = n_1 = n_2 = 25$ and $n_0 = n_1 = n_2 = 100$). In the Monte Carlo study for two groups, the relative performance of discriminant analysis and logistic regression was similar in the case of mixed and small sample sizes. Therefore only the small and large sample cases were included in the three group study.

The separation between three groups can be described in terms of three Mahalanobis distances, Δ_{01} , Δ_{02} and Δ_{12} . There are of course many ways in which these distances can be varied. For the purpose of this study, attention was restricted to the equidistant case, with $\Delta_{01} = \Delta_{02} = \Delta_{12} = \Delta$ (say). The following values of Δ^2 were used: $\Delta^2 = 0, 0.5, 1, 1.5, 2, 3$ and 4 . To achieve these distances in the case of uncorrelated feature variables, the elements of μ_0 , μ_1 and μ_2 were chosen as follows:

$$\mu_{0j} = 0, \quad j = 1, \dots, k,$$

$$\mu_{11} = \Delta \quad \text{and} \quad \mu_{1j} = 0, \quad j = 2, \dots, k,$$

$$\mu_{21} = \Delta/2 \quad \text{and} \quad \mu_{2j} = \sqrt{3} \Delta / (2\sqrt{k-1}), \quad j = 2, \dots, k.$$

In the equicorrelated case, as in (2.4.1), the following choices were made:

$$\mu_{0j} = 0, \quad j = 1, \dots, k,$$

$$\mu_{11} = \Delta \quad \text{and} \quad \mu_{1j} = 0, \quad j = 2, \dots, k,$$

$$\mu_{21} = a \quad \text{and} \quad \mu_{2j} = b, \quad j = 2, \dots, k,$$

with

$$a = \Delta/2 - (k-1)\sigma^{ij}b/\sigma^{ii}$$

$$b = \sqrt{\frac{\frac{3}{4}\Delta^2\sigma^{ii}}{(k-1)\sigma^{ij}\{1 - [(k-1)(\sigma^{ij})^2/(\sigma^{ii})^2] + [(k-2)\sigma^{ij}/\sigma^{ii}]\}}}$$
 (2.5.1.1)

In these equations σ^{ii} represents any diagonal element of Σ^{-1} (for Σ defined as in (2.4.1), all the diagonal elements of Σ^{-1} are equal) and σ^{ij} represents any off-diagonal element of Σ^{-1} (all the off-diagonal elements are equal). Data were generated from the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ by means of the IMSL routine DRNMVN, and the relevant components of μ_1 and μ_2 were added to the data from groups Π_1 and Π_2 .

A selection of boxplots of the simulation output is given in Figs. 2.15 - 2.18. On each graph the following coding is used to denote the respective actual error rates of discriminant analysis and logistic regression for the eight different cases: DA_1 and LR_1 for small samples ($n_0 = n_1 = 25$) and $k = 2$; DA_2 and LR_2 for large samples ($n_0 = n_1 = 100$) and $k = 2$; DA_3 and LR_3 for small samples and $k = 10$ and DA_4 and LR_4 for large samples and $k = 10$. Tables 2.7 and 2.8 contain the means and standard deviations of the actual error rates. The conclusions drawn from investigation of these graphs are similar to those in the two group normal case. The only cases where a significant difference between the error rates of discriminant analysis and logistic regression is observed, occur in small sample cases with $k = 10$, at moderate to large separation between the populations ($\Delta^2 = 2, 3$ and 4) (see Figs. 2.16 and 2.18 for cases where $\Delta^2 = 3$). For normal feature data, the use of the normal linear discriminant rule is recommended, since it never performs significantly worse than the logistic discriminant rule, and significantly outperforms it in some cases.

As in the two group normal case, the introduction of correlation between the feature variables had little effect on the error rates. The influence of an increase in the sample size and a change in the number of feature variables is the same as in the two group case. As is to be expected, a comparison of the error rates of corresponding two group and three group cases, shows that the error rates are larger for three groups.

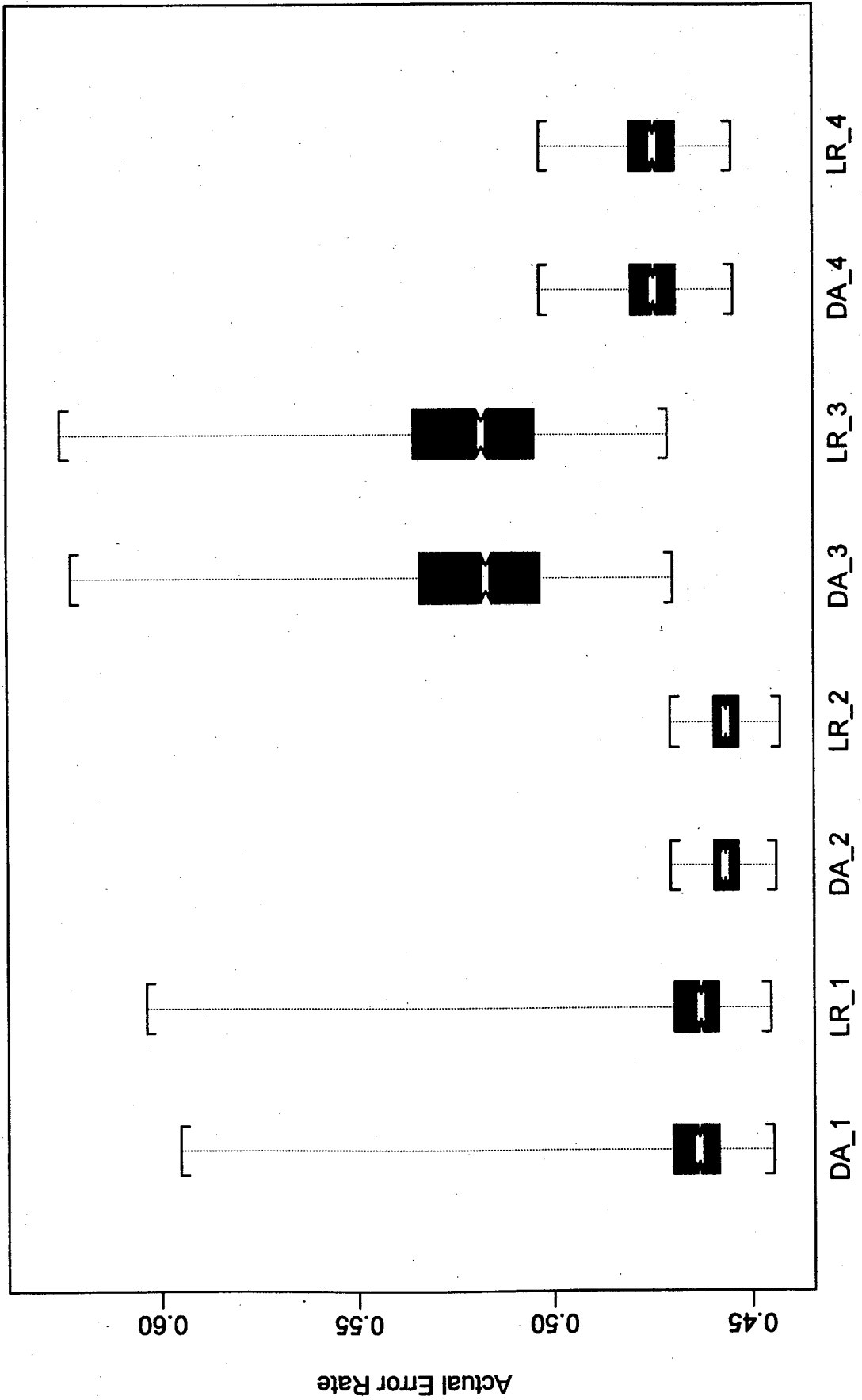


FIG. 2.15: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, NORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1

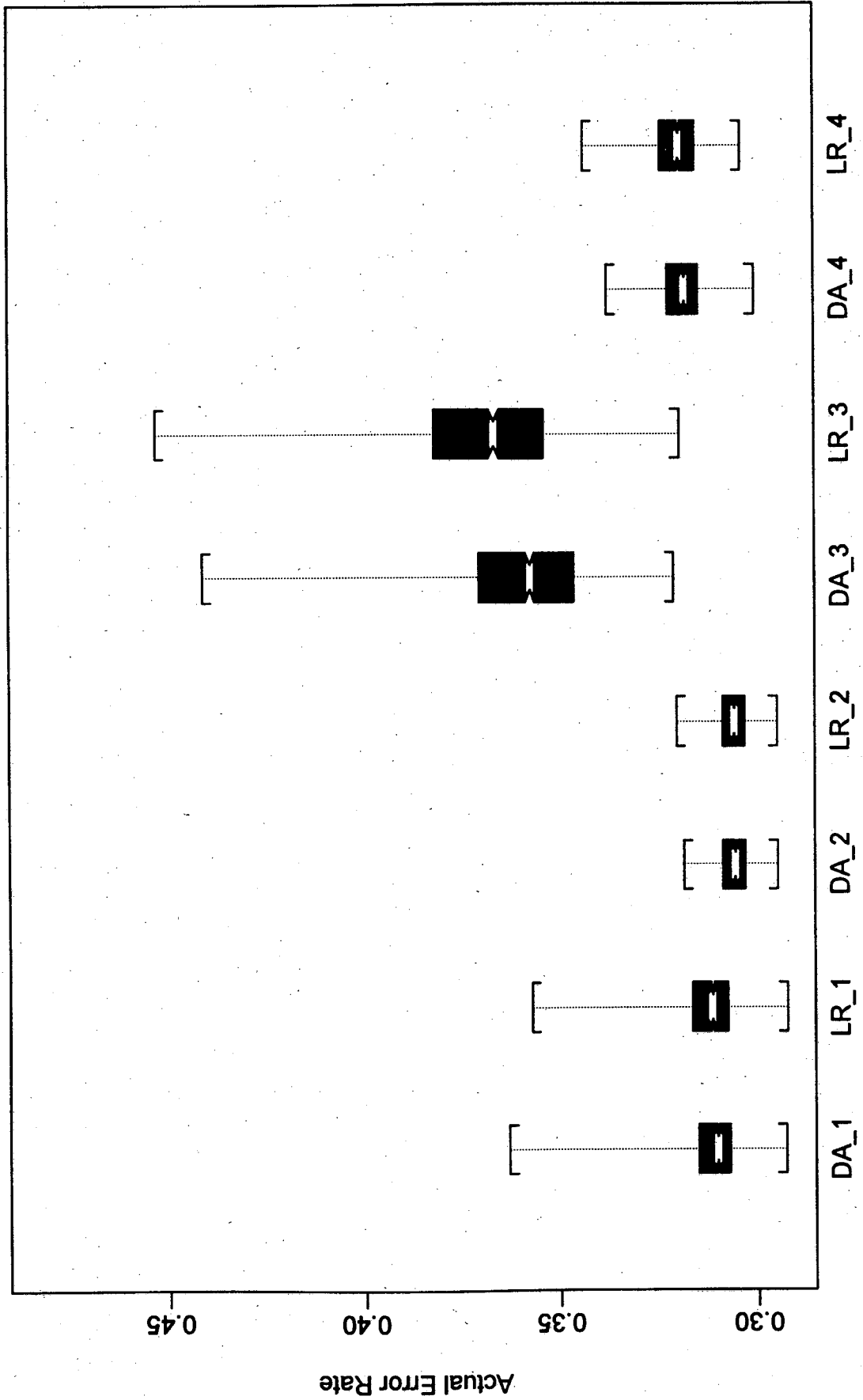


FIG. 2.16: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, NORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 3

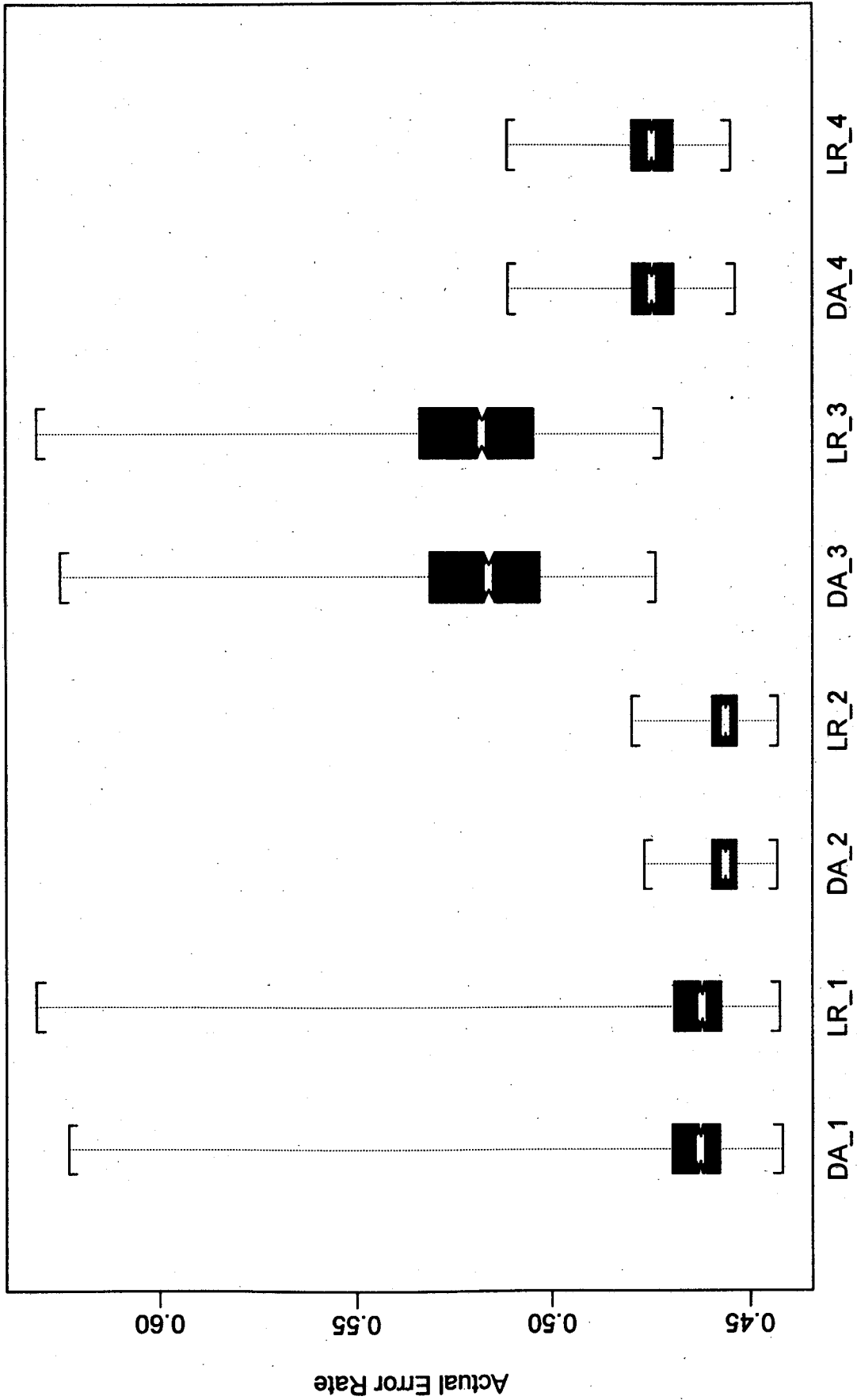


FIG. 2.17: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, NORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 1

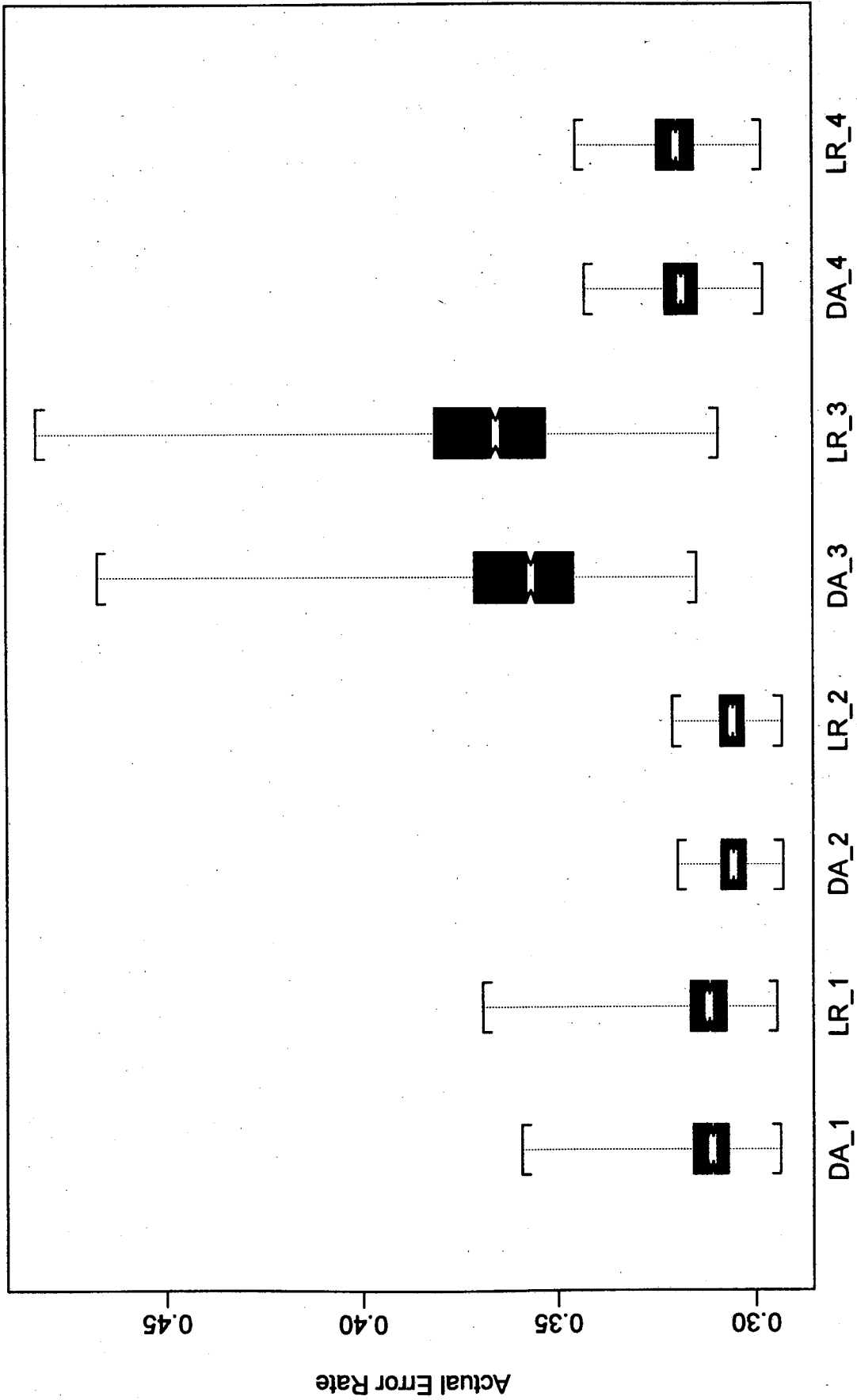


FIG. 2.18: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, NORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 3

TABLE 2.7 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES THREE GROUPS, NORMAL DATA ($\rho = 0$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66658 (.00372)	.66659 (.00372)	.66655 (.00366)	.66654 (.00365)
1	.46576 (.01351)	.46541 (.01341)	.45637 (.00463)	.45640 (.00464)
2	.37514 (.00797)	.37555 (.00806)	.36844 (.00420)	.36854 (.00423)
3	.31156 (.00739)	.31256 (.00789)	.30568 (.00399)	.30589 (.00405)
4	.26257 (.00657)	.26448 (.00775)	.25671 (.00374)	.25710 (.00382)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66661 (.00378)	.66661 (.00381)	.66668 (.00387)	.66670 (.00388)
1	.51981 (.02249)	.52109 (.02283)	.47469 (.00823)	.47486 (.00830)
2	.42799 (.01996)	.43284 (.02096)	.38362 (.00708)	.38437 (.00729)
3	.35948 (.01847)	.36876 (.02417)	.31846 (.00597)	.31970 (.00635)
4	.30926 (.01891)	.32226 (.03295)	.26825 (.00567)	.27044 (.00625)

TABLE 2.8 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES THREE GROUPS, NORMAL DATA ($\rho = 0.9$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66690 (.00375)	.66690 (.00377)	.66656 (.00377)	.66654 (.00377)
1	.46484 (.01221)	.46456 (.01208)	.45650 (.00474)	.45649 (.00470)
2	.37545 (.00817)	.37586 (.00821)	.36838 (.00431)	.36848 (.00434)
3	.31185 (.00736)	.31285 (.00807)	.30557 (.00417)	.30578 (.00415)
4	.26263 (.00710)	.26439 (.00841)	.25692 (.00395)	.25720 (.00408)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66659 (.00382)	.66654 (.00383)	.66682 (.00384)	.66684 (.00386)
1	.51879 (.02217)	.52074 (.02263)	.47473 (.00804)	.47489 (.00805)
2	.42686 (.02050)	.43204 (.02199)	.38295 (.00701)	.38371 (.00708)
3	.35953 (.01932)	.36829 (.02152)	.31861 (.00630)	.31985 (.00665)
4	.30877 (.01832)	.32315 (.03494)	.26811 (.00572)	.27006 (.00630)

2.5.2 THE DOUBLE EXPONENTIAL CASE

The methods described in Section 2.4.2 for the two group double exponential case with uncorrelated and correlated feature variables respectively, were also used to generate data for the three group double exponential case. The same eight cases included in the study of the three group normal case were investigated and the same values of Δ^2 were used. The required separation between the groups was obtained by using the parameterisation described in Section 2.5.1 for uncorrelated and correlated feature variables respectively.

The actual error rates were summarised by means of boxplots, of which a selection appears in Figs. 2.19 - 2.22. Tables 2.9 and 2.10 contain the means and standard deviations of the actual error rates.

As in the two group double exponential case, there is little difference between the error rates of the two techniques, except in the small sample cases with $k = 10$. In these cases linear discriminant analysis significantly outperformed logistic regression at moderate to large values of Δ^2 . This effect is somewhat more pronounced when the feature variables are correlated. The error rates are smaller in the correlated cases than in the corresponding cases with uncorrelated feature variables. The error rates are also smaller than the error rates in corresponding normal cases.

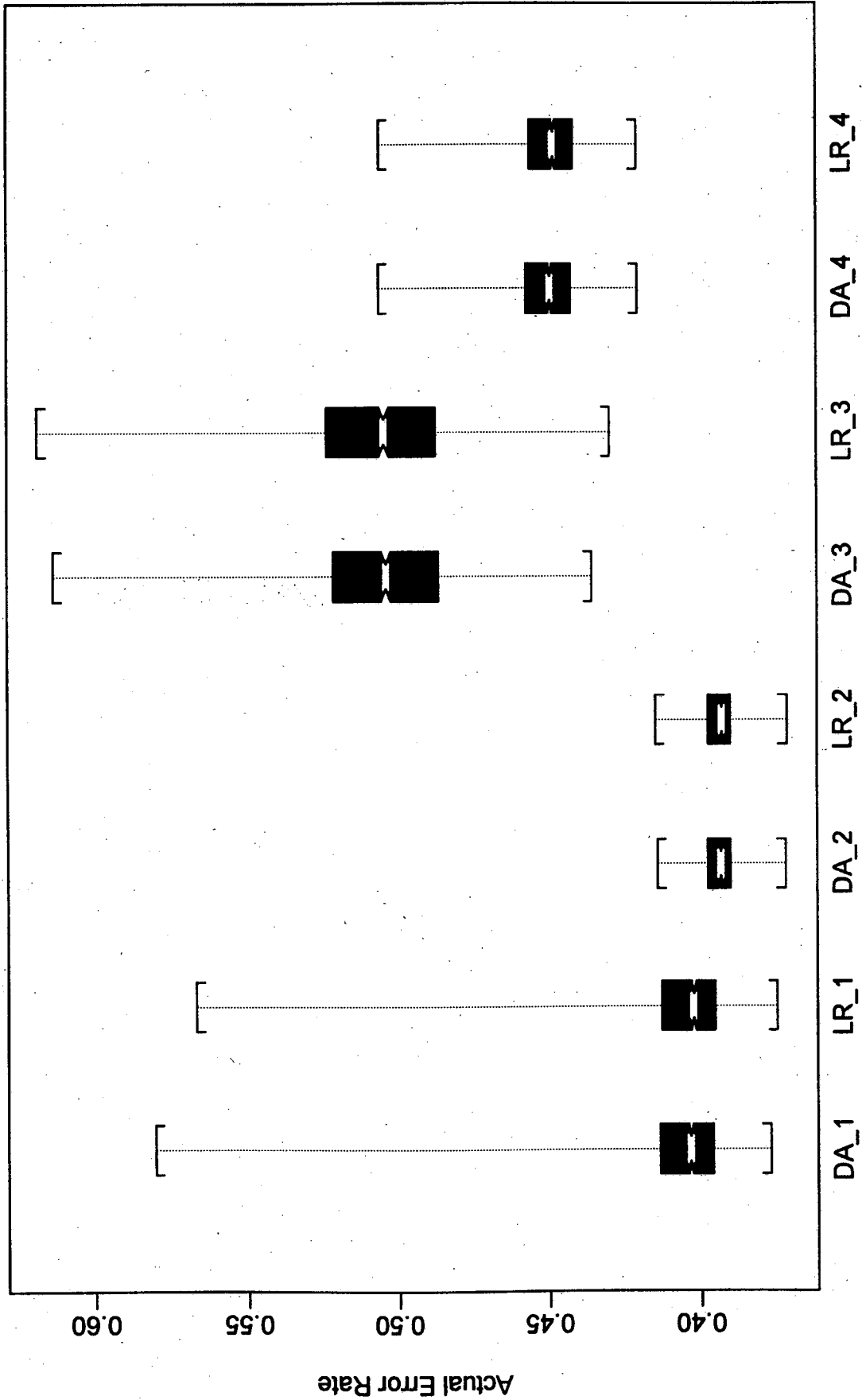


FIG. 2.19: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, DOUBLE EXPONENTIAL DATA
CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1

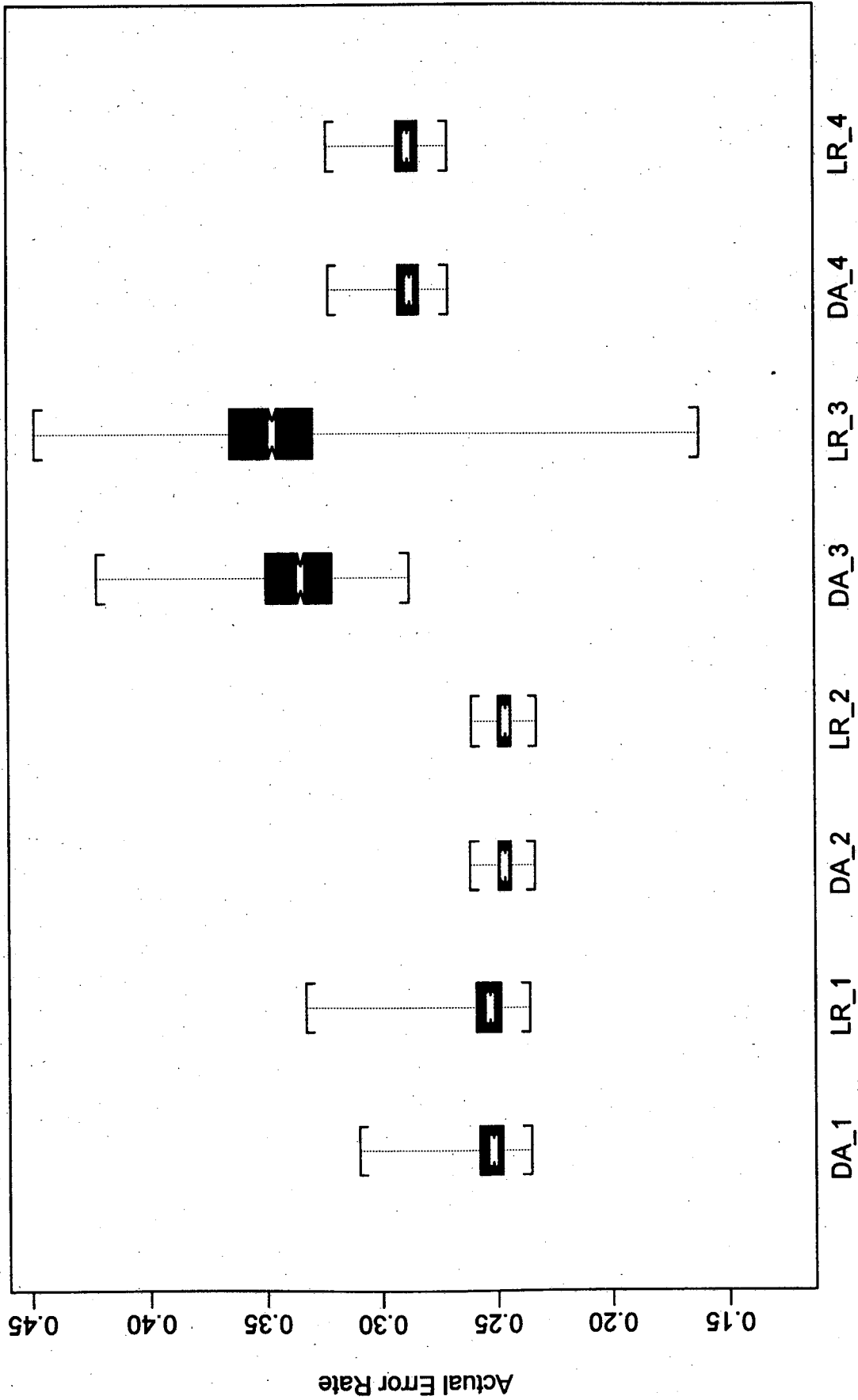


FIG. 2.20: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, DOUBLE EXPONENTIAL DATA
CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 3

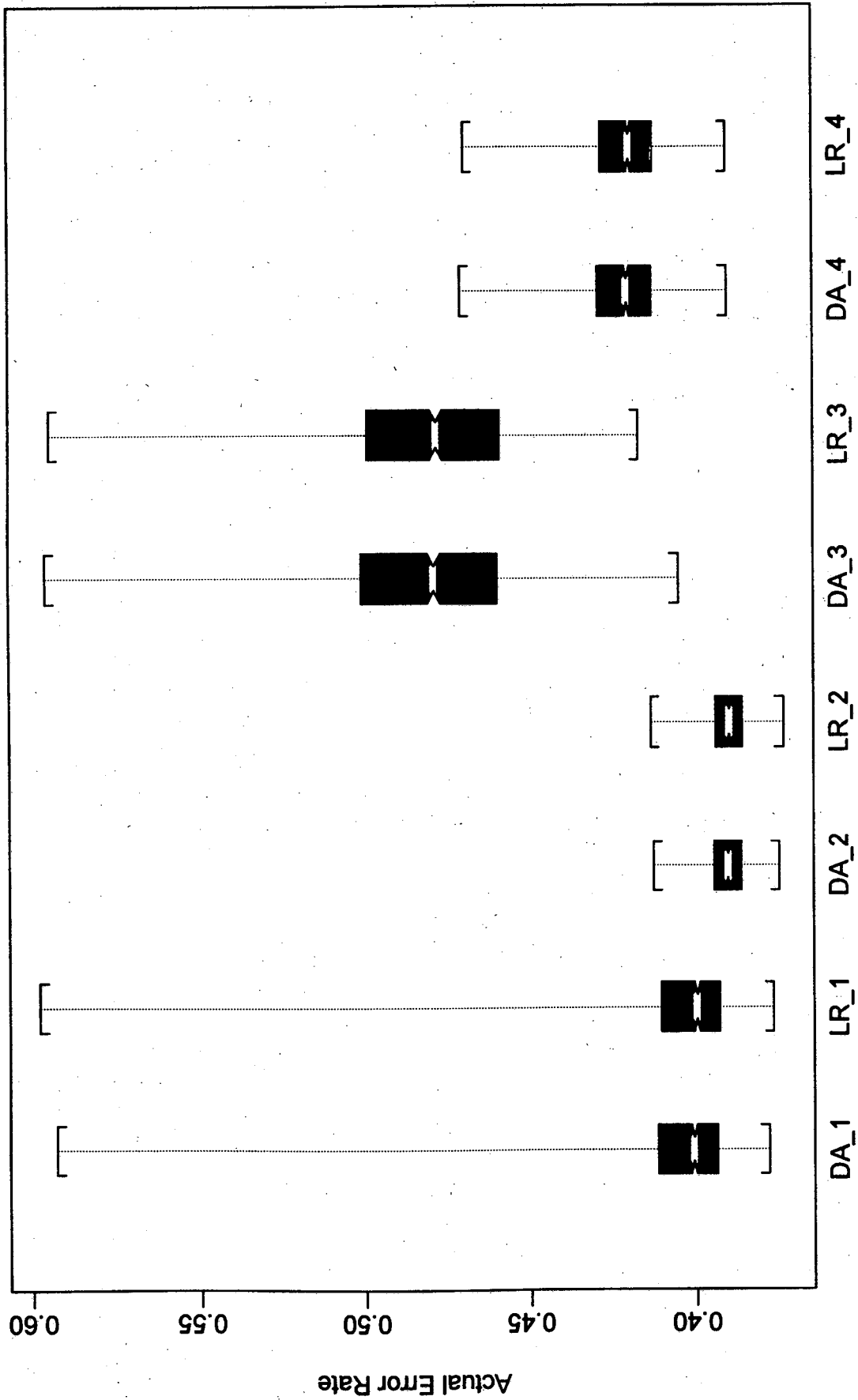


FIG. 2.21: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, DOUBLE EXPONENTIAL DATA
CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 1

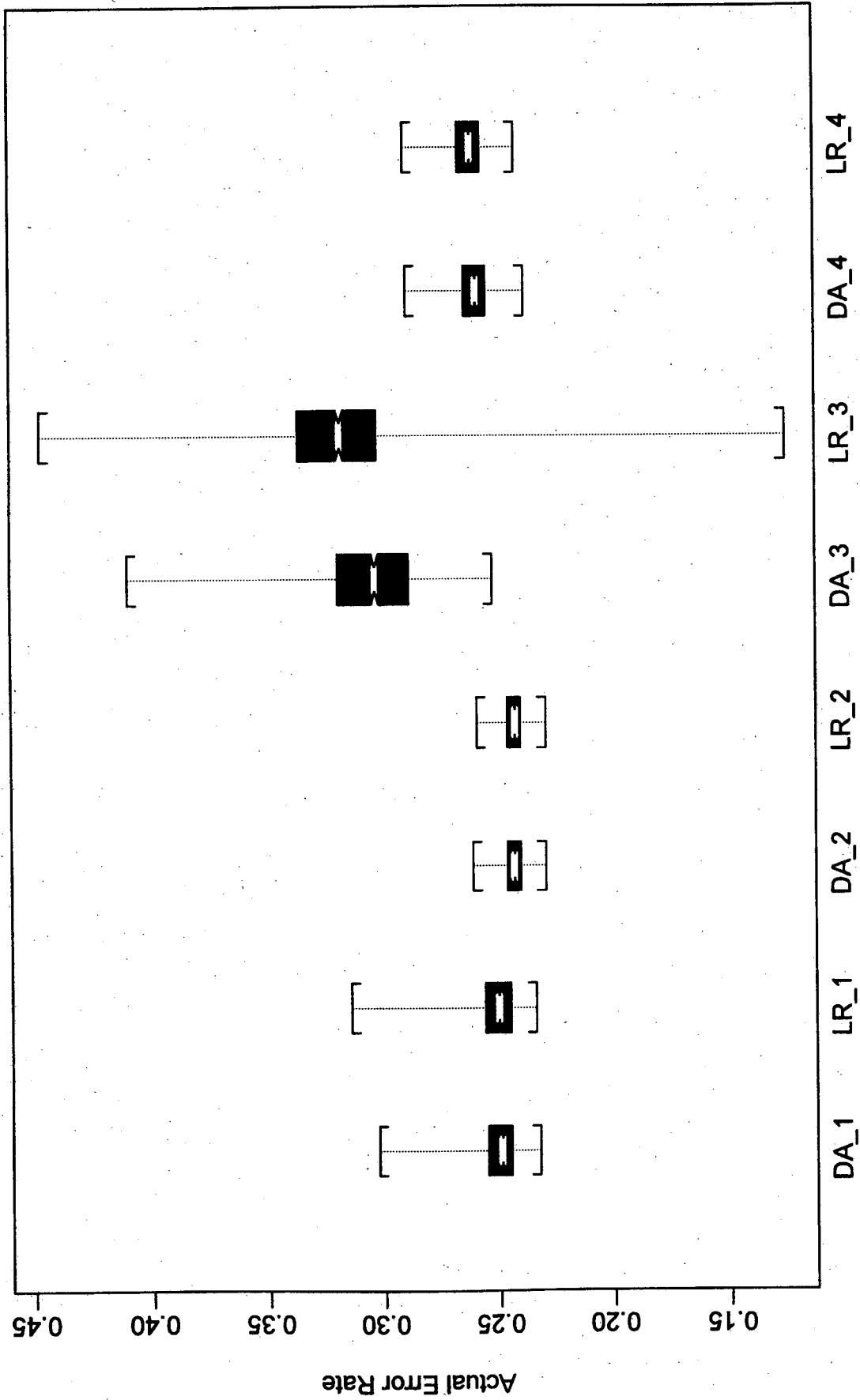


FIG. 2.22: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, DOUBLE EXPONENTIAL DATA
CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 3

TABLE 2.9 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES THREE GROUPS, DOUBLE EXPONENTIAL DATA ($\rho = 0$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66675 (.00370)	.66677 (.00373)	.66645 (.00367)	.66643 (.00369)
1	.40690 (.01709)	.40606 (.01659)	.39401 (.00562)	.39374 (.00569)
2	.31277 (.01032)	.31253 (.01069)	.30327 (.00491)	.30312 (.00508)
3	.25360 (.00859)	.25483 (.00971)	.24646 (.00423)	.24642 (.00433)
4	.21197 (.00720)	.21415 (.00863)	.20568 (.00378)	.20607 (.00391)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66687 (.00368)	.66687 (.00366)	.66682 (.00387)	.66683 (.00386)
1	.50582 (.02626)	.50656 (.02712)	.45017 (.01143)	.44921 (.01144)
2	.40574 (.02492)	.41117 (.02624)	.35216 (.00908)	.35176 (.00905)
3	.33618 (.02249)	.34842 (.02697)	.28768 (.00700)	.28840 (.00722)
4	.28501 (.02174)	.30365 (.03170)	.23971 (.00637)	.24194 (.00674)

TABLE 2.10 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES THREE GROUPS, DOUBLE EXPONENTIAL DATA ($\rho = 0.9$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66666 (.00383)	.66663 (.00383)	.66645 (.00384)	.66643 (.00386)
1	.40494 (.01768)	.40388 (.01734)	.39090 (.00628)	.39040 (.00608)
2	.31019 (.01131)	.30975 (.01145)	.29993 (.00498)	.29968 (.00488)
3	.25063 (.00875)	.25174 (.00988)	.24302 (.00439)	.24302 (.00444)
4	.20971 (.00740)	.21191 (.00890)	.20257 (.00371)	.20301 (.00384)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66661 (.00384)	.66662 (.00396)	.66663 (.00376)	.66663 (.00377)
1	.48186 (.02970)	.48069 (.02955)	.42150 (.01214)	.42103 (.01192)
2	.37672 (.02760)	.38262 (.02829)	.32059 (.00864)	.32130 (.00871)
3	.30597 (.02374)	.32056 (.02785)	.25941 (.00704)	.26158 (.00742)
4	.25732 (.02073)	.27819 (.03368)	.21630 (.00592)	.22008 (.00710)

2.5.3 THE LOGNORMAL CASE

The Johnson translation system, described in Section 2.4.3 for the two group lognormal case, was used to generate data for the three group lognormal case. The same eight cases included in the study of the three group normal and double exponential cases were investigated, and the parameterisation described in Section 2.5.1, was used to obtain the required separation between the groups.

For uncorrelated feature variables, the following choices of the parameters λ_{ij} and ξ_{ij} were made:

$$\lambda_{ij} = 1/\sqrt{e^2 - e}, \quad i = 0, 1, 2; j = 1, 2, \dots, k,$$

$$\xi_{0j} = -1/\sqrt{e-1}, \quad j = 1, 2, \dots, k,$$

$$\xi_{11} = \Delta - 1/\sqrt{e-1} \text{ and } \xi_{1j} = -1/\sqrt{e-1}, \quad j = 2, \dots, k,$$

$$\xi_{21} = \Delta/2 - 1/\sqrt{e-1} \text{ and } \xi_{2j} = \sqrt{3\Delta}/(2\sqrt{k-1}) - 1/\sqrt{e-1}, \quad j = 2, \dots, k.$$

These choices yield lognormal variables with

$$\mu_{0j} = 0, \quad j = 1, \dots, k,$$

$$\mu_{11} = \Delta, \quad \mu_{1j} = 0, \quad j = 2, \dots, k,$$

$$\mu_{21} = \Delta/2, \quad \mu_{2j} = \sqrt{3\Delta}/(2\sqrt{k-1}), \quad j = 2, \dots, k,$$

$$\text{and } \sigma_{ij}^2 = 1, \quad j = 1, 2, \dots, k.$$

For correlated feature variables, the parameters λ_{ij} and ξ_{ij} were chosen as follows:

$$\lambda_{ij} = 1/\sqrt{e^2 - e}, \quad i = 0, 1, 2; j = 1, 2, \dots, k,$$

$$\xi_{0j} = -1/\sqrt{e-1}, \quad j = 1, \dots, k,$$

$$\xi_{11} = \Delta - 1/\sqrt{e-1} \text{ and } \xi_{1j} = -1/\sqrt{e-1}, \quad j = 2, \dots, k,$$

$$\xi_{21} = a - 1/\sqrt{e-1} \text{ and } \xi_{2j} = b - 1/\sqrt{e-1}, \quad j = 2, \dots, k,$$

with a and b given by (2.5.1.1).

These choices yield lognormal variables with

$$\mu_{0j} = 0, \quad j = 1, \dots, k,$$

$$\mu_{11} = \Delta \quad \text{and} \quad \mu_{1j} = 0, \quad j = 2, \dots, k,$$

$$\mu_{21} = a \quad \text{and} \quad \mu_{2j} = b, \quad j = 2, \dots, k.$$

As in the two group lognormal case, the IMSL routine DRNMVN was used to generate multivariate normal variables with mean $\mathbf{0}$ and covariance matrix Σ given by (2.4.1). A ρ -value of 0.935 for the normal variables yielded lognormal variables with covariance matrix given by (2.4.1) with ρ approximately equal to 0.9.

A selection of boxplots of the simulation output of the three group lognormal case is given in Figs. 2.23 - 2.26, and Tables 2.11 and 2.12 provide the means and standard deviations of the actual error rates. The following conclusions are made.

The error rates of the logistic discriminant rule are significantly lower than that of the normal linear discriminant rule for small to moderate values of Δ^2 (see Figs. 2.23 and 2.25 for cases where $\Delta^2 = 1$). The difference seems to decrease with increasing separation between the groups. For large values of Δ^2 , ($\Delta^2 = 4$), discriminant analysis outperformed logistic regression in the small sample case with $k = 10$ (see Figs. 2.24 and 2.26 for cases where $\Delta^2 = 4$). Logistic regression should therefore be used for lognormal data, except in cases where the number of variables is large relative to the sample size. The effect of the presence of correlation is the same as in the two group lognormal case. The error rates obtained in cases where the feature variables are correlated, are markedly lower than the error rates of the corresponding cases with uncorrelated feature variables, especially for $k = 10$.

The problem of non-existence of maximum likelihood estimates of the parameters of the logistic discriminant function was more prevalent in the three group lognormal case than in any of the other cases included in the study, occurring as much as 20% of the time at $\Delta^2 = 4$. The reason for this is that, due to the shape of the lognormal distribution, complete separation between populations will be more likely to occur at any given value of Δ^2 , than in the case of the normal or double exponential distributions.

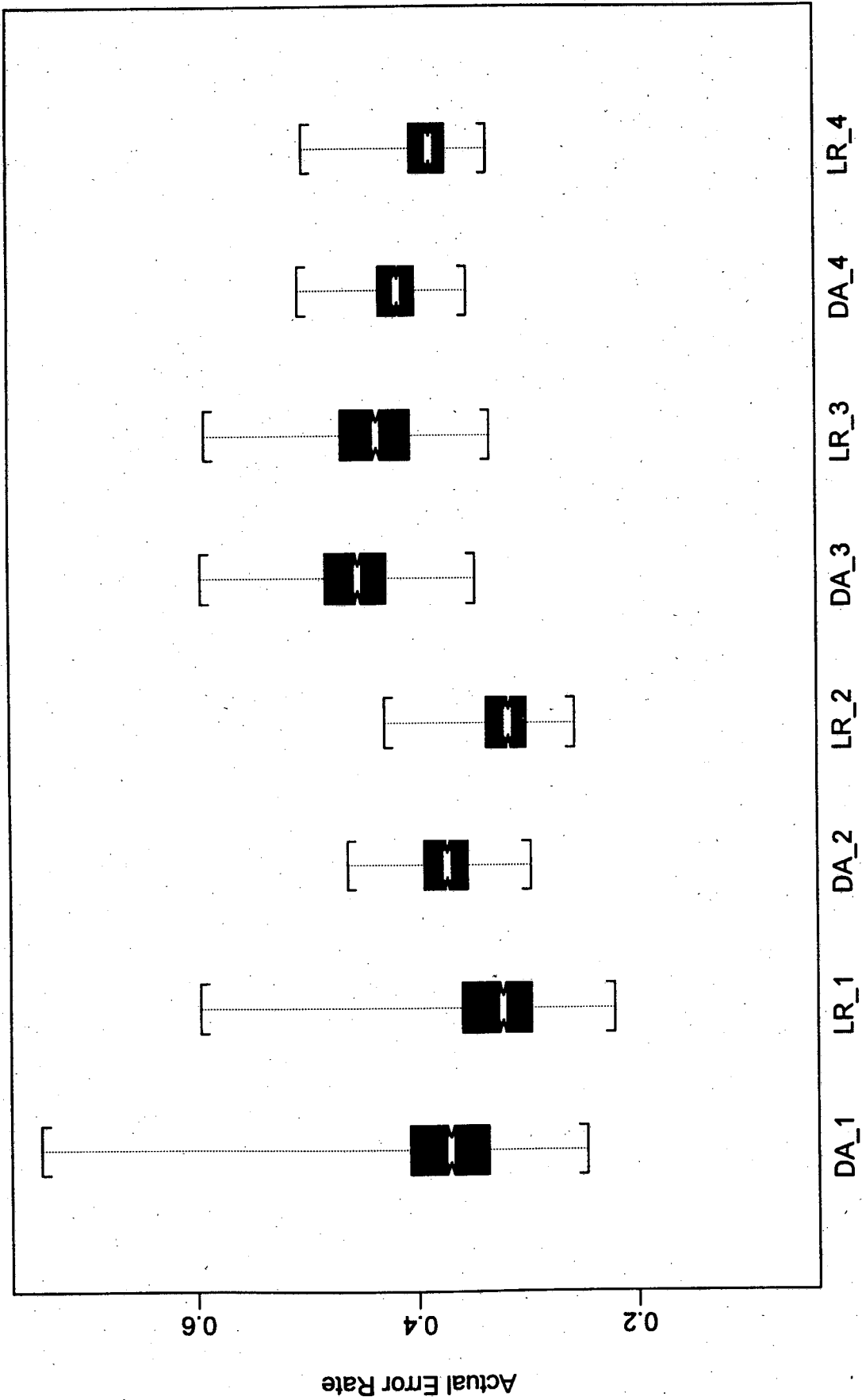


FIG. 2.23: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, LOGNORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1

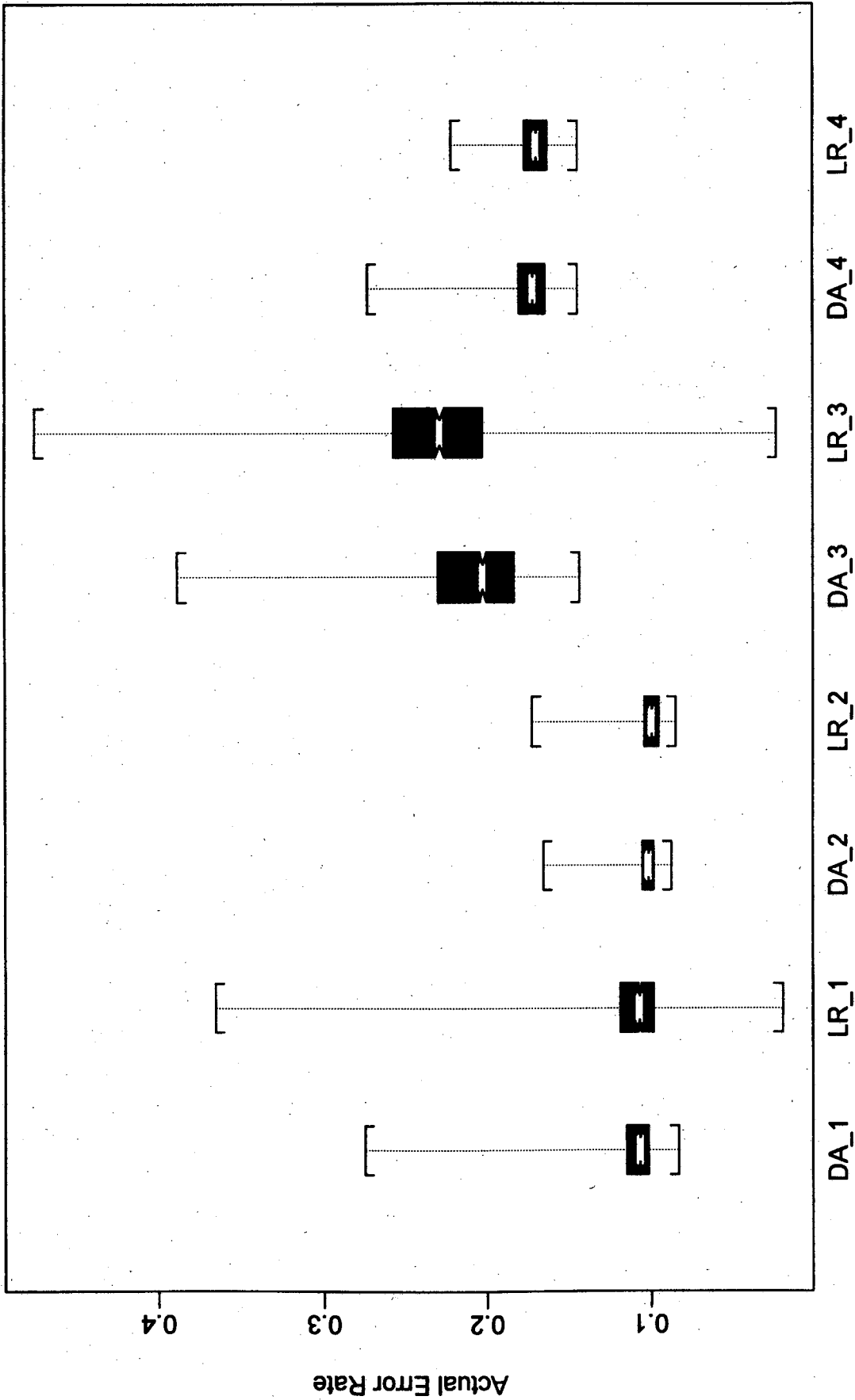


FIG. 2.24: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, LOGNORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 4

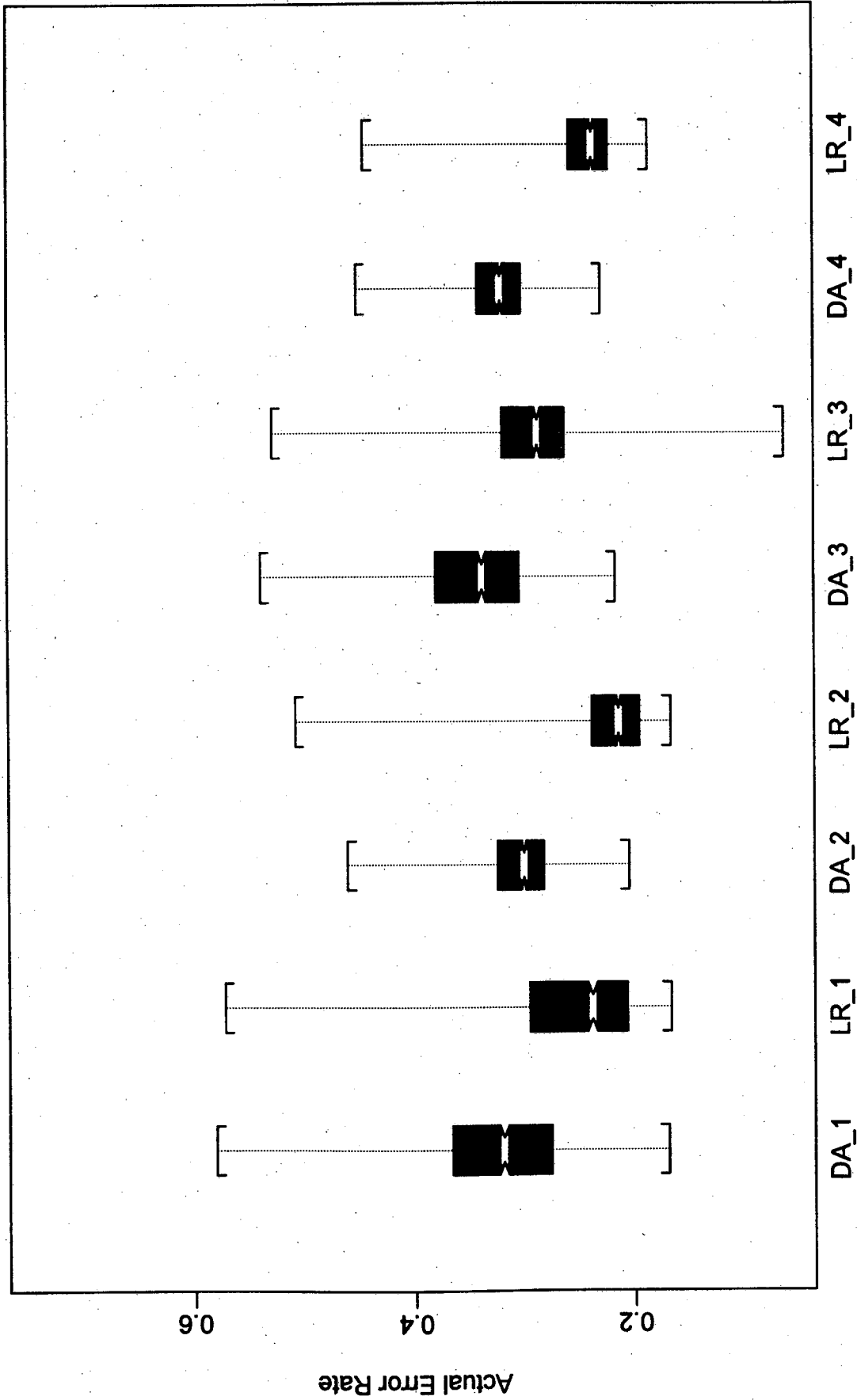


FIG. 2.25: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, LOGNORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 1

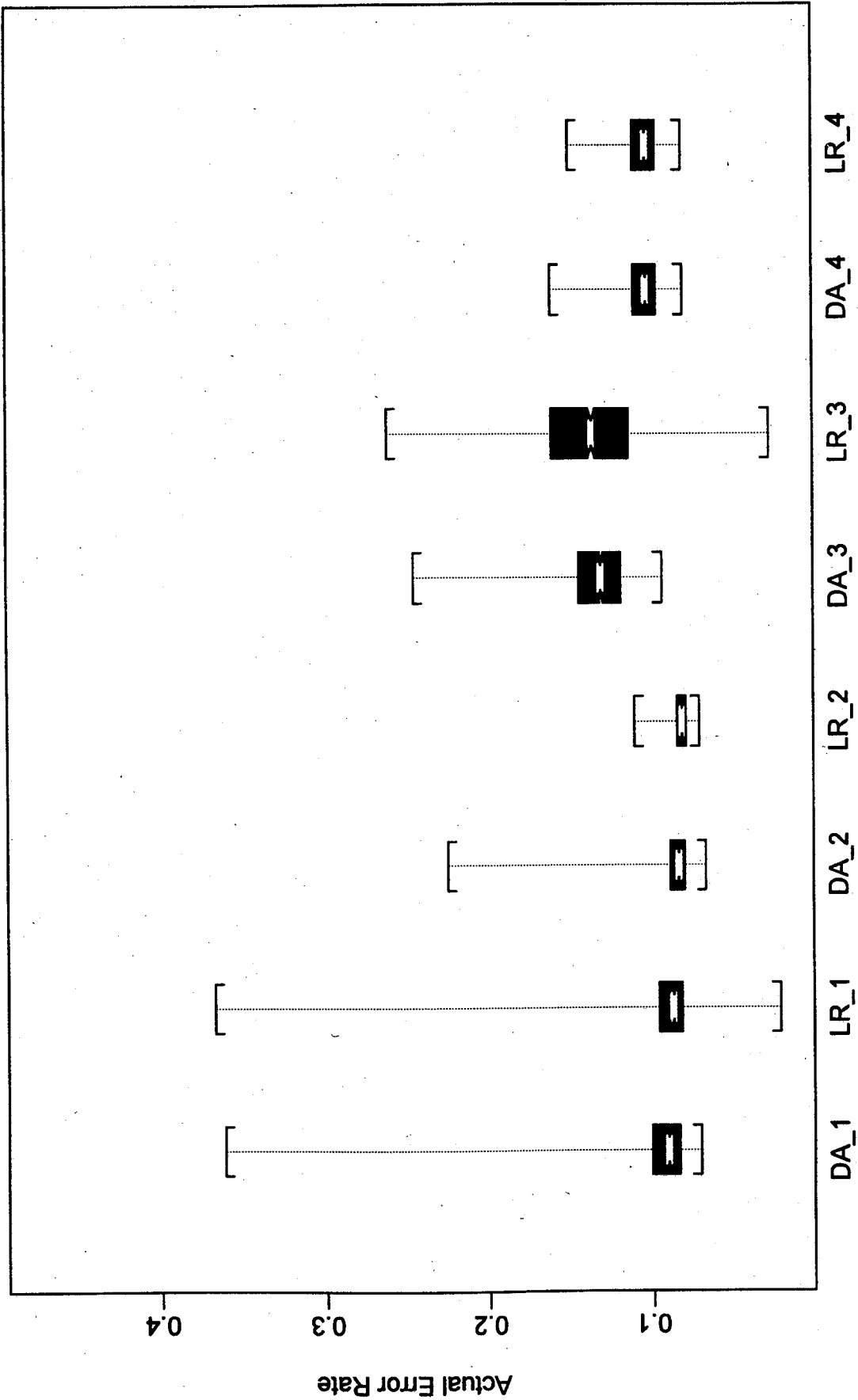


FIG. 2.26: ACTUAL ERROR RATES OF DA AND LR, 3 GROUPS, LOGNORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 4

TABLE 2.11 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES THREE GROUPS, LOGNORMAL DATA ($\rho = 0$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66663 (.00351)	.66659 (.00361)	.66660 (.00368)	.66659 (.00366)
1	.36973 (.04905)	.33146 (.05113)	.37315 (.02845)	.32058 (.03063)
2	.22244 (.04655)	.20163 (.03776)	.20545 (.02512)	.18267 (.01812)
3	.14846 (.03538)	.14376 (.03106)	.13109 (.01481)	.12748 (.01244)
4	.11245 (.02360)	.11203 (.02599)	.10233 (.00730)	.10026 (.00780)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66683 (.00370)	.66687 (.00366)	.66664 (.00372)	.66666 (.00375)
1	.45387 (.04402)	.43835 (.04605)	.41617 (.02460)	.38810 (.02517)
2	.34215 (.04914)	.33541 (.04637)	.29299 (.02469)	.27303 (.01681)
3	.26606 (.04710)	.27020 (.05534)	.22074 (.01876)	.21130 (.01341)
4	.21205 (.04328)	.22721 (.05602)	.17390 (.01454)	.17058 (.01083)

TABLE 2.12 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES THREE GROUPS, LOGNORMAL DATA ($\rho = 0.9$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66659 (.00352)	.66658 (.00355)	.66662 (.00356)	.66664 (.00362)
1	.31890 (.07225)	.26258 (.07788)	.30268 (.04101)	.22394 (.04442)
2	.18433 (.06111)	.14865 (.04240)	.15391 (.03362)	.12411 (.00709)
3	.12644 (.04382)	.11221 (.03052)	.10393 (.01703)	.09810 (.00525)
4	.09928 (.03019)	.09241 (.02207)	.08568 (.01064)	.08219 (.00414)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	DA	LR	DA	LR
0	.66685 (.00367)	.66677 (.00379)	.66675 (.00367)	.66675 (.00363)
1	.33846 (.05711)	.29040 (.05400)	.31819 (.03187)	.24268 (.03090)
2	.21216 (.04535)	.20121 (.04764)	.17495 (.02881)	.15299 (.01300)
3	.15838 (.02867)	.16286 (.04573)	.12351 (.01393)	.12170 (.01048)
4	.13353 (.02209)	.13677 (.03624)	.10347 (.01022)	.10414 (.01009)

2.6 COMPARISON OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED BINARY LOGISTIC REGRESSION

In the Monte Carlo simulation study investigating the classification performance of discriminant analysis and logistic regression for the three group case, the parameters β_{0i} and β_{1i} , $i = 1, 2$, of the logistic discriminant rule were estimated in two different ways. Firstly, a fully polychotomous analysis was performed, in which the parameters were estimated from the training data by means of maximum likelihood. Secondly, the strategy recommended by Begg and Gray (1984), in which two separate binary logistic regressions were performed to obtain estimates of the parameters, was implemented. In this section, the error rates of the discriminant rules obtained from these two methods, will be compared.

A representative selection of boxplots of the error rates attained by the logistic discriminant rules obtained from the fully polychotomous approach (coded FP on the graphs) and the individualised binary logistic regressions (coded IR on the graphs) appears in Figs. 2.27 - 2.28 for normal feature variables, in Figs. 2.29 - 2.30 for double exponential feature variables and in Figs. 2.31 - 2.34 for the lognormal case. Tables 2.13 - 2.16 contain the averages and the standard deviations of the logistic discriminant rule actual error rates for both approaches.

If logistic regression is used for the classification of entities into more than two available groups, the fully polychotomous approach should strictly be used. However, as pointed out by Begg and Gray (1984), and by Hosmer and Lemeshow (1989), the unavailability of software to implement this approach might necessitate use of the alternative approach based on individualised binary logistic regressions. An important question that deserves attention is: what price is paid in terms of classification performance if this alternative approach is used? An inspection of Figs. 2.27 - 2.34 and the entries in Tables 2.13 - 2.16 provide a partial answer to this question for the cases of normal, double exponential and lognormal feature variables.

Consider first the normal case. Since the correlated case is very similar to the uncorrelated case, only the latter is represented in Figs. 2.27 and 2.28 and in Table 2.13. In most cases the difference in error rates is very small, except for the small sample case with $k = 10$, where the fully polychotomous approach is significantly better. In general therefore, for the cases considered in the Monte Carlo study, using the individualised approach will lead to a significant deterioration in classification performance only when the number of variables becomes large relative to the sample size. It should be noted that these are exactly the previously identified cases where the fully polychotomous logistic regression generally has a significantly larger error rate than the normal linear discriminant rule. Therefore the practitioner who adopts the individualised approach in these cases, is in fact using an inferior classification rule. Another disadvantage of the individualised approach is that the error rates are highly variable, especially in the small sample cases at large values of Δ^2 . This accounts for the apparent contradiction in conclusions reached when considering medians and

averages of the actual error rates (see Fig. 2.28 for cases FP_3 and IR_3, and the corresponding entries in Table 2.13).

For the double exponential case, the simulation output for the correlated and uncorrelated cases is very similar, and therefore only the uncorrelated case is represented in the graphs and table. Perusal of Figs. 2.29 - 2.30 and Table 2.14 for the double exponential case, shows that the conclusions reached above are also valid here.

The results displayed in Figs. 2.31 - 2.34 and in Tables 2.15 - 2.16 for the lognormal case are much more erratic. Consider first the small sample cases. For uncorrelated feature variables the individualised approach outperforms the fully polychotomous approach, especially at larger values of Δ^2 (see Figs. 2.31 and 2.32 and Table 2.15). In the case of correlated feature variables, this trend is reversed for $k=2$, but in general not for $k=10$, except at $\Delta^2=4$ (see Figs. 2.33 and 2.34 and Table 2.16). It is difficult to offer an explanation for this behaviour. For large samples, the approaches are practically equivalent when the feature variables are uncorrelated. For correlated feature variables, the fully polychotomous approach performs better. This is true for cases with $k=2$ and $k=10$.

In conclusion, for normal and double exponential data, the fully polychotomous approach is preferable to the individualised approach, but in these cases the normal linear discriminant rule outperforms polychotomous logistic regression and should be the method of choice. For the lognormal case, where polychotomous logistic regression often outperforms the normal linear discriminant rule, there are a number of configurations for which the binary approach should be the method of choice. It should also be mentioned that the problem of non-existence of the maximum likelihood estimates of the parameters at large separations between lognormal populations is appreciably more serious in the fully polychotomous approach.

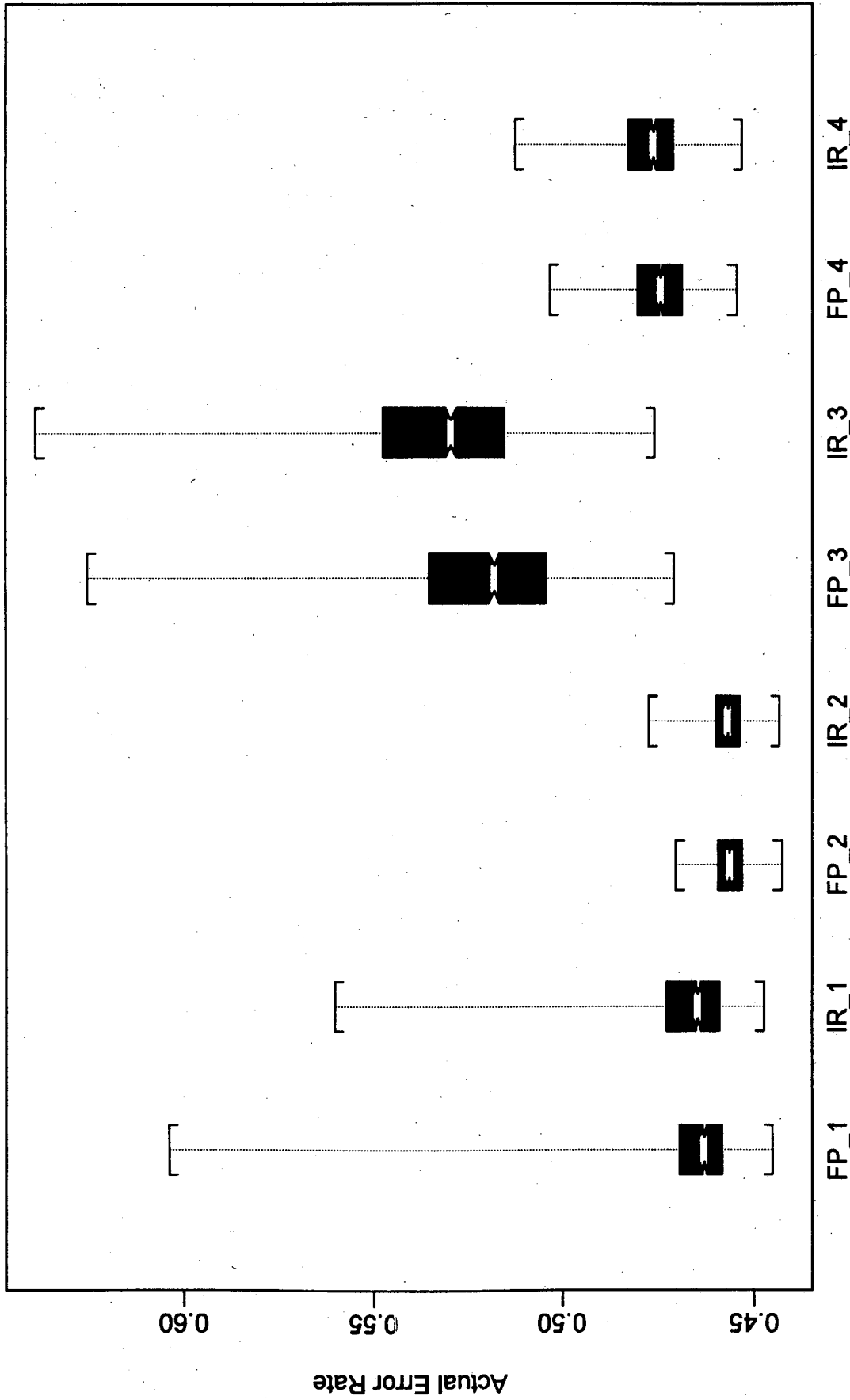


FIG. 2.27: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, NORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1

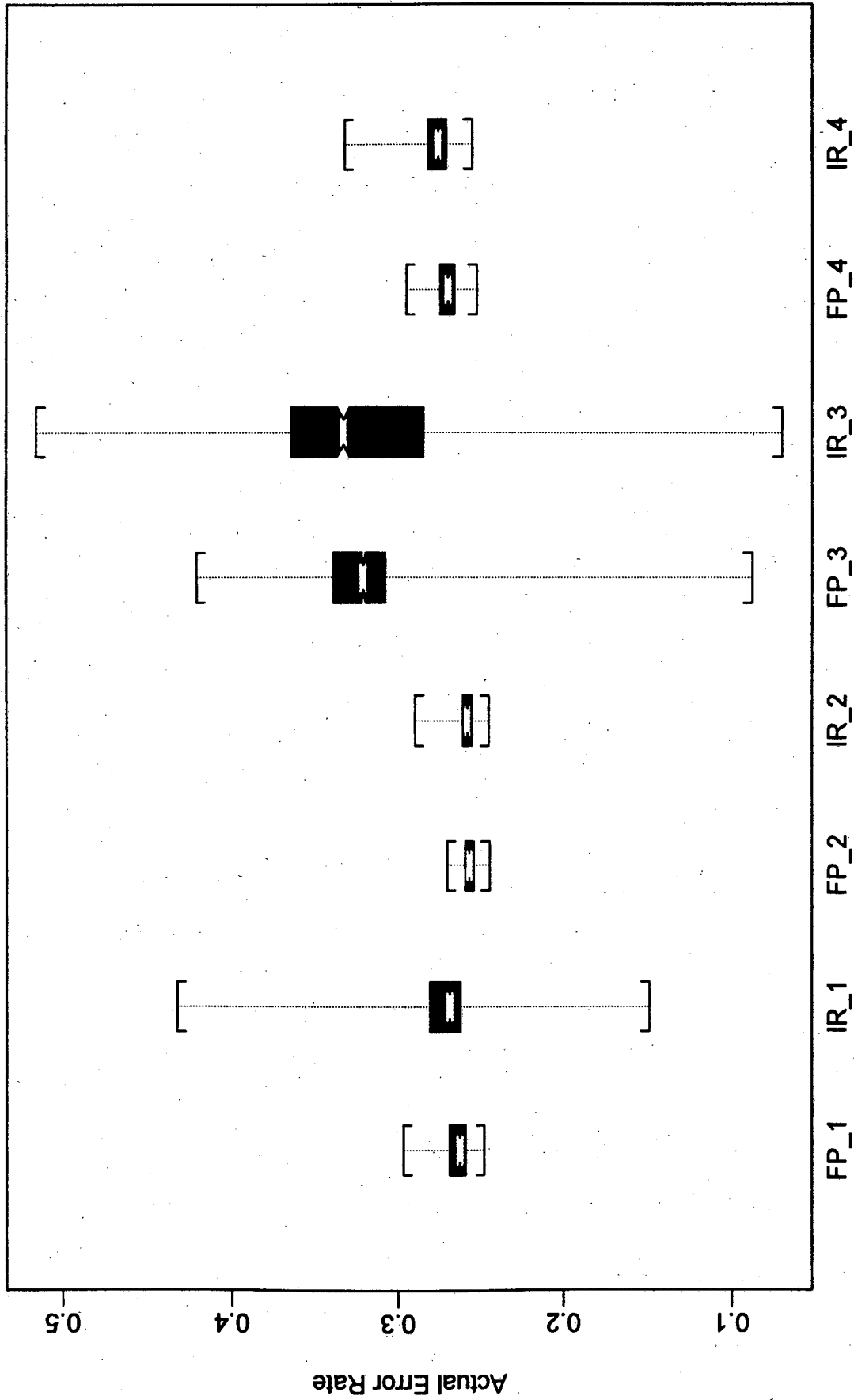


FIG. 2.28: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, NORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 4

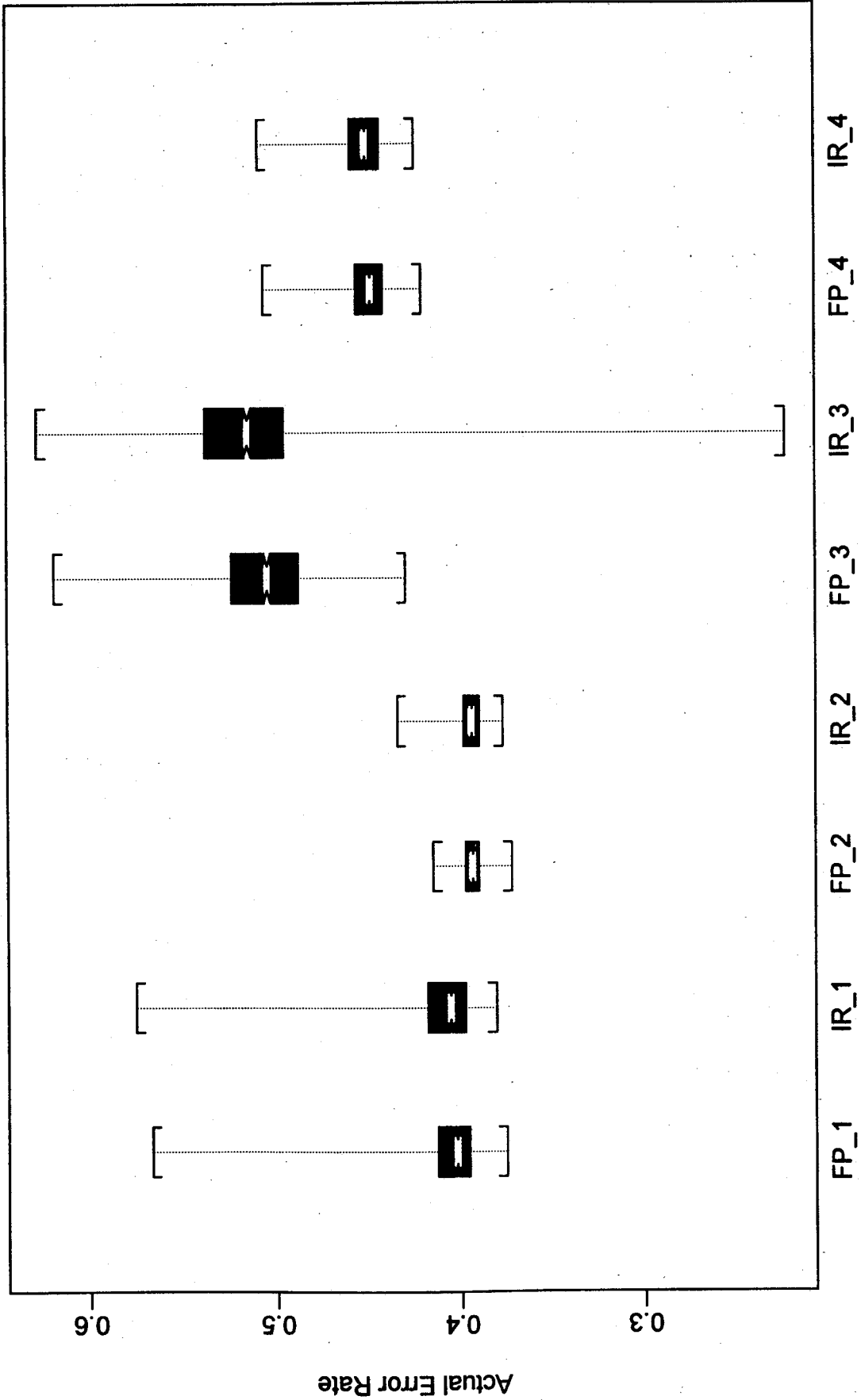


FIG. 2.29: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, DOUBLE EXPONENTIAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1

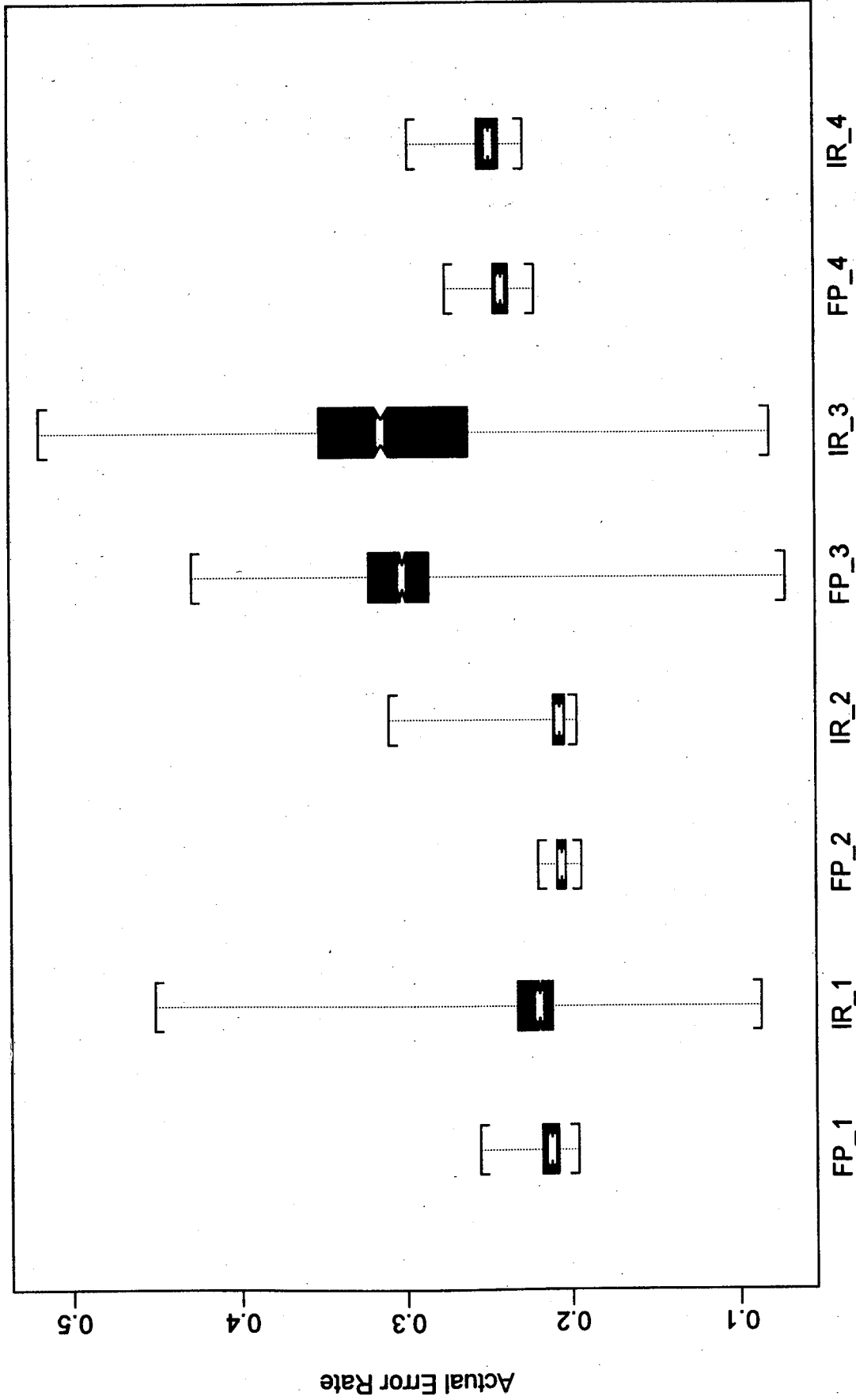


FIG. 2.30: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, DOUBLE EXPONENTIAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 4

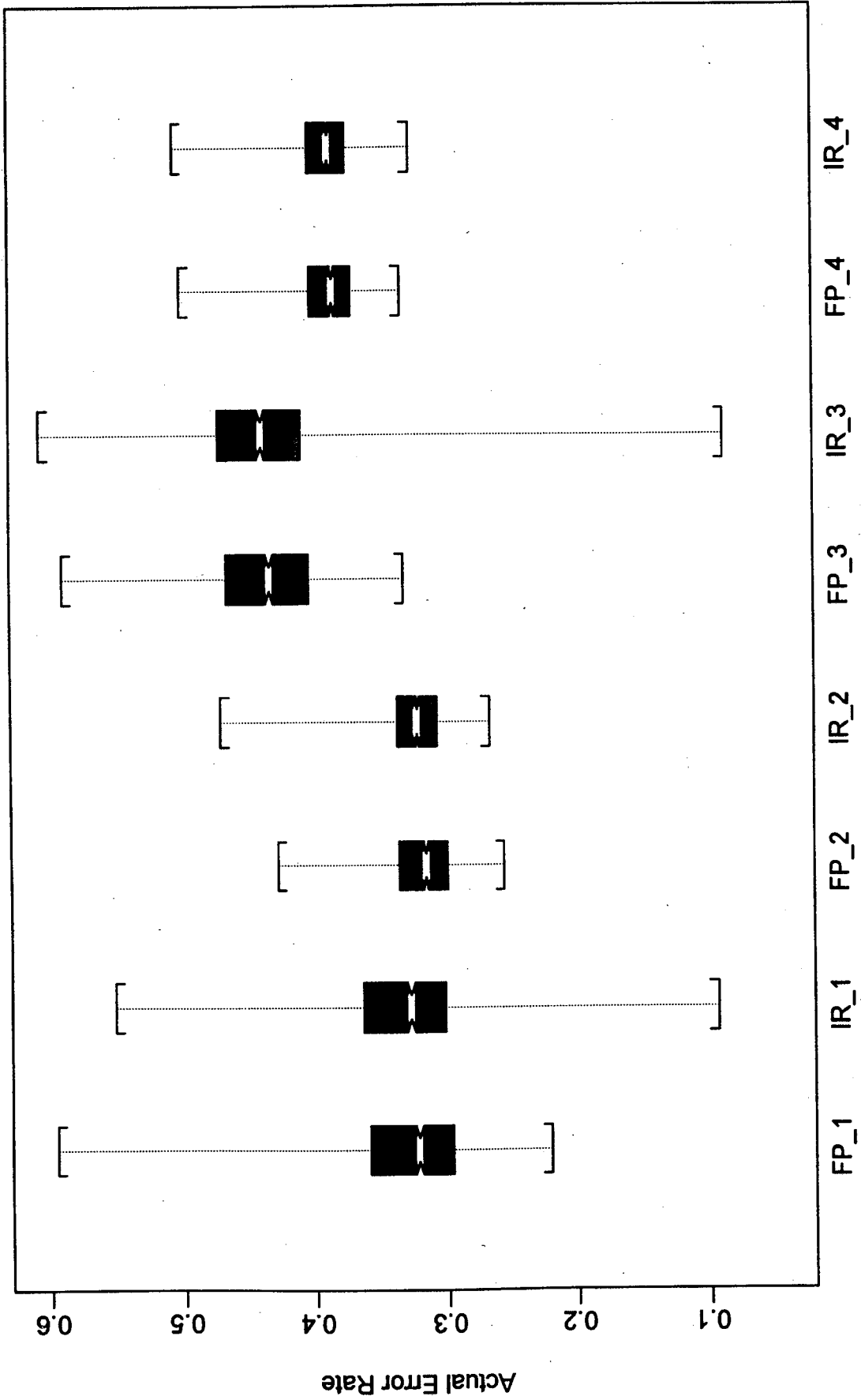


FIG. 2.31: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, LOGNORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 1

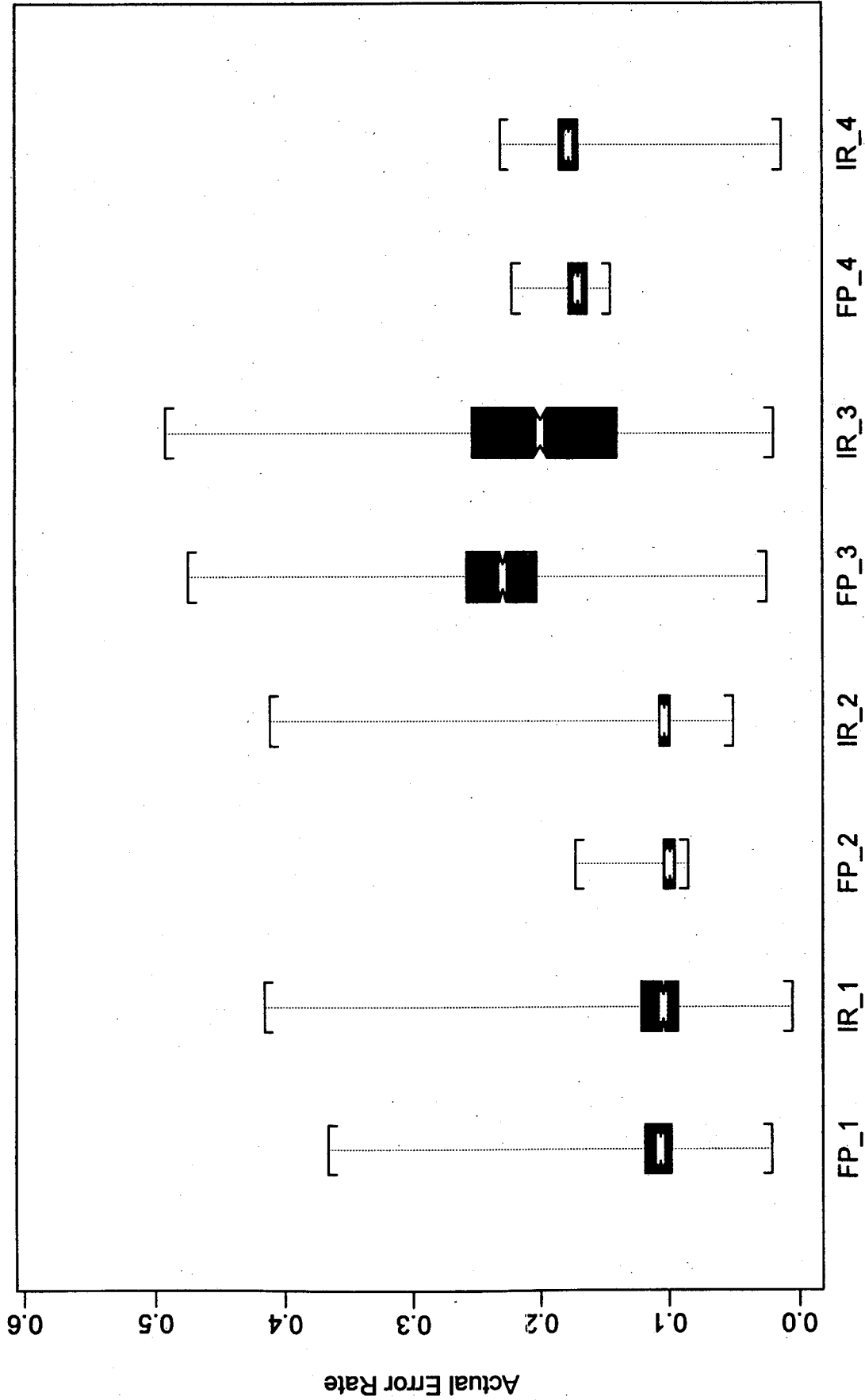


FIG. 2.32: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, LOGNORMAL DATA, CORRELATION = 0, SQUARED MAHALANOBIS DISTANCE = 4

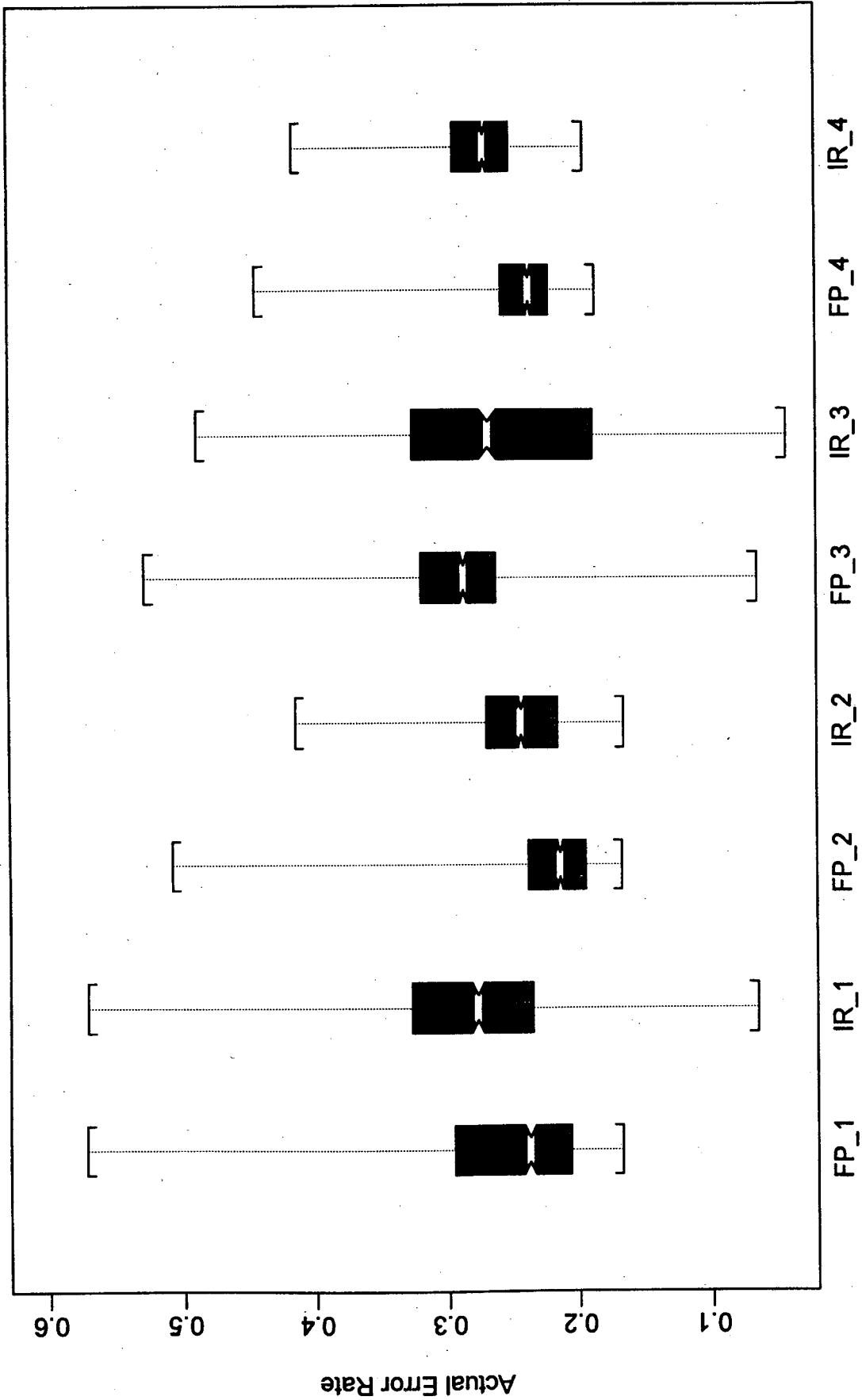


FIG. 2.33: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, LOGNORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 1

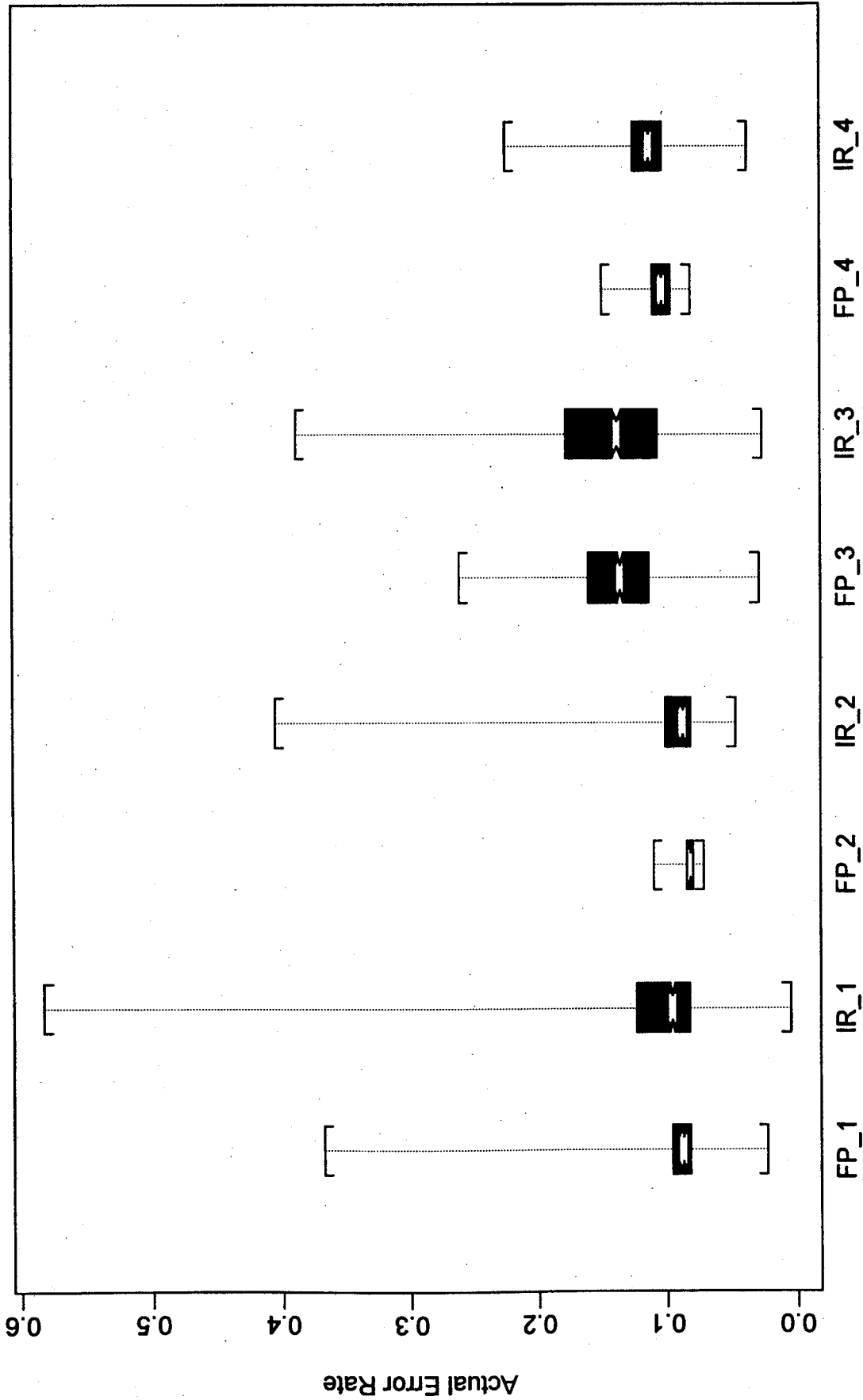


FIG. 2.34: ACTUAL ERROR RATES OF FULLY POLYCHOTOMOUS AND INDIVIDUALISED LR, LOGNORMAL DATA, CORRELATION = 0.9, SQUARED MAHALANOBIS DISTANCE = 4

TABLE 2.13 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES FULLY POLYCHOTOMOUS AND INDIVIDUALISED BINARY APPROACHES NORMAL DATA ($\rho = 0$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66659 (.00372)	.66650 (.00367)	.66654 (.00365)	.66660 (.00381)
1	.46541 (.01341)	.46799 (.01420)	.45640 (.00464)	.45698 (.00490)
2	.37555 (.00806)	.38056 (.01265)	.36854 (.00423)	.36950 (.00487)
3	.31256 (.00789)	.31995 (.01762)	.30589 (.00405)	.30705 (.00477)
4	.26448 (.00775)	.27526 (.02291)	.25710 (.00382)	.25863 (.00492)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66661 (.00381)	.66678 (.00390)	.66670 (.00388)	.66648 (.00391)
1	.52109 (.02283)	.53221 (.02504)	.47486 (.00830)	.47715 (.00913)
2	.43284 (.02096)	.44549 (.04928)	.38437 (.00729)	.38777 (.00814)
3	.36876 (.02417)	.37733 (.06488)	.31970 (.00635)	.32516 (.00813)
4	.32226 (.03295)	.31929 (.07149)	.27044 (.00625)	.27693 (.00849)

TABLE 2.14 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES FULLY POLYCHOTOMOUS AND INDIVIDUALISED BINARY APPROACHES DOUBLE EXPONENTIAL DATA ($\rho = 0$)

$k=2$	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66677 (.00373)	.66673 (.00381)	.66643 (.00369)	.66651 (.00364)
1	.40606 (.01659)	.41203 (.02409)	.39374 (.00569)	.39464 (.00685)
2	.31253 (.01069)	.32142 (.02574)	.30312 (.00508)	.30494 (.00580)
3	.25483 (.00971)	.26722 (.03052)	.24642 (.00433)	.24826 (.00594)
4	.21415 (.00863)	.22855 (.03540)	.20607 (.00391)	.20802 (.00652)

$k=10$	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66687 (.00366)	.66659 (.00369)	.66683 (.00386)	.66690 (.00387)
1	.50656 (.02712)	.51596 (.03550)	.44921 (.01144)	.45151 (.01209)
2	.41117 (.02624)	.42359 (.05599)	.35176 (.00905)	.35609 (.01007)
3	.34842 (.02697)	.35392 (.06882)	.28840 (.00722)	.29368 (.00962)
4	.30365 (.03170)	.30275 (.07651)	.24194 (.00674)	.24979 (.01008)

TABLE 2.15 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES FULLY POLYCHOTOMOUS AND INDIVIDUALISED BINARY APPROACHES LOGNORMAL DATA ($\rho = 0$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66659 (.00361)	.66672 (.00377)	.66659 (.00366)	.66672 (.00361)
1	.33146 (.05113)	.33353 (.04786)	.32058 (.03063)	.32299 (.02433)
2	.20163 (.03776)	.19652 (.04913)	.18267 (.01812)	.18099 (.01515)
3	.14376 (.03106)	.13754 (.05090)	.12748 (.01244)	.13008 (.01460)
4	.11203 (.02599)	.11124 (.05212)	.10026 (.00780)	.10434 (.01427)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66687 (.00366)	.66687 (.00368)	.66666 (.00375)	.66651 (.00382)
1	.43835 (.04605)	.43230 (.07269)	.38810 (.02517)	.39022 (.02345)
2	.33541 (.04637)	.30656 (.08845)	.27303 (.01681)	.27801 (.01824)
3	.27020 (.05534)	.23137 (.08616)	.21130 (.01341)	.21635 (.01317)
4	.22721 (.05602)	.19841 (.07684)	.17058 (.01083)	.17741 (.01367)

TABLE 2.16 MEANS AND STANDARD DEVIATIONS OF ACTUAL ERROR RATES FULLY POLYCHOTOMOUS AND INDIVIDUALISED BINARY APPROACHES LOGNORMAL DATA ($\rho = 0.9$)

k=2	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66658 (.00355)	.66660 (.00349)	.66664 (.00362)	.66682 (.00344)
1	.26258 (.07788)	.28878 (.07326)	.22394 (.04442)	.24572 (.04165)
2	.14865 (.04240)	.17923 (.06966)	.12411 (.00709)	.14617 (.03621)
3	.11221 (.03052)	.13888 (.06918)	.09810 (.00525)	.11769 (.04095)
4	.09241 (.02207)	.11525 (.07017)	.08219 (.00414)	.10445 (.04775)

k=10	SMALL SAMPLES		LARGE SAMPLES	
Δ^2	FP	IR	FP	IR
0	.66677 (.00379)	.66671 (.00378)	.66675 (.00363)	.66678 (.00370)
1	.29040 (.05400)	.25883 (.08837)	.24268 (.03090)	.27235 (.03133)
2	.20121 (.04764)	.18414 (.06721)	.15299 (.01300)	.17353 (.02337)
3	.16286 (.04573)	.16271 (.06689)	.12170 (.01048)	.13744 (.02528)
4	.13677 (.03624)	.15134 (.06610)	.10414 (.01009)	.11269 (.02471)

2.7 CONCLUSIONS AND RECOMMENDATIONS

Sections 2.4 - 2.6 contain a report on a comparison of the classification performance of the linear discriminant function and the logistic discriminant function, as measured in terms of their actual error rates for a number of cases where the feature variables are continuous.

In Section 2.4, the two group case received attention, while the case of three groups was discussed in Section 2.5. Two approaches for estimating the coefficients of the logistic discriminant function in the case of more than two groups, were compared in Section 2.6. The main conclusions were: for normal and double exponential data, the linear discriminant function outperforms the logistic discriminant function. The differences are slight in large sample cases, but quite large in cases where the number of feature variables is large compared to the sample size; for lognormal data, the logistic discriminant rule should be preferred, except for cases where the number of feature variables is large relative to the sample size. For the distributions investigated in this chapter, the use of an individualised binary approach instead of a fully polychotomous approach in the case of more than two groups should only be considered when the feature variables are lognormally distributed. Finally, logistic regression suffers from a disadvantage that was encountered especially at large separations between lognormal populations, viz. the non-existence of the maximum likelihood estimates of the parameters in the logistic regression function. This adds more weight to the general conclusion that discriminant analysis seems to be a better option than logistic regression.

CHAPTER 3

VARIABLE SELECTION AND THE CLASSIFICATION PERFORMANCE OF THE LINEAR DISCRIMINANT FUNCTION

3.1 INTRODUCTION

In many statistical applications data are available on a large number of potentially important variables. Variable selection is often used as the first step in the analysis of such data to identify a model that contains only a subset of the available variables and that is optimal in some appropriate sense. This is in line with the principle of parsimony formulated by Box and Jenkins (1972, p.17) as selecting the “smallest number of parameters for adequate representation”. A simple model is not only easier to interpret, but it also requires fewer variables to be measured than a more complex model, which can be an important cost saving factor.

Many variable selection techniques have been proposed in the literature, frequently with a view to application in regression analysis. An excellent review of this topic is provided by Miller (1990). Examples of selection procedures in regression that immediately come to mind are various stepwise procedures and the use of criteria such as Mallows' C_p (Mallows, 1973). These selection techniques are also often applied in other areas, such as discriminant analysis and logistic regression. In this chapter attention will be restricted to selection of variables for inclusion in a linear discriminant function. A variable selection technique that can be used in linear discriminant analysis as well as in logistic regression, will be proposed in the next chapter.

Selecting a subset of the available variables for use in subsequent analyses typically consists of two closely linked stages. Define the *dimension of a model* as the number of variables it includes. Then the first stage in the application of a selection technique is to identify an optimal subset of the available variables for each possible model dimension. The second stage entails comparing the optimal models of different dimensions in order to make a unique choice. In the first stage, it is necessary to define what is meant by an optimal model of given dimension. This is most frequently done in terms of a measure of lack of fit or error, and the optimal model of a given dimension is the model that minimises this measure. The second stage is more difficult, since it requires comparing the optimal models of different dimensions with respect to two contrasting aspects: model dimension or complexity, and lack of fit. Typically, the lack of fit decreases as the model dimension increases. Therefore, if lack of fit was the only aspect taken into account, it would lead to choosing the model with the highest possible dimension. The disadvantage is that overfitting typically occurs when using models of high dimension. As a result of this overfitting, it often happens that a model

of lower dimension fitting the available data less well, performs better in terms of prediction based on new data. This is an illustration of the frequently occurring bias-variance trade-off, with predictions for new cases based on a simple model typically having larger bias and smaller variance than those based on a more complex model.

In Section 3.2 an overview of variable selection techniques used in discriminant analysis is provided. Thereafter a number of aspects regarding the first stage of variable selection within a two group discriminant analysis context, are investigated. Firstly, in Section 3.3, the effect of model dimension on the classification performance of the linear discriminant rule, as reflected in the actual error rate, is studied. In this part of the study, no variable selection takes place: the actual error rate is merely determined for fixed subsets containing different numbers of feature variables. This is followed in Section 3.4 by a comparison of the properties of a number of different criteria that can be used to select a subset of feature variables of a pre-specified size. This is done by considering two groups that differ with respect to five out of a total of ten available feature variables. Different criteria are then used to select optimal models of dimension five, and the performance of these criteria is then compared in terms of the error rates of the associated discriminant functions. In this part of the study, the criteria are therefore forced to select a subset of the correct size. The results of this part of the study are used to reduce the number of selection criteria. In Section 3.5 a much more extensive investigation is undertaken into the properties of the criteria previously identified in Section 3.4. These criteria are used to select subsets of variables of all possible dimensions $(1, 2, \dots, k)$. As in Section 3.3, the classification performance of the resulting linear discriminant functions is studied. It should be noted that the difference between the studies in Sections 3.3 and 3.5 is that no variable selection takes place in Section 3.3, whereas different criteria are used to select optimal models of dimension 1 to k in Section 3.5. The investigation reported in Section 3.5 stops short of a full investigation into the properties of different variable selection criteria, since the criteria that are discussed, are not used to choose a final model from the optimal models that have been identified for each possible model dimension. This aspect is addressed in Chapter 4. The chapter closes in Section 3.6 with a number of conclusions and recommendations. Throughout Chapter 3 only two underlying distributions for the feature variables are investigated: the normal distribution, representative of the symmetric case, and the lognormal distribution, representing the asymmetrical case.

An important and notoriously difficult problem associated with variable selection in discriminant analysis is not addressed in this chapter: estimation of the post selection actual error rate. This receives attention in Chapter 4.

3.2 OVERVIEW OF TECHNIQUES USED FOR VARIABLE SELECTION IN DISCRIMINANT ANALYSIS

In this section a number of methods that are used for the selection of feature variables in discriminant analysis are discussed. These include methods that consider all possible subsets of variables, stepwise procedures commonly used in practice, simultaneous test procedures and error rate based procedures.

The following notation will be used in this section and throughout the remainder of the chapter. Consider a $(G+1)$ -group homoscedastic normal model, as described in Section 2.1. Assume that training samples of sizes n_0, n_1, \dots, n_G are available from the k -dimensional populations $\Pi_0, \Pi_1, \dots, \Pi_G$ respectively. Denote the sample vectors by \mathbf{x}_{ij} for $i=1, \dots, n_j$; $j=0, 1, \dots, G$. Each of these $n = \sum_{j=0}^G n_j$ vectors contains the observations on the k available feature variables for an entity of known origin. The sample mean vectors are

$$\bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}_{ij}, \quad j=0, 1, \dots, G, \quad (3.2.1)$$

with corresponding sample covariance matrices

$$\mathbf{S}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)', \quad j=0, 1, \dots, G. \quad (3.2.2)$$

For the homoscedastic model,

$$\mathbf{S} = \frac{1}{n - G - 1} \sum_{j=0}^G (n_j - 1) \mathbf{S}_j, \quad (3.2.3)$$

is the pooled sample covariance matrix, which is an unbiased estimator of the common population covariance matrix Σ . The population mean vectors are denoted by $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G$.

There exists an analogy between discriminant analysis and regression analysis, and an implication of this analogy is that techniques that are commonly used for variable selection in regression, can also be applied in discriminant analysis. The following exposition of this analogy for the case $G=1$, is based on Kshirsagar (1972, p. 206-214). Let Y be a dummy variable, with $Y = \lambda_0$ if an entity belongs to Π_0 , and $Y = \lambda_1$ if it belongs to Π_1 . Consider an entity with k -dimensional feature vector \mathbf{X} . Then the expected value of \mathbf{X} can be expressed as

$$E(\mathbf{X}) = \alpha + \beta Y \quad (3.2.4)$$

where

$$\alpha = \frac{\lambda_0 \mu_1 - \lambda_1 \mu_0}{\lambda_0 - \lambda_1}; \quad \beta = \frac{\mu_0 - \mu_1}{\lambda_0 - \lambda_1} \quad (3.2.5)$$

Equation (3.2.4) represents a model for the regression of \mathbf{X} on Y . Ordinarily, one would use (3.2.4) to predict the value of \mathbf{X} from that of Y . In discriminant analysis, however, the situation is reversed, since the problem is to predict the group membership, Y , of an entity with feature vector \mathbf{X} . It therefore makes sense to rather consider the regression of Y on \mathbf{X} . Let $\mathbf{X}_0: n_0 \times k$ be the matrix with rows the vectors \mathbf{x}_{i0} , $i=1, \dots, n_0$, and similarly for $\mathbf{X}_1: n_1 \times k$, with rows \mathbf{x}_{i1} , $i=1, \dots, n_1$. The matrix of corrected sums of squares and cross products of all n observations is

$$\mathbf{A}: k \times k = \mathbf{A}_0 + \mathbf{A}_1 + c^2(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)' \quad (3.2.6)$$

where $c^2 = \frac{n_0 n_1}{n_0 + n_1}$ and $\mathbf{A}_i = \mathbf{X}_i'(\mathbf{I}_{n_i} - \frac{1}{n_i} \mathbf{1}_{n_i} \mathbf{1}_{n_i}') \mathbf{X}_i$ for $i=0,1$, and with \mathbf{I}_{n_i} the $n_i \times n_i$ identity matrix and $\mathbf{1}_{n_i}$ an n_i -dimensional vector with all elements equal to 1. The vector of corrected sums of products of \mathbf{X} and Y is

$$\mathbf{a}: k \times 1 = c^2(\lambda_0 - \lambda_1)(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1), \quad (3.2.7)$$

while the sum of squares of Y is

$$q_Y = c^2(\lambda_0 - \lambda_1)^2. \quad (3.2.8)$$

It follows from standard regression theory that the least squares estimate of the vector of regression coefficients of Y on \mathbf{X} is given by

$$\mathbf{b} = \mathbf{A}^{-1} \mathbf{a} \quad (3.2.9)$$

and by using matrix manipulations, this can be written in the form

$$\mathbf{b} = \frac{c^2(\lambda_0 - \lambda_1)}{1 + c^2(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)'(\mathbf{A}_0 + \mathbf{A}_1)^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)} (\mathbf{A}_0 + \mathbf{A}_1)^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1). \quad (3.2.10)$$

It now readily follows that

$$\mathbf{b}'\mathbf{x} = -\frac{\gamma}{n_0 + n_1 - 2} \left\{ W(\mathbf{x}; \mathbf{t}) + \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)' \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0) \right\} \quad (3.2.11)$$

where $\gamma = \frac{c^2(\lambda_0 - \lambda_1)}{1 + c^2(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)'(\mathbf{A}_0 + \mathbf{A}_1)^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)}$ is independent of \mathbf{x} , and $W(\mathbf{x}; \mathbf{t})$ is the Anderson classification statistic defined in (2.1.7). Now $\mathbf{b}'\mathbf{x}$ is the predicted value of the group membership variable Y based on the observed feature vector \mathbf{x} , and from (3.2.11) it is clear that the classification implied by this prediction will be equivalent to classification based on the statistic $W(\mathbf{x}; \mathbf{t})$.

Summarising, a two-group discriminant analysis can therefore be carried out by performing a regression analysis with the dummy variable Y as dependent variable and the independent variables contained in the feature vector \mathbf{X} . Consequently, variable selection techniques that are commonly used in regression analysis can also be applied in discriminant analysis by merely using the above Y and \mathbf{X} as dependent and independent variables respectively. An aspect that deserves some more attention is the fact that the dependent variable, Y , is a dummy variable that does not satisfy the normality assumption usually made in regression analysis. This turns out however not to be a stumbling block, as argued by Kshirsagar (1972, p. 211-214), and the F-based techniques commonly used in variable selection in regression, are valid here also.

An analogy similar to the one above exists for the more general case of $G + 1$ groups. Then G dummy variables Y_1, \dots, Y_G are required, where $Y_i = 1$ if and only if the entity belongs to Π_i , and $Y_i = 0$ otherwise. Hence, the vector $\mathbf{Y}: G \times 1 = [Y_1, \dots, Y_G]'$ of dummy variables will have unity in the i -th position if the entity belongs to $\Pi_i, i = 1, \dots, G$, and zero elsewhere, while for an entity belonging to $\Pi_0, \mathbf{Y} = \mathbf{0}$. Kshirsagar (1972, Chapter 9) provides more details in this regard.

Turning to variable selection criteria that are applied in discriminant analysis, these can be grouped into two broad classes depending on whether the separatory (descriptive) or allocatory (predictive) aspect of the analysis is emphasised. If the separatory aspect is of primary interest, selection techniques that choose subsets of variables that best separate the two populations, should be used. Examples of criteria belonging to this class are the squared multiple correlation coefficient R^2 , Mallows' C_p -statistic and F-based stepwise criteria. If the classification of future entities is the primary concern, i.e. the allocatory aspect is the focus of attention, selection techniques that in some way make use of an error rate estimator, should be used.

McKay and Campbell (1982a,b) provide a good overview of selection techniques, addressing selection based on separatory criteria in the first paper, and concentrating on allocatory criteria in the second paper. Many of the procedures using separatory criteria are based on the F-statistic of the test for no additional information, defined by Rao (1965), which is now briefly explained.

Consider the $(G + 1)$ -group homoscedastic normal model, as described in Section 2.1. Let V denote the set of all k potential variables, and consider a subset V_1 , containing $p < k$ variables and its complement V_2 with $k - p$ variables. Assume without loss of generality that the variables in V_1 correspond to the indices $1, \dots, p$. Partition a typical vector of observations on the k variables as $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2)'$, where \mathbf{x}_1 contains observations on the p variables in V_1 and \mathbf{x}_2 contains observations on the $k - p$ variables in V_2 . Let

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_{i1} \\ \boldsymbol{\mu}_{i2} \end{bmatrix}, \quad i = 0, 1, \dots, G \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \quad (3.2.12)$$

be the group mean vectors and common covariance matrix partitioned in the same way. The concept of *no additional information* provided by the variables in V_2 in the presence of the variables in V_1 , is used in many of the separatory variable selection techniques mentioned above. To explain this concept, consider the two groups Π_i and Π_j , $i \neq j = 0, 1, \dots, G$. The squared Mahalanobis distance between these groups is

$$\begin{aligned} \Delta_{ij}^2 &= (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ &= (\boldsymbol{\mu}_{i1} - \boldsymbol{\mu}_{j1})' \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_{i1} - \boldsymbol{\mu}_{j1}) + \\ &\quad [\boldsymbol{\mu}_{i2} - \boldsymbol{\mu}_{j2} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_{i1} - \boldsymbol{\mu}_{j1})]' \boldsymbol{\Sigma}_{22.1}^{-1} [\boldsymbol{\mu}_{i2} - \boldsymbol{\mu}_{j2} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_{i1} - \boldsymbol{\mu}_{j1})] \end{aligned} \quad (3.2.13)$$

where $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$. Clearly, the variables in V_2 do not contribute to Δ_{ij}^2 if and only if $\boldsymbol{\mu}_{i2} - \boldsymbol{\mu}_{j2} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_{i1} - \boldsymbol{\mu}_{j1}) = \mathbf{0}$. Based on these considerations, the null hypothesis that the variables in V_2 do not provide any additional separation between any two of the $G + 1$ groups, can be formulated as

$$H_0: \boldsymbol{\mu}_{i2} - \boldsymbol{\mu}_{j2} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_{i1} - \boldsymbol{\mu}_{j1}) = \mathbf{0}, \quad i \neq j; i, j = 0, 1, \dots, G. \quad (3.2.14)$$

A test statistic for H_0 can be based on two matrices: the matrix \mathbf{B} of between-group sums of squares and cross products, and the matrix \mathbf{W} of within-group sums of squares and cross products. These matrices are given by $\mathbf{B} = \sum_{i=0}^G n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$, and $\mathbf{W} = (n - G - 1)\mathbf{S}$. Partition these matrices as in (3.2.12) and let

$$\mathbf{W}_{22.1} = \mathbf{W}_{22} - \mathbf{W}_{21} \mathbf{W}_{11}^{-1} \mathbf{W}_{12} \quad (3.2.15)$$

and

$$\mathbf{B}_{22.1} = \mathbf{B}_{22} + \mathbf{W}_{22} - (\mathbf{B}_{21} + \mathbf{W}_{21})(\mathbf{B}_{11} + \mathbf{W}_{11})^{-1}(\mathbf{B}_{12} + \mathbf{W}_{12}) - \mathbf{W}_{22.1}. \quad (3.2.16)$$

As pointed out by McLachlan (1992, p.394), test statistics for H_0 similar to those used in MANOVA can be based on the matrices $\mathbf{W}_{22.1}$ and $\mathbf{B}_{22.1}$.

In the two group case with $G = 1$, (3.2.14) becomes

$$H_0: \boldsymbol{\mu}_{02} - \boldsymbol{\mu}_{12} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_{01} - \boldsymbol{\mu}_{11}) = \mathbf{0}, \quad (3.2.17)$$

and this hypothesis can be tested by using the statistic

$$F = \frac{n_0 n_1 (n - k - 1)}{k - p} \frac{D^2 - D_1^2}{n_0 n_1 D_1^2 + n(n - 2)}, \quad (3.2.18)$$

where D^2 is the squared sample Mahalanobis distance between the two populations based on all the variables in V , and D_1^2 is the same distance based only on the p variables in V_1 . It can be shown that the test statistic in (3.2.18) has an F-distribution with $k - p$ and $n - k - 1$ degrees of freedom when the null hypothesis in (3.2.17) is true. The null hypothesis is rejected if the test statistic has a value exceeding a critical value from the relevant F-distribution. It is then concluded that the variables in the subset V_2 provide additional information, and these variables are therefore added to the model.

Stepwise procedures for variable selection in discriminant analysis make repeated use of the test for no additional information. These procedures are commonly used, and are available in most standard statistical packages. There are however many disadvantages when variables are selected in a stepwise manner. A brief description of stepwise selection methods is now given, followed by a discussion of some of the associated disadvantages. In a forward selection, the first variable to enter the model is determined by calculating the univariate analysis of variance F-statistic for each of the potential variables :

$$F_i = (n - G - 1)(1 - \Lambda_i) / G \Lambda_i, \quad i = 1, \dots, k,$$

where Λ_i , $i = 1, \dots, k$ is the Wilks' Λ -statistic associated with each of the potential variables. Here $\Lambda_1 = |\mathbf{W}_{11}| / |\mathbf{W}_{11} + \mathbf{B}_{11}|$ is obtained by partitioning the matrices \mathbf{B} and \mathbf{W} as in (3.2.12) with $p = 1$, and the Λ_i , $i = 2, \dots, k$, are obtained similarly. This F-statistic has an F-distribution with G and $n - G - 1$ degrees of freedom if the hypothesis that the i -th feature variable does not contribute to the separation between the two groups, is valid. The variable corresponding to the maximum value of the F-

statistic is entered into the model first, provided that it exceeds a specified threshold value.

For the selection of additional variables, the procedure is as follows. Consider a stage where p ($p = 1, \dots, k-1$) variables have already been entered into the model. Without loss of generality, the p variables that have been entered can be relabelled $1, 2, \dots, p$. Consider the Wilks' Λ -statistic based on the subset containing these p variables,

$$\Lambda_{1, \dots, p} = \frac{|\mathbf{W}_{1, \dots, p}|}{|\mathbf{W}_{1, \dots, p} + \mathbf{B}_{1, \dots, p}|} \quad (3.2.19)$$

and the increment if a variable, which can be labelled $p+1$, is added to the model:

$$\Lambda_{(p+1)} = \Lambda_{1, \dots, p, p+1} / \Lambda_{1, \dots, p} \quad (3.2.20)$$

The associated F-statistic that can be used to evaluate the additional separation between the groups provided by the $p+1$ -th variable in the presence of variables $1, \dots, p$, is given by

$$F = \frac{n - p - G - 1}{G} \frac{1 - \Lambda_{(p+1)}}{\Lambda_{(p+1)}} \quad (3.2.21)$$

This statistic has an F-distribution with G and $n - p - G - 1$ degrees of freedom, if the $(p+1)$ -th variable does not provide any additional separation. In implementing the forward selection process, this statistic is calculated for each variable that has not been entered into the model at that stage, and the variable corresponding to the maximum F-statistic is entered provided that this maximum exceeds a threshold value. The selection process terminates if the maximum F-statistic is smaller than the threshold.

A serious defect in this procedure is of course that maximisation of the F-statistic at each step results in the F-distribution no longer being appropriate. This has the effect that the test at each stage is not performed at the nominal significance level and that the true significance level is unknown. Hawkins (1976) provides guidelines that can be adopted with the F-based forward selection process to ensure that the overall probability of including a seemingly irrelevant variable, will be less than a pre-specified level α . Another problem is that these tests are not independent, and the simultaneous significance level of the tests is difficult to obtain. A further disadvantage of the forward selection procedure is that it does not allow for a variable to be discarded from the model once it has been entered. Because of the forward selection, the full model is never considered and therefore no indication of the performance of the selected subset relative to that of the full set of variables is obtained. Another problem results from the way in which variables are considered one at a time. It is entirely possible that two variables may not individually discriminate well between groups, but jointly they may contribute to the discrimination. If variables are considered one at a time, such variables may never be entered into the model.

Backward elimination proceeds along the same lines as forward selection. It firstly considers the full model, containing all the variables. For each variable, the F-statistic in (3.2.21) is calculated, and at each step the variable corresponding to the minimum F-statistic is removed from the model, provided that this minimum is less than a threshold value. If the minimum at any stage exceeds the threshold, the process terminates. Problems similar to those discussed in the previous paragraph for forward selection, are also present if backward elimination is used.

A fully stepwise procedure contains elements of both forward selection and backward elimination. The first two variables to be entered into the model are determined in exactly the same way as for forward selection, but in subsequent steps possible addition of a variable that has not been entered as well as deletion of a variable that has already been entered, are considered by evaluating the F-statistics defined by (3.2.21). The process terminates when no further additions or deletions can be made. The criticism of forward selection with respect to the relevance of the F-test at each stage, and with respect to the joint significance level attained by the sequence of tests, also applies to the fully stepwise procedure.

An alternative to stepwise variable selection that is gaining in popularity as computing power increases, is to evaluate all possible subsets of variables and to choose the optimal subset of each dimension according to a criterion, such as R^2 , Mallows' C_p or Wilks' lambda, defined by $\Lambda = |W_{11}|/|B_{11} + W_{11}|$. This is especially feasible if the total number of feature variables is not too large. To choose between the optimal models of each dimension, the test of no additional information can be repeatedly performed, until a stage is reached where an increase in the model dimension will not increase the separation between the groups. Since repeated hypothesis tests are performed, it is a problem to choose the critical values of the individual tests to attain a specified overall significance level. When there is a large number of potential feature variables, it is not always possible to examine all possible subsets of variables in order to find the best subset. Then recourse has to be taken to an appropriate stepwise procedure.

For the two group case, a procedure that overcomes the problems mentioned in connection with the stepwise variable selection procedures, but does not require evaluation of all possible subsets, was proposed by McKay (1976). He developed a procedure to find all subsets of variables that do not discriminate significantly worse than the entire set of variables under consideration. The advantage of this method is that the Type I family error rate can be controlled and that the significance level of each test can be determined, which is not the case in the stepwise procedures. He proposed a simultaneous test procedure similar to the procedure developed by Gabriel (1969) to find all subsets of variables for which there is a difference in the mean vectors between the populations. In the simultaneous test procedure that he proposes, McKay (1976) uses the union-intersection principle of Roy (1953) and the test for no additional information of Rao (1965). McKay (1977) also extended this procedure to the multiple group situation.

McLachlan (1976a) suggested constructing a tolerance interval for the increase in the conditional risk when a subset of variables is deleted from the discriminant function. If equal costs of misclassification are assumed, this conditional risk is the same as the conditional or actual error rate. He suggested using the difference in the asymptotic error rate estimator (cf. McLachlan, 1974) associated with the full set of variables and that for the reduced set, to estimate the increase in the risk. He then derived the asymptotic distribution of the difference between the estimator of increased risk and the true increased risk, and used this to construct a tolerance interval for the true increased risk. The confidence coefficient corresponding to no increase in the risk is regarded as an indication of the additional discrimination value of the deleted variables.

McLachlan (1980a) combined separatory and allocatory considerations and investigated the relationship between the F-test and the overall error rate for variable selection in the two group case with the assumption of a homoscedastic normal model. He compared selection based on the F-test for no additional information with selection based on a criterion that considers the asymptotic probability of no increase in the overall error rate if a subset of variables is deleted. He analysed several data sets and concluded that there is 'a fairly high degree of confidence' that the overall error rate will not increase if selection of variables is based on the F-test, provided that the significance level of the F-test is not 'too conservative'.

Variable selection techniques that take allocatory considerations into account, typically entail minimisation of an estimate of the (actual) error rate that is calculated for each model under consideration. Habbema and Hermans (1977) expressed the opinion that selection procedures using error rate as selection criterion should be employed when the aim of the discriminant analysis is that of allocating future cases. They developed an algorithm, called ALLOC-1, in which they perform a stepwise analysis similar to the F-based stepwise analysis, but each time adding the variable that results in the smallest estimated leave-one-out error rate. The procedure terminates if the decrease in the error rate when an additional variable is added, is less than a certain threshold value. Their algorithm does not require multivariate normality, but estimates the density functions by means of the kernel method. Habbema and Hermans (1977) consider more than two groups and compare the allocation performance of this procedure to that of the usual F-based forward selection as well as all possible subsets selection, using two example data sets and forcing all the procedures to continue until all variables are selected. The order in which the variables enter the model is completely different for the error rate based procedure than for the other two procedures. There are also differences in selection order between the F-based forward selection and the all possible subsets procedures, but these two procedures are more in agreement with one another than with the error rate based procedure. The error rates, as estimated by the apparent error rate as well as the leave-one-out error rate, of the models selected by each of the methods for each model size, are also compared, and the estimated error rates attained by the error rate based procedure are lower for each model size than that of the other two procedures. It must however be mentioned that one of the data sets used, consisted of twelve populations with a sample size of four per population, and 9 variables. Various authors warn against the use of stepwise selection in such

circumstances. A much more detailed study is required to properly evaluate the relative merit of selection procedures.

Two points of criticism can be levelled against the approach proposed by Habbema and Hermans (1977). Firstly, specification of the threshold that determines termination of the stepwise procedure is problematic, and the authors provide little guidance in this respect. Their proposal to compare the reduction in the estimated error rate when an additional variable enters the model to an absolute threshold value, seems unrealistic, since the magnitude of the estimated error rates fluctuates widely depending on the separation between the groups. Using a threshold value at each step that is expressed relative to the estimated error rate at that step, seems a better option. Another solution to this problem is to replace the stepwise approach by an all possible subsets approach and to select the model that leads to the global minimum estimated error rate. The authors discount this option on the basis that it would be too time consuming. The second problem with Habbema and Hermans' approach is that it can often happen that different models of the same size give the same estimate of actual error rate, thus making a unique decision at each stage of the process problematic. This is relevant if a 0-1 loss function is used. A solution to this problem is to use a different loss function, and the authors briefly refer to using the posterior probabilities of group membership in the selection process.

More recently Ganeshanandam and Krzanowski (1989) also investigated the use of leave-one-out error rate as variable selection criterion. They assume that the required model dimension, $p \leq k$ is fixed. They then select a best subset of p variables by means of a fully stepwise procedure, at each step using estimated error rate to decide on inclusion or deletion of a variable. They also propose a method of assessing the classification performance of the final rule, and this will be discussed in Chapter 4. A point of criticism against their approach is that the difficult and important problem of choosing between different model dimensions, is not addressed. Furthermore, the use of an error rate estimator employing a 0-1 loss function, can result in the non-uniqueness problem referred to in the previous paragraph.

3.3 THE EFFECT OF MODEL DIMENSION ON THE PROPERTIES OF THE RESULTING CLASSIFICATION RULE (NO SELECTION)

Consider two groups, Π_0 and Π_1 , with equal prior probabilities. Training data consisting of observations on k feature variables for a total of n entities of known origin are available. Denote this training data by t , as defined in Section 2.1. If a linear discriminant analysis approach is used, an entity of unknown origin with feature vector x can be classified by using the Anderson classification statistic, $W(x; t)$, given in (2.1.7). In this section the actual error rate as defined in (2.2.8), will be used to evaluate the classification performance of this rule.

The following further notation is required. Denote a subset of the set of indices $\mathcal{K} = \{1, \dots, k\}$ by \mathcal{J} , and suppose the number of elements in \mathcal{J} is $p \leq k$. The Anderson statistic based only on the p variables corresponding to the indices in \mathcal{J} will be denoted by $W_p(\mathbf{x}; t(\mathcal{J}))$. In this notation, the statistic based on all k feature variables is $W(\mathbf{x}; t) = W_k(\mathbf{x}; t(\mathcal{K}))$. If the subset \mathcal{J} and its cardinality p are determined from the training data, as is the case when variable selection is performed, the resulting classification statistic will be denoted by $W_{p(t)}(\mathbf{x}; t(\mathcal{J}(t)))$.

An important objective in this thesis is to evaluate variable selection techniques that are currently used in discriminant analysis, and to propose new techniques for this purpose that perform better than the currently used techniques in the sense that classification statistics with lower actual error rates are obtained. At some stage therefore it will be necessary to investigate the error rate behaviour of statistics of the form $W_{p(t)}(\mathbf{x}; t(\mathcal{J}(t)))$, where $\mathcal{J}(t)$ and $p(t)$ are found by applying some variable selection technique to the training data. In this section though, attention is restricted to an investigation into the error rate behaviour of statistics $W_p(\mathbf{x}; t(\mathcal{J}))$, i.e. cases where \mathcal{J} and p are specified beforehand, independent of the training data. The purpose is to study the effect of model dimension (i.e. the value of p) and the variables that are included in the linear discriminant function on the error rate of this function. By keeping \mathcal{J} and p independent of t , the possible effect of the selection step on the error rate behaviour of the resulting linear discriminant function is eliminated. The results of this investigation may also provide valuable guidelines to the way in which an eventual variable selection technique should be structured in order to ensure that discriminant functions obtained from application of such a technique, have good error rate behaviour.

Details of the simulation study that was undertaken in the above context, are now provided. Two different distributions for the feature variables x_1, \dots, x_k were studied: as an example of a symmetric distribution, the case of normally distributed feature variables, and as an example of a skewed distribution, the case where these variables are lognormally distributed. For each of these two cases, two sample sizes were used: $n_0 = n_1 = 25$ (small samples) and $n_0 = n_1 = 100$ (large samples). Here n_i is used to denote the number of entities in the training sample from $\Pi_i, i=0,1$. In the discussion below, NS and NL will respectively refer to the small and large sample cases with normal feature variables, and similarly for the lognormal case, where LS and LL will be used. The value $k=10$ was used throughout. With respect to the covariance structure, the choices $\Sigma = \mathbf{I}$ (representing uncorrelated variables with unit variances) and Σ given by (2.4.1) (representing equi-correlated variables with unit variances) were made. The ρ -values $-0.1, 0.4$ and 0.9 were used. These choices represent a wide range of correlation: from a fairly small negative correlation through the uncorrelated case, to moderate and large positive correlation. Note that the condition $-1/(k-1) < \rho < 1$ has to be satisfied in the equi-correlated case for Σ to be positive definite. Extending the coding that was introduced above, NS1 - NS4 will be

used to refer to the four different cases with $\rho = -0.1, 0, 0.4, 0.9$ respectively, for normal feature variables and small sample sizes. The codes NL1-NL4, LS1-LS4 and LL1-LL4 are defined similarly. The final factor that was varied in the study was the number r , of feature variables with respect to which the two populations were assumed to differ. These variables will informally be referred to as relevant. Values $r = 1$, $r = 5$ and $r = 10$ were used. Extending the coding still further, NS11, NS12 and NS13 will refer to the cases of normal feature variables, small samples, $\rho = -0.1$ and $r = 1$, $r = 5$ and $r = 10$ respectively. A similar coding is used for the other cases.

Throughout the study it is assumed that the feature vector \mathbf{X} has mean vector $\boldsymbol{\mu}_0 = \mathbf{0}$ in Π_0 . Separation between the two populations was obtained by assuming non-zero values for r of the elements of $\boldsymbol{\mu}_1$, the mean vector of \mathbf{X} in Π_1 . It is a convenient and widely accepted practice (cf. McLachlan, 1992, p. 25) to describe the separation between Π_0 and Π_1 in terms of Δ^2 , the squared Mahalanobis distance between these two groups. The values $\Delta^2 = 1, 2, 3, 4$ were used. To obtain these distances, the following parameterisation was used for the elements of $\boldsymbol{\mu}_1$. For the cases where $r = 1, 5$,

$$\mu_{1l} = \begin{cases} \Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r \sigma^{ij}} & , \quad l = 1, \dots, r \\ 0 & , \quad l = r + 1, \dots, 10, \end{cases}$$

while for $r = 10$:

$$\mu_{1l} = \Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r \sigma^{ij}} , \quad l = 1, \dots, 10.$$

Here σ^{ij} are the elements of Σ^{-1} . It should be noted that the above specifications for the elements of $\boldsymbol{\mu}_1$ yield the pre-specified values of Δ as the Mahalanobis distance between Π_0 and Π_1 taking all k feature variables into account. Also, in all cases, the non-zero elements of $\boldsymbol{\mu}_1$ are equal. Finally, in each case, the variables with respect to which the two populations differ, correspond to the first r indices in \mathcal{K} .

The factors discussed above, identify a total of 48 different cases. In each of these cases, the expected actual error rate, i.e. the unconditional error rate, associated with $W_p(\mathbf{x}; t(\mathcal{J}))$ was estimated using simulation, for $p = 1, \dots, k$. For each of the values of p , the indices in \mathcal{J} were taken to be $1, \dots, p$. Consequently, for $p \leq r$ the linear discriminant function contained only seemingly relevant variables, and for $p > r$, it contained all the seemingly relevant variables and one or more seemingly irrelevant variables. Of course, for any given value of p there are many other ways to specify the indices contained in \mathcal{J} , but these were not considered in the study. Summarising, the

results that are discussed below illustrate the resulting error rate behaviour if a practitioner, confronted with two k -dimensional populations at a Mahalanobis distance Δ apart, decides on subjective or a priori grounds to use only a subset of $p \leq k$ of the available variables in the classification function.

Van Ness and Simpson (1976) studied the effect of dimension, i.e. the number of variables in the classification function, on actual error rate for five discriminant rules, including the linear discriminant rule. They considered the case of k uncorrelated normal feature variables, and assumed that the two populations differ only with respect to a single variable, i.e. in the notation introduced above, they took $r = 1$, $\mu_0 = \mathbf{0}$ and $\mu_{11} = \Delta$, $\mu_{1l} = 0$, $l = 2, \dots, k$. The values used for k were: 1, 2, 3, 5, 10, 20 and 30. Sample sizes $n_0 = n_1 = 10$ and $n_0 = n_1 = 20$ were investigated. Although these authors concentrate mainly on a comparison of the behaviour of the different discriminant rules as dimension changes, the results that they obtain for the linear discriminant function are in agreement with the results described below for the corresponding cases. It should be noted that they did not study any cases where the feature variables are correlated, where the two populations differ with respect to more than one feature variable, or cases where the feature variables are not normally distributed.

3.3.1 THE NORMAL CASE

If the feature variables are normally distributed, the actual error rate associated with $W_p(\mathbf{x}; t(j))$ was calculated using (2.2.9). It should be noted that for this purpose the quantities in (2.2.9) were calculated using only the p variables with indices in j . To estimate the required unconditional error rates, 5000 Monte Carlo repetitions were used. For each repetition a training data set was generated from the two relevant normal distributions, and the actual error rate associated with $W_p(\mathbf{x}; t(j))$ was calculated from (2.2.9) for $p = 1, \dots, k$. The unconditional error rates were estimated by averaging these quantities.

McLachlan (1992, p. 18) provides the following asymptotic expression that can be used to calculate approximate values of the unconditional error rates for the cases considered in this section:

$$\Phi\left(-\frac{1}{2}\Delta_p\right) + n^{-1} \left\{ \phi\left(\frac{1}{4}\Delta_p\right) / 4 \right\} \left\{ p\Delta_p + 4(p-1)\Delta_p^{-1} \right\}. \quad (3.3.1.1)$$

In this expression, ϕ is the probability density function of the standard normal distribution, and

$$\Delta_p^2 = (\mu_{1p} - \mu_{0p})' \Sigma_p^{-1} (\mu_{1p} - \mu_{0p})$$

is the squared Mahalanobis distance between Π_0 and Π_1 , based only on the p variables with indices in J . Strictly, expression (3.3.1.1) is valid only in cases of very large samples, but it should also provide an approximate indication of the true unconditional error rate values for smaller sample sizes. This is confirmed by a comparison of the values calculated from (3.3.1.1) with the simulation study results (see Tables 3.1 and 3.2). The reason for referring to the expression at this point is that it provides an indication of the way in which the unconditional error rate varies with n , p and Δ_p . For constant n , (3.3.1.1) is a function of p and Δ_p . If Δ_p should remain constant with changes in p , (3.3.1.1) is monotone increasing in p . This is true in cases NS21 and NL21, and for cases NS22 and NL22 when $p \geq 5$. In all the other cases considered, Δ_p changes with p , and the effect of a change in the value of p on the unconditional error rate is more complex. Specifically, it is clear from (3.3.1.1) that the unconditional error rate will no longer necessarily be a monotone increasing function of p .

TABLE 3.1 : ERROR RATES FOR SMALL SAMPLE CASE ($n_0 = n_1 = 25$)

Δ_p^2	Expression (3.3.1.1)	Simulation
1	0.3974	0.3695
2	0.3140	0.2950
3	0.2624	0.2450
4	0.2256	0.2067

TABLE 3.2 : ERROR RATES FOR LARGE SAMPLE CASE ($n_0 = n_1 = 100$)

Δ_p^2	Expression (3.3.1.1)	Simulation
1	0.3307	0.3271
2	0.2584	0.2548
3	0.2105	0.2062
4	0.1754	0.1703

The results of the simulation study were summarised by means of graphs, of which a representative selection appears in Figs. 3.1 - 3.4. Since the results for the large sample cases are largely similar to those for the small sample cases, both large and small sample results are only given for the case where $r = 1$ (see Figs. 3.1 and 3.2). For the cases where $r = 5$ and $r = 10$, only the small sample cases are shown (see Figs. 3.3 and 3.4). Each of the graphs in these figures shows the unconditional error rate as a function of p for one of the normal cases defined above. Four values of $\Delta^2 = \Delta_k^2$, the squared Mahalanobis distance between Π_0 and Π_1 , based on all k feature variables, are represented in every graph. The following general conclusions are evident from an inspection of these graphs.

1. If the feature variables are uncorrelated, the unconditional error rate is a minimum at $p = r$, i.e. all the seemingly relevant variables and none of the seemingly irrelevant variables should be included in the classification function.
2. If the feature variables are positively correlated and $r < k$, inclusion of one or more seemingly irrelevant variables into the classification function leads to a decrease in the error rate. This effect becomes more pronounced as the correlation increases (see Fig. 3.1 for case NS41, Fig. 3.2 for case NL41 and Fig. 3.3 for cases NS32 and NS42).
3. If the correlation between the feature variables has a large positive value (the cases where $\rho = 0.9$), the unconditional error rate reaches a maximum at $p = r$, i.e. the worst possible option is to use a classification rule based on all the seemingly relevant variables, without any seemingly irrelevant variables. A striking feature of the graphs for the cases where $\rho = 0.9$ with $r < k$, is the sharp reduction in the unconditional error rates if a single seemingly irrelevant variable is added to the classification function containing all the seemingly relevant variables (see Fig. 3.1 for case NS41, Fig. 3.2 for case NL41 and Fig. 3.3 for case NS42).
4. From an inspection of the graphs for the cases where $\rho = 0.4$, and a comparison of these graphs with those for $\rho = 0$ and $\rho = 0.9$, it is clear that the change in error rate behaviour from $\rho = 0$ to $\rho = 0.9$ takes place gradually.
5. If the feature variables are negatively correlated, the unconditional error rate is a minimum at $p = k$, irrespective of the value of r . The only exception to this is at $\Delta^2 = 1$ in case NS11 (see Fig. 3.1), where the minimum error rate is attained at $p = r$.
6. In general the uncorrelated cases seem more favourable than cases where the feature variables are correlated, in the sense that the lowest unconditional error rate attainable by appropriately choosing p in uncorrelated cases, is lower than the corresponding values for correlated cases.
7. Obviously, the error rates decrease with increasing sample size, and also with an increase in the value of Δ_k^2 .

Conclusions 1 and 3 above deserve some more comment. At first sight it may seem somewhat strange that inclusion of variables with respect to which two populations do not differ, can actually reduce the error rate of the classification function. A partial explanation for this phenomenon lies in the fact that addition of such a variable does in fact cause the Mahalanobis distance between the two populations to increase, provided that the variable being added is correlated with the variables already in the linear discriminant function. To see why addition of a so-called irrelevant variable can cause the Mahalanobis distance to increase, consider the k -dimensional feature vector \mathbf{X} with mean μ_i and covariance matrix Σ in group $\Pi_i, i = 0, 1$. Let $\eta = \mu_1 - \mu_0$. Then the squared Mahalanobis distance between the two groups based on all k variables is

$$\Delta^2 = \Delta_k^2 = \eta' \Sigma^{-1} \eta.$$

Partition $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1: p \times 1 \\ \mathbf{X}_2: (k-p) \times 1 \end{bmatrix}$, and partition η and Σ correspondingly.

Then it readily follows that

$$\Delta^2 = \eta_1' \Sigma_{11}^{-1} \eta_1 + (\eta_2 - \Sigma_{21} \Sigma_{11}^{-1} \eta_1)' \Sigma_{22.1}^{-1} (\eta_2 - \Sigma_{21} \Sigma_{11}^{-1} \eta_1), \quad (3.3.1.2)$$

where $\Delta_p^2 = \eta_1' \Sigma_{11}^{-1} \eta_1$ is the squared Mahalanobis distance based only on the p variables in \mathbf{X}_1 , and $\Delta_{k,p}^2 = (\eta_2 - \Sigma_{21} \Sigma_{11}^{-1} \eta_1)' \Sigma_{22.1}^{-1} (\eta_2 - \Sigma_{21} \Sigma_{11}^{-1} \eta_1)$ is the increase in the squared Mahalanobis distance brought about by adding the $k-p$ variables in \mathbf{X}_2 . If all the variables in \mathbf{X}_2 are seemingly irrelevant, then $\eta_2 = \mathbf{0}$, and adding these variables will lead to an increase in the squared Mahalanobis distance if and only if $\Sigma_{21} \neq \mathbf{0}$, i.e. if and only if \mathbf{X}_1 and \mathbf{X}_2 are correlated. An interesting special case is when addition of a single variable is considered. Then Σ_{21} becomes a row vector σ'_{12} of covariances, $\Sigma_{22.1} = \sigma_{p+1,p+1} - \sigma'_{12} \Sigma_{11}^{-1} \sigma_{12}$ and

$$\Delta_{p+1,p}^2 = (\eta_{p+1} - \sigma'_{12} \Sigma_{11}^{-1} \eta_1)^2 / (\sigma_{p+1,p+1} - \sigma'_{12} \Sigma_{11}^{-1} \sigma_{12}). \quad (3.3.1.3)$$

If the variable being added is seemingly irrelevant, $\eta_{p+1} = 0$ and

$$\Delta_{p+1,p}^2 = (\sigma'_{12} \Sigma_{11}^{-1} \eta_1)^2 / (\sigma_{p+1,p+1} - \sigma'_{12} \Sigma_{11}^{-1} \sigma_{12}),$$

and this will be positive if and only if the $(p+1)$ -th variable is correlated with the p variables already in the linear discriminant function. It is possible to write (3.3.1.3) in another interesting form, viz.

$$\Delta_{p+1,p}^2 = [(\mu_{1,p+1} - \mu_{0,p+1}) - \sigma'_{12} \Sigma_{11}^{-1} (\mu_{11} - \mu_{01})]^2 / \sigma_{p+1,p+1} (1 - \rho_{p+1.1\dots p}^2),$$

where $\rho_{p+1.1\dots p}^2$ is the squared population multiple correlation coefficient between \mathbf{X}_{p+1} and \mathbf{X}_1 . Flury (1989) draws attention to the points made above for the special case $p = 1$, and he presents illustrations that aid in the interpretation of these and other similar phenomena.

The above argument offers only a partial explanation of the change in actual error rate as (seemingly irrelevant) variables are added to the linear discriminant function, since the actual error rate is not a monotone function of the squared Mahalanobis distance. As McLachlan (1992, p. 391) points out, it may happen that addition of a variable to

the linear discriminant function causes only a slight increase in the squared Mahalanobis distance, and that this is offset by the need to estimate an additional parameter, causing the overall actual error rate to increase. This is illustrated in Figs. 3.1 - 3.4 by the behaviour of the actual error rate if the process of adding seemingly irrelevant variables is continued beyond $p = r + 1$.

The following simple two-dimensional example may help to further explain the decrease in actual error rate if an seemingly irrelevant variable is added to the variable(s) already in the linear discriminant function. Suppose the feature vector $\mathbf{X} = [X_1, X_2]'$ is normally distributed, with $E(\mathbf{X}) = \mu_0 = \mathbf{0}$ in Π_0 , and $E(\mathbf{X}) = \mu_1 = [\Delta\sqrt{1-\rho^2}, 0]'$ in Π_1 , and with common covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

with $\rho \neq 0$. The above parameterisation for μ_1 ensures that the Mahalanobis distance between Π_0 and Π_1 , based on both variables, will equal Δ . It is assumed that training samples of equal sizes are available from Π_0 and Π_1 , and these samples yield the mean vectors $\bar{\mathbf{x}}_0 = (\bar{x}_{01}, \bar{x}_{02})'$, $\bar{\mathbf{x}}_1 = (\bar{x}_{11}, \bar{x}_{12})'$ and the pooled covariance matrix

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \quad \text{with inverse } \mathbf{S}^{-1} = \begin{bmatrix} s^{11} & s^{12} \\ s^{21} & s^{22} \end{bmatrix}.$$

The Anderson classification statistic based only on X_1 is given by:

$$W_1(x_1) = [x_1 - \frac{1}{2}(\bar{x}_{11} + \bar{x}_{01})](\bar{x}_{11} - \bar{x}_{01})/s_{11},$$

where $\mathbf{x} = (x_1, x_2)'$ is the feature vector of an entity of unknown origin. Without loss of generality, assume that $\bar{x}_{11} - \bar{x}_{01} > 0$ and that $\mathbf{x} \in \Pi_1$ is misclassified, i.e. $W_1(x_1) \leq 0$. This is equivalent to $x_1 \leq \frac{1}{2}(\bar{x}_{11} + \bar{x}_{01})$. Now consider classification of this entity using the Anderson classification statistic based on both X_1 and X_2 , viz.

$$W_2(\mathbf{x}) = [\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_0)]' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0).$$

Since $\mu_{02} = \mu_{12}$, it seems reasonable to make the simplifying assumption that $\bar{x}_{02} - \bar{x}_{12} \approx 0$. This implies that

$$W_2(\mathbf{x}) \approx x_1(\bar{x}_{11} - \bar{x}_{01})s^{11} - \frac{1}{2}(\bar{x}_{11}^2 - \bar{x}_{01}^2)s^{11} \\ + x_2(\bar{x}_{11} - \bar{x}_{01})s^{21} - \frac{1}{2}(\bar{x}_{12} + \bar{x}_{02})(\bar{x}_{11} - \bar{x}_{01})s^{21}.$$

Using $W_2(\mathbf{x})$, the given entity will be classified correctly if $W_2(\mathbf{x}) > 0$. This is easily seen to be equivalent to

$$s^{21}x_2 > s^{11}\left[\frac{1}{2}(\bar{x}_{11} + \bar{x}_{01}) - x_1\right] + s^{21}(\bar{x}_{12} + \bar{x}_{02})/2.$$

For moderate to large positive values of ρ , σ^{21} will be a large negative number, and hence s^{21} will also be negative with large probability. Hence $W_2(\mathbf{x}) > 0$ is equivalent to

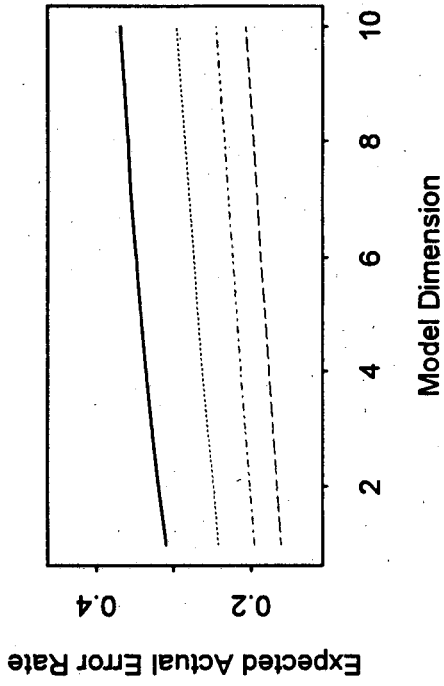
$$x_2 < s^{11}\left[\frac{1}{2}(\bar{x}_{11} + \bar{x}_{01}) - x_1\right]/s^{21} + (\bar{x}_{12} + \bar{x}_{02})/2. \quad (3.3.1.4)$$

It was assumed above that $W_1(\cdot)$ classified the given entity incorrectly, i.e. that $x_1 \leq \frac{1}{2}(\bar{x}_{11} + \bar{x}_{01})$ was observed for X_1 in Π_1 . Consider a case where $\frac{1}{2}(\bar{x}_{11} + \bar{x}_{01}) - x_1$ is small, i.e. a case where the classification decision is marginal. Since $\mu_{11} > \frac{1}{2}(\mu_{11} + \mu_{01})$, the fact that $x_1 \leq \frac{1}{2}(\bar{x}_{11} + \bar{x}_{01})$ implies that x_1 is in this case most probably appreciably below μ_{11} . The large positive correlation between X_1 and X_2 therefore implies that with high probability X_2 will be observed appreciably below μ_{12} . Since $\mu_{12} = \mu_{02}$, $\frac{1}{2}(\bar{x}_{12} + \bar{x}_{02}) \approx \mu_{12}$. With $\frac{1}{2}(\bar{x}_{11} + \bar{x}_{01}) - x_1$ small, the event in (3.3.1.4) will also occur with high probability, and this is equivalent to a correct classification using $W_2(\cdot)$.

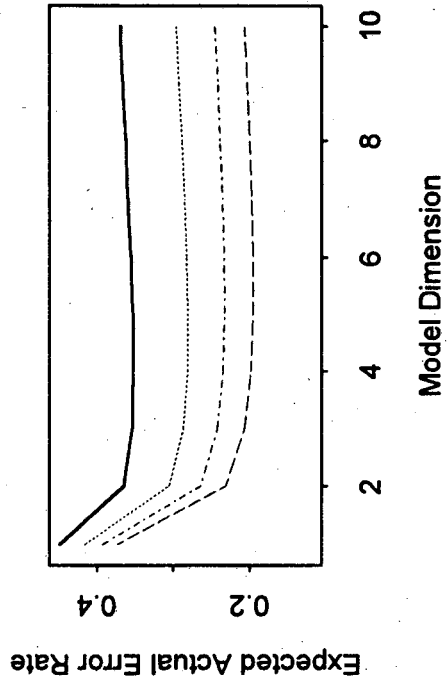
The above argument certainly does not prove that addition of an seemingly irrelevant variable to the variables in a classification function will always reduce the associated error rate, but it does provide an intuitive motivation why this phenomenon could occur.

The following may help to strengthen this motivation. Consider once more two 2-dimensional populations, with feature variables X_1 and X_2 which are strongly positively correlated. Assume that X_1 separates the populations well, and that the two populations do not differ with respect to X_2 . Without loss of generality, assume that $E(X_1|\Pi_0) < E(X_1|\Pi_1)$, but that there exists a region where the two populations overlap with respect to X_1 . If an entity of unknown origin has to be classified based only on an observation of X_1 , misclassification can easily occur if this observation lies in the region of overlap. Note that this corresponds either to an entity belonging to group Π_0 yielding a large value of X_1 , or to an entity belonging to group Π_1 yielding a small value of X_1 . Since X_1 and X_2 are highly positively correlated, this would imply either a fairly large value of X_2 if the entity belongs to Π_0 , or a fairly small value of X_2 if the entity belongs to Π_1 . Clearly therefore, including X_2 in the classification function will make correct classification of the entity more probable.

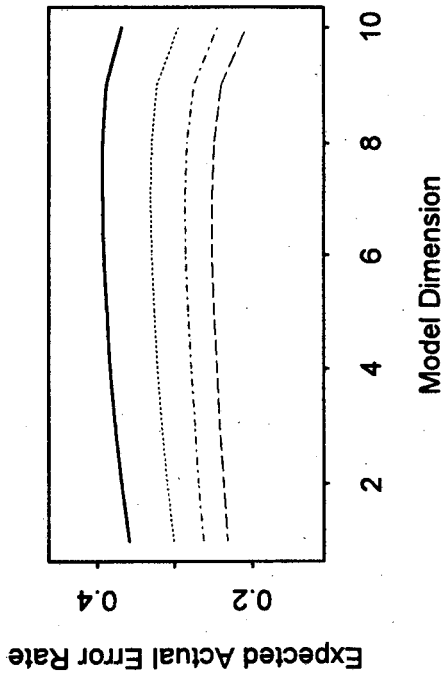
CASE NS21



CASE NS41



CASE NS11



CASE NS31

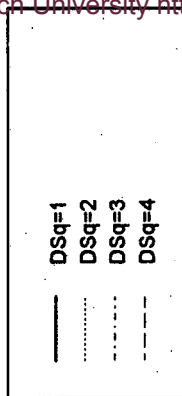
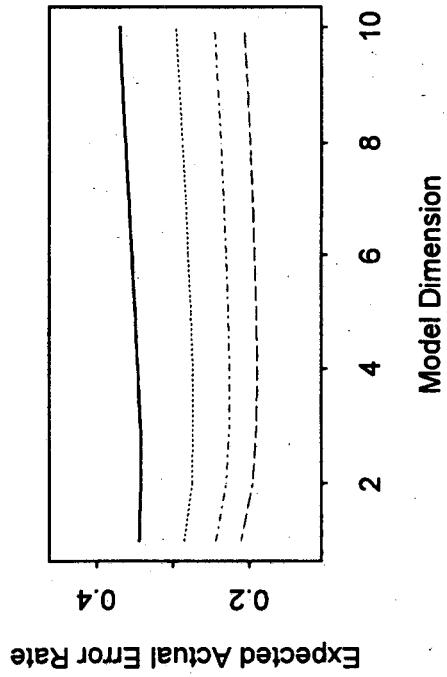
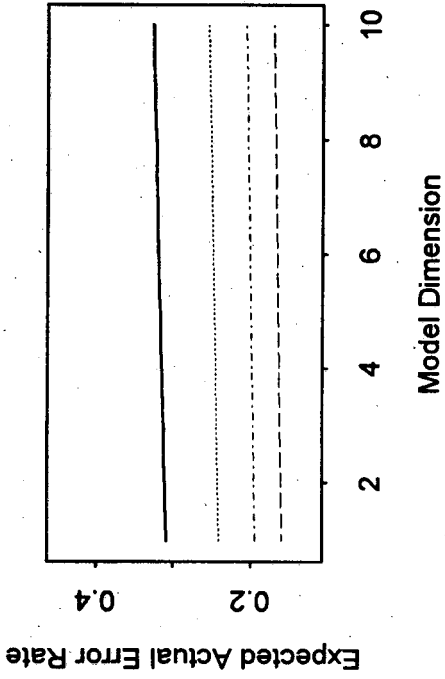
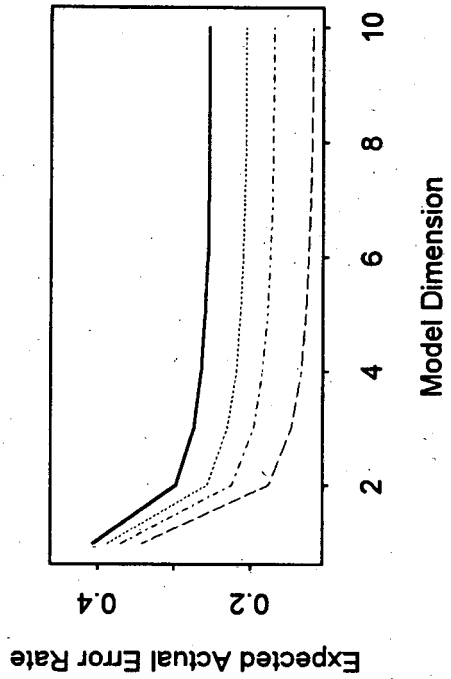


FIG. 3.1: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, NORMAL DATA, SMALL SAMPLES, $r = 1$

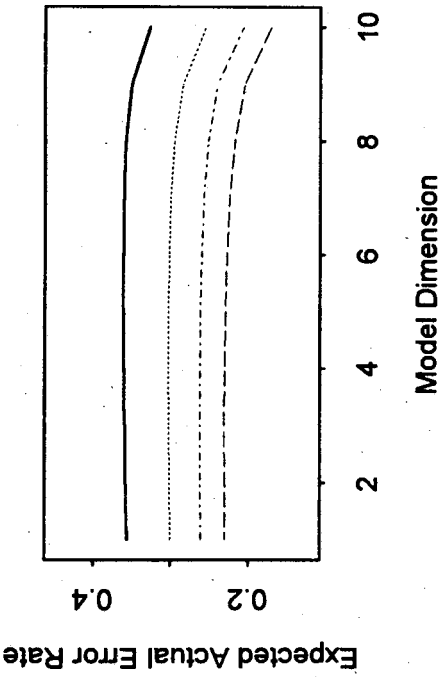
CASE NL21



CASE NL41



CASE NL11



CASE NL31

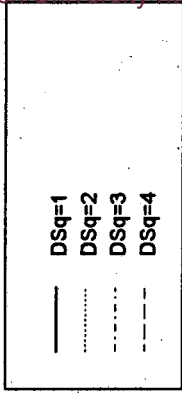
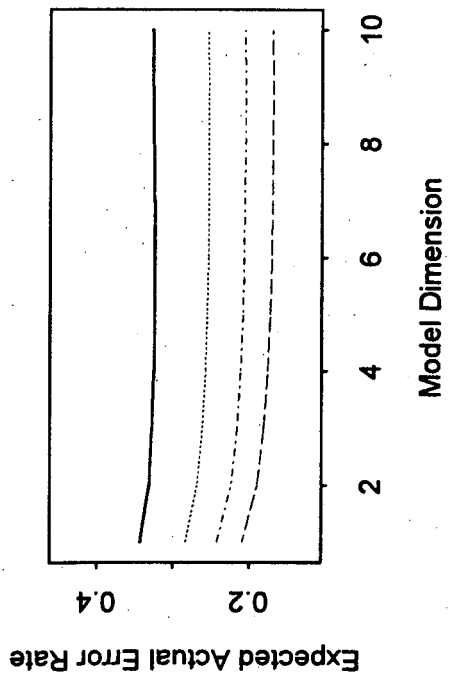
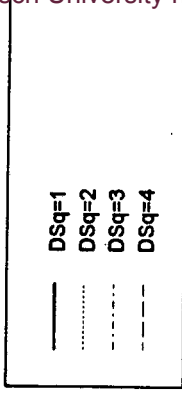
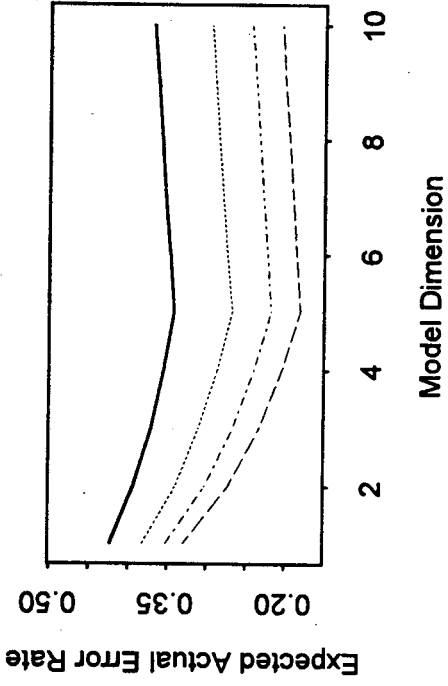
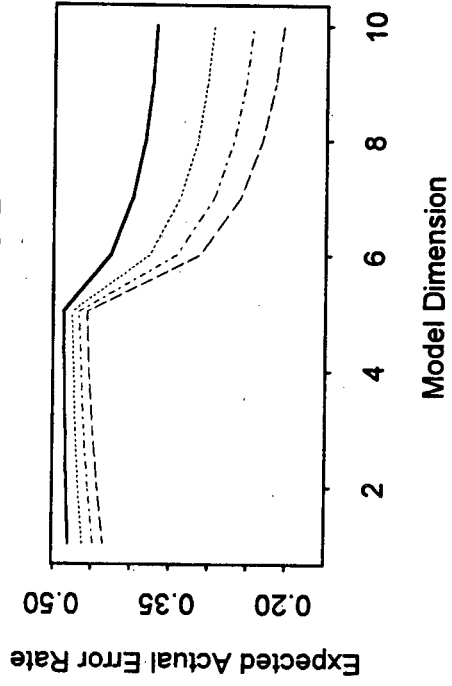


FIG. 3.2: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, NORMAL DATA, LARGE SAMPLES, $r = 1$

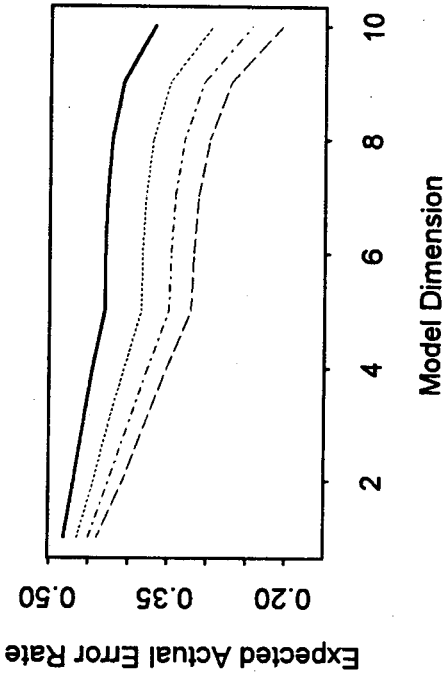
CASE NS22



CASE NS42



CASE NS12



CASE NS32

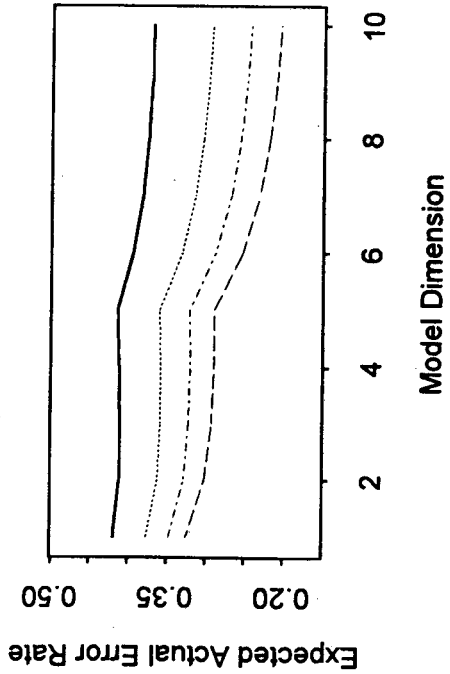
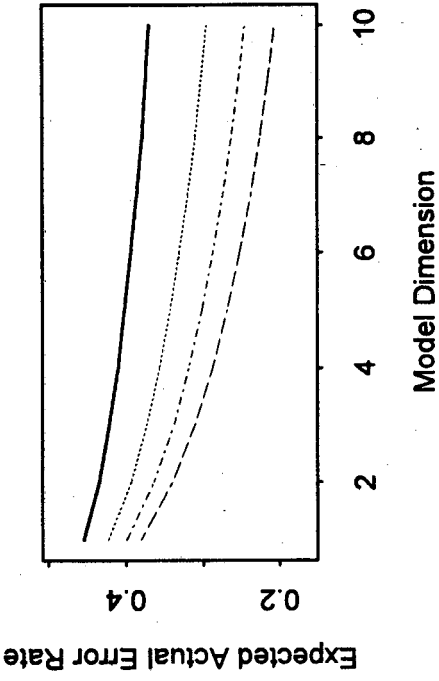
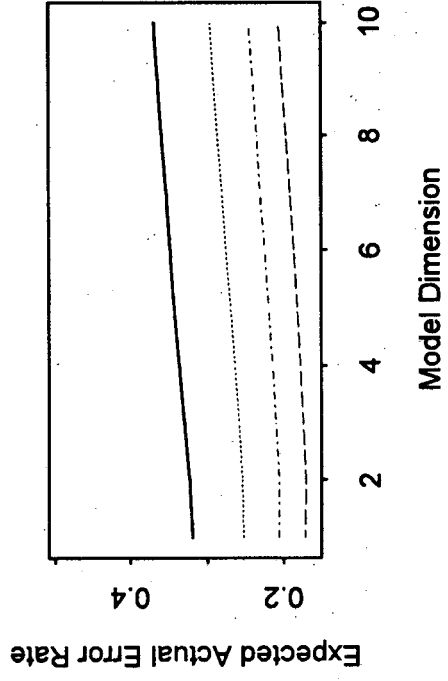


FIG. 3.3: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, NORMAL DATA, SMALL SAMPLES, $\tau = 5$

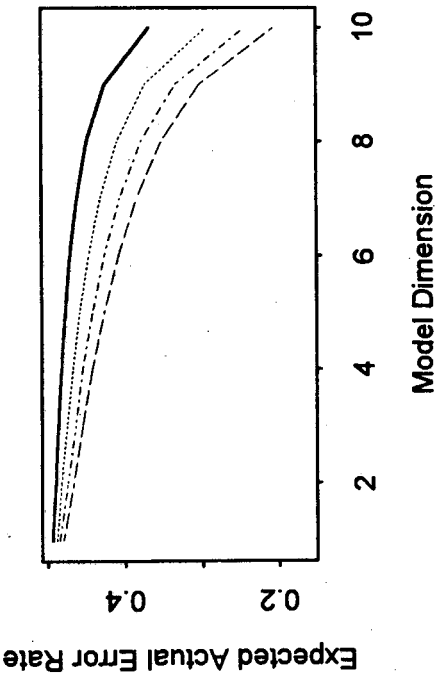
CASE NS23



CASE NS43



CASE NS13



CASE NS33

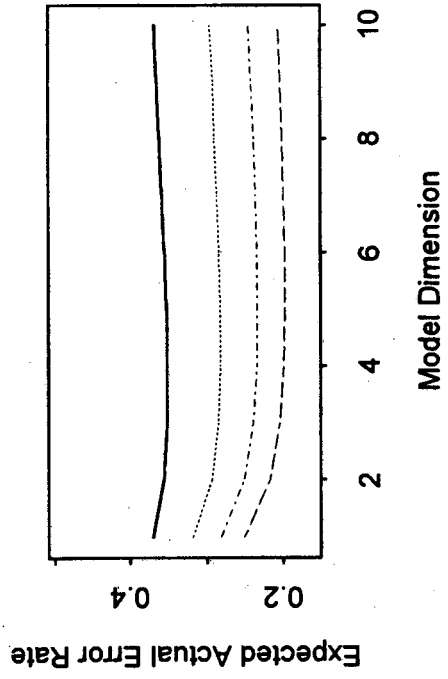


FIG. 3.4: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, NORMAL DATA, SMALL SAMPLES, $r = 10$

3.3.2 THE LOGNORMAL CASE

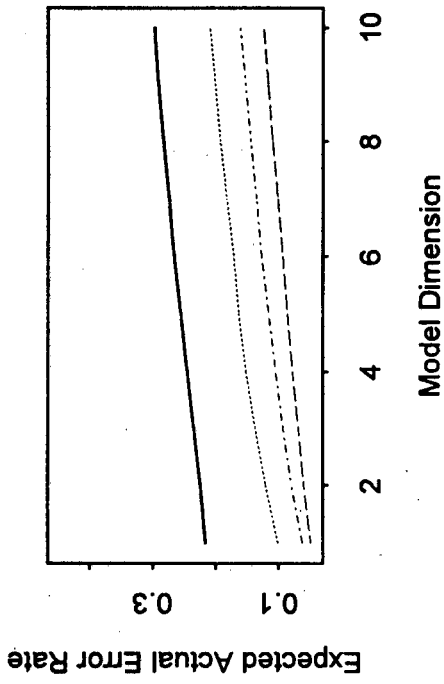
For lognormal feature variables, the actual error rates associated with $W_p(\mathbf{x}; \mathbf{t}(\mathcal{J}))$ were obtained by means of simulation. To estimate the required unconditional error rates, 5000 Monte Carlo repetitions were used. For each repetition a training data set was generated from the two relevant lognormal distributions and the Anderson classification statistics $W_p(\mathbf{x}; \mathbf{t}(\mathcal{J}))$ were calculated for $p = 1, \dots, k$. To estimate the actual error rate associated with each $W_p(\mathbf{x}; \mathbf{t}(\mathcal{J}))$, $p = 1, \dots, k$, a large number (1000) of cases from each group were generated independently of the training data, and classified using the classification statistic $W_p(\mathbf{x}; \mathbf{t}(\mathcal{J}))$, $p = 1, \dots, k$. To obtain estimates of the expected actual error rate, the actual error rates associated with each dimension p , $p = 1, \dots, k$, were averaged over the 5000 Monte Carlo repetitions.

The results of the simulation study were summarised by means of graphs, of which a representative selection appears in Figs. 3.5 - 3.8. Since the results for the large sample cases are largely similar to those for the small sample cases, both large and small sample results are only given for the case where $r = 1$ (see Figs. 3.5 and 3.6). For the cases where $r = 5$ and $r = 10$, only the small sample cases are shown (see Figs. 3.7 and 3.8). Each of the graphs in these figures shows the unconditional error rate as a function of p for one of the lognormal cases defined above. Four values of $\Delta^2 = \Delta_k^2$, the squared Mahalanobis distance between Π_0 and Π_1 , based on all k feature variables, are represented in every graph. Perusal of these graphs leads to the following conclusions.

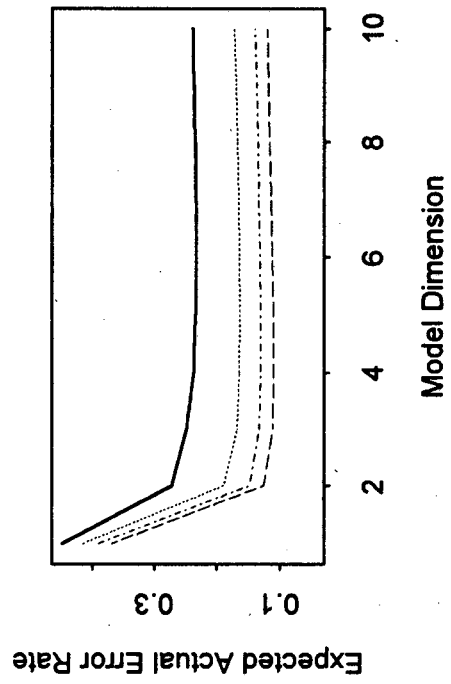
1. If the feature variables are uncorrelated, the unconditional error rate is a minimum at $p = r$, i.e. all the seemingly relevant variables and none of the seemingly irrelevant variables should be included in the classification function (see Fig. 3.5 for case LS21, Fig. 3.6 for case LL21, Fig 3.7 for case LS22 and Fig 3.8 for case LS23).
2. If the feature variables are positively correlated and $r < k$, the error rate decreases when one or more seemingly irrelevant variables are included in the classification function. This effect becomes more pronounced as the correlation increases (see Fig. 3.5 for cases LS31 and LS41, Fig. 3.6 for cases LL31 and LL41, and Fig. 3.7 for cases LS32 and LS42).
3. For the cases where $\rho = 0.9$ and $r = 1$ or 10 (cases LS41, LL41 and LS43), the unconditional error rate reaches a maximum at $p = r$, i.e. the worst possible option is to use a classification rule based on all the seemingly relevant variables, without any seemingly irrelevant variables. As in the normal case, there is a sharp reduction in the unconditional error rate for the cases where $\rho = 0.9$ when a single seemingly irrelevant variable is added to the classification function containing all the seemingly relevant variables (see Fig. 3.5 for case LS41, Fig. 3.6 for case LL41 and Fig. 3.7 for case LS42).

4. When comparing graphs for the cases where $\rho = 0.4$, to graphs of cases where $\rho = 0$ and $\rho = 0.9$, it is evident that the change in error rate behaviour from $\rho = 0$ to $\rho = 0.9$ takes place gradually.
5. If the feature variables are negatively correlated, the unconditional error rate is a minimum at $p = r$, irrespective of the value of r . (see Fig. 3.5 for case LS11, Fig. 3.6 for case LL11, Fig. 3.7 for cases LS12 and Fig. 3.8 for case LS13).
6. The minimum unconditional error rate that is achieved by appropriately choosing p in uncorrelated cases, is lower than the corresponding values for correlated cases.
7. An increase in sample size and in the value of Δ_k^2 lead to a decrease in the expected actual error rates.

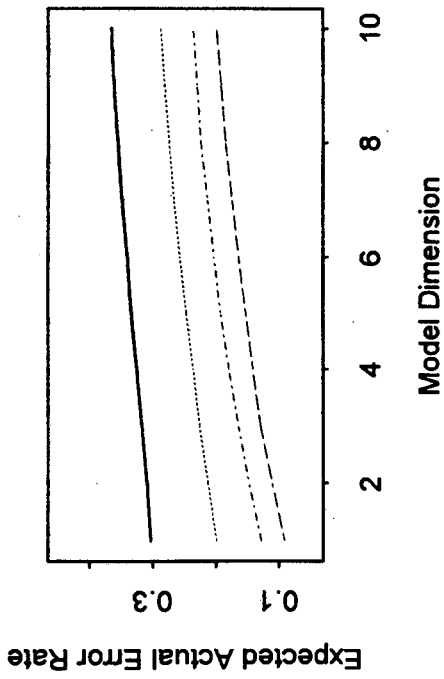
CASE LS21



CASE LS41



CASE LS11



CASE LS31

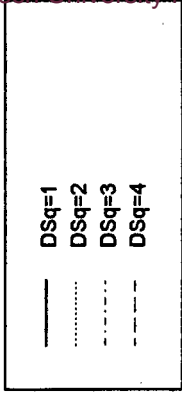
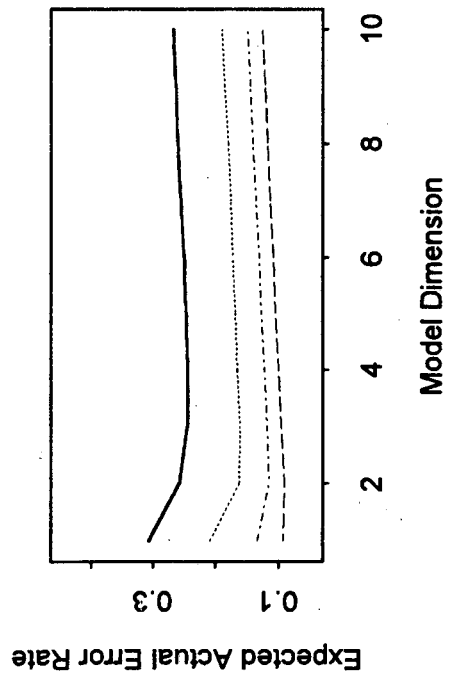
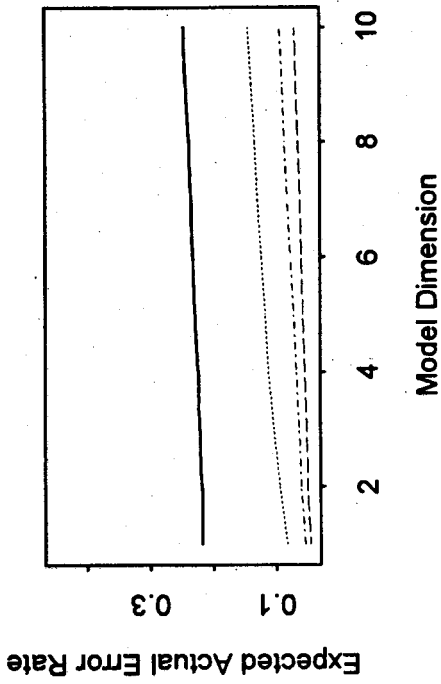
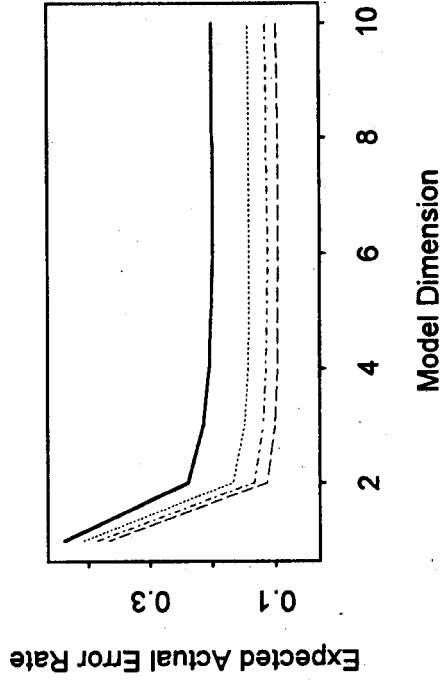


FIG. 3.5: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, LOGNORMAL DATA, SMALL SAMPLES, $r = 1$

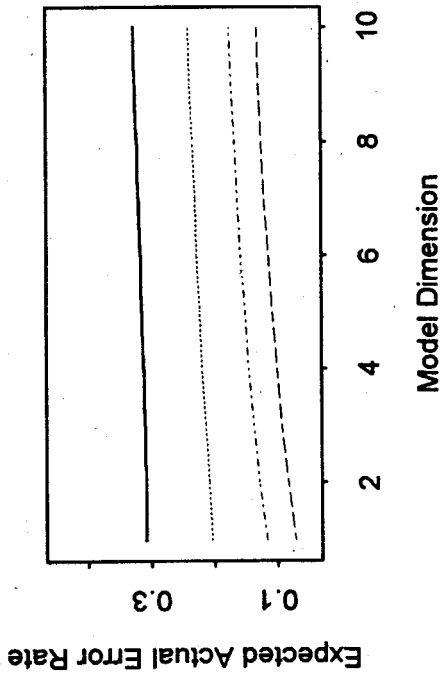
CASE LL21



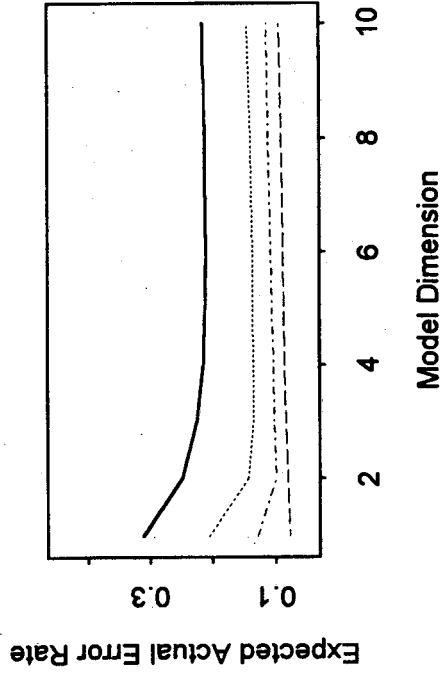
CASE LL41



CASE LL11



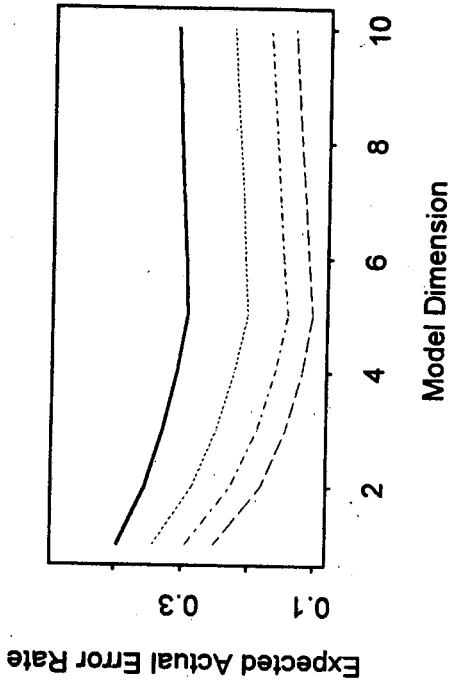
CASE LL31



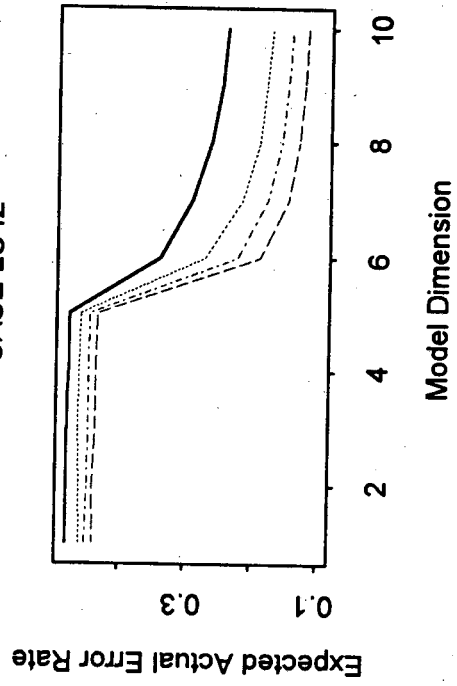
DSq=1
DSq=2
DSq=3
DSq=4

FIG. 3.6: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, LOGNORMAL DATA, LARGE SAMPLES, $\tau = 1$

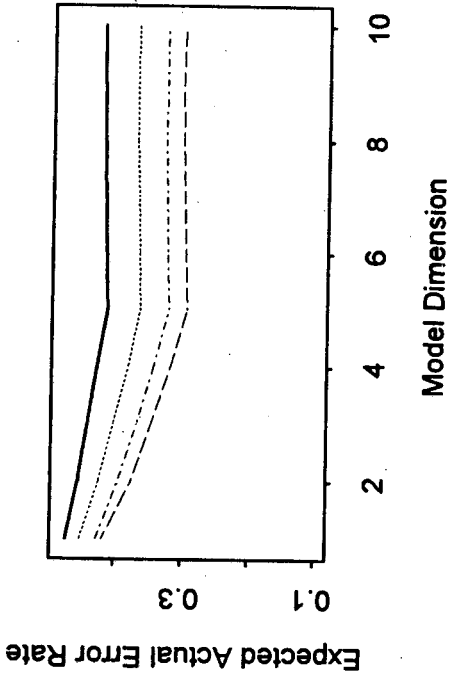
CASE LS22



CASE LS42



CASE LS12



CASE LS32

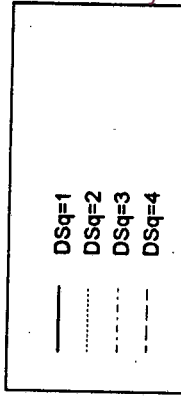
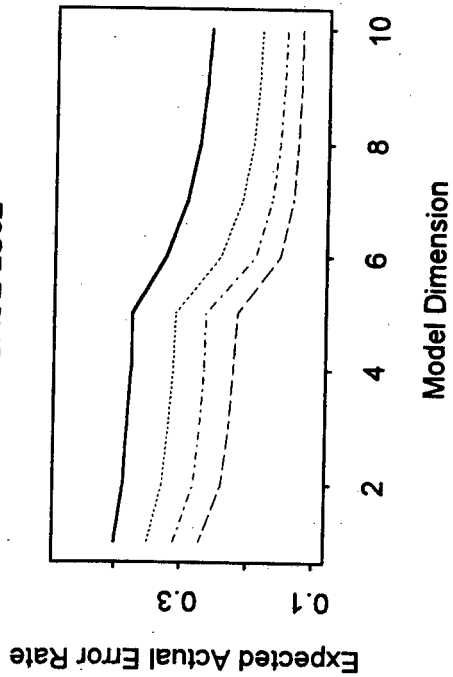
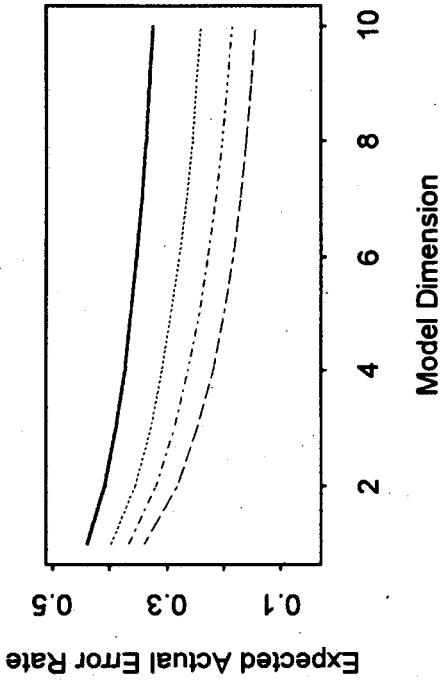
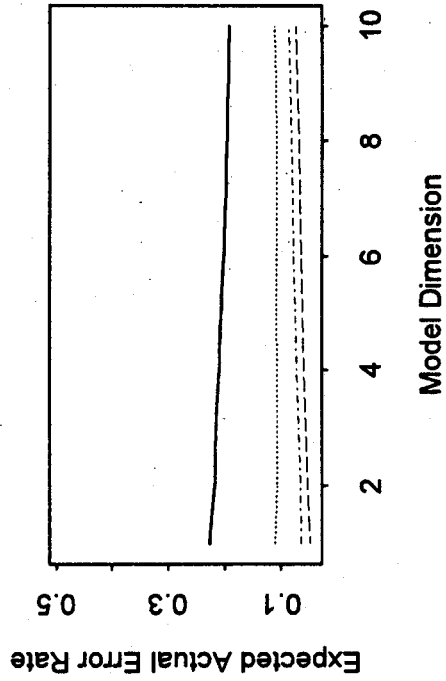


FIG. 3.7: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, LOGNORMAL DATA, SMALL SAMPLES, $r = 5$

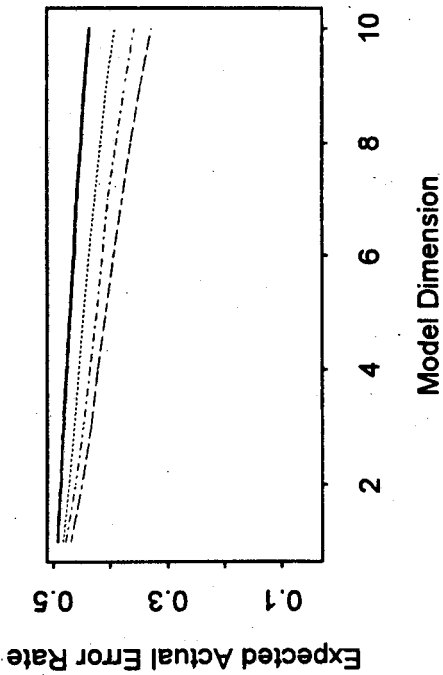
CASE LS23



CASE LS43



CASE LS13



CASE LS33

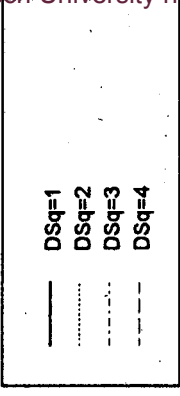
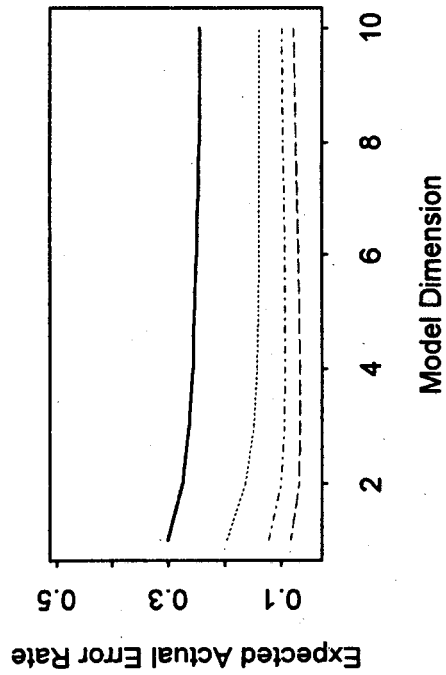


FIG. 3.8: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, LOGNORMAL DATA, SMALL SAMPLES, $r = 10$

3.4 COMPARISON OF DIFFERENT METHODS TO SELECT A PRE-SPECIFIED NUMBER OF VARIABLES

As pointed out in the introduction to Chapter 3, the first stage of variable selection often consists of identifying for each possible model dimension a subset of the available variables that is in some sense optimal. This is followed during the second stage by making a unique choice from these optimal models of different dimensions. In this chapter the first stage of the process is emphasised within a discriminant analysis context. Four criteria that can be used to identify an optimal subset of given size of the available variables are now compared within the following setting.

Consider a two-group situation, with populations Π_0 and Π_1 , and suppose that $k = 10$ feature variables have been observed for the entities in samples of sizes n_0 and n_1 from these two populations respectively. Assume further that the two populations differ from each other only with respect to the first $r = 5$ feature variables. Suppose that each of the four selection criteria is applied to the available data to identify an optimal subset of five feature variables. In this section the actual error rates of the discriminant functions based on the subsets identified by each of the criteria will be investigated in a simulation study. The aim is to reduce the number of potential criteria, with a view to a much more extensive study along the same lines, which will be described in Section 3.5.

Selection criteria from the separatory as well as the allocatory class are investigated in this section. If only models of a fixed dimension are considered, as is the case in this section, all the separatory criteria such as R^2 , C_p and F-based criteria, are equivalent.

Therefore R^2 is the only member of this class that will be included in the study. Using different error rate estimators to select a subset of fixed size from the available feature variables does not in general lead to the same variables being selected. The following error rate estimators were therefore included in the study as representative examples from the allocatory class of selection criteria: the *apparent error rate* (cf. (2.2.13)), the *leave-one-out error rate* (cf. (2.2.15)), and the *posterior probability error rate estimator* (cf. (2.2.19)). Each of the criteria was used in an all possible subsets approach to identify a best subset (i.e. the subset with the maximum value of R^2 or the minimum value of each of the three error rate estimators) containing five variables. The Anderson classification statistics based on the variables in these subsets, are denoted by $W_5(\mathbf{x}; \mathbf{t}(J_i(\mathbf{t})))$, $i = 1, 2, 3, 4$, referring to selection by means of R^2 , the apparent error rate, the leave-one-out error rate and the posterior probability error rate estimator, in that order.

Details of the Monte Carlo simulation study that was undertaken to evaluate the performance of the selection criteria in terms of the estimated expected actual error rate of the resulting discriminant functions, are now provided. Two distributions for the feature variables were used, viz. the normal distribution and the lognormal

distribution. In each case, two sample sizes were considered: $n_0 = n_1 = 25$ (small samples) and $n_0 = n_1 = 100$ (large samples). As in the previous section, the coding NS and NL will be used to denote the small sample and large sample normal cases respectively, while LS and LL will be used similarly for the case of lognormal feature variables. Regarding the covariance structure, the matrices $\Sigma = \mathbf{I}$ and Σ given by (2.4.1) with $\rho = 0.9$ were used. The values $k = 10$ and $r = 5$ were used throughout. Using coding similar to that in Section 3.3, the cases studied in this section will be referred to as NS22, NS42, NL22 and NL42, with similar coding for the lognormal case. The cases denoted by e.g. NS11 - NS13, NS23, NS31 - NS33, NS41 and NS43 in Section 3.3, are not studied in this section, but are included in the extended study described in Section 3.5.

It is assumed that the feature vector \mathbf{X} has mean vector $\mu_0 = \mathbf{0}$ in Π_0 , and that the first $r = 5$ elements of μ_1 , the mean vector of \mathbf{X} in Π_1 , differ from zero. The same parameterisation used in Section 3.3 for the cases where $r = 5$, was used for the elements of μ_1 :

$$\mu_{1l} = \begin{cases} \Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r \sigma^{ij}} & , \quad l = 1, \dots, 5 \\ 0 & , \quad l = 6, \dots, 10 \end{cases}$$

The values $\Delta^2 = 0, 1, 2, 3, 4, 6, 9$ were used for the squared Mahalanobis distance between the two populations based on all the available feature variables. The factors discussed above, identify a total of eight different cases. For each case, the expected actual error rates associated with $W_5(\mathbf{x}; t(J_i(t)))$, $i = 1, 2, 3, 4$, were estimated at each value of Δ^2 , using simulation.

3.4.1 THE NORMAL CASE

For normally distributed feature variables, (2.2.9) was used to calculate the actual error rate associated with $W_5(\mathbf{x}; t(J_i(t)))$, $i = 1, 2, 3, 4$. In each case the quantities in (2.2.9) were calculated using only the five variables with indices in $J_i(t)$, $i = 1, 2, 3, 4$. To estimate the expected actual error rates, 5000 Monte Carlo repetitions were used. For each repetition a training data set was generated from the two relevant normal distributions. Each of the four selection criteria was then applied to this training data set to select a best subset containing five variables. At each value of Δ^2 , the actual error rates associated with the Anderson classification statistic $W_5(\mathbf{x}; t(J_i(t)))$ based on each of these selected best subsets, were calculated from (2.2.9) for $i = 1, 2, 3, 4$. To estimate the expected actual error rate associated with each $W_5(\mathbf{x}; t(J_i(t)))$, the relevant 5000 actual error rates were averaged.

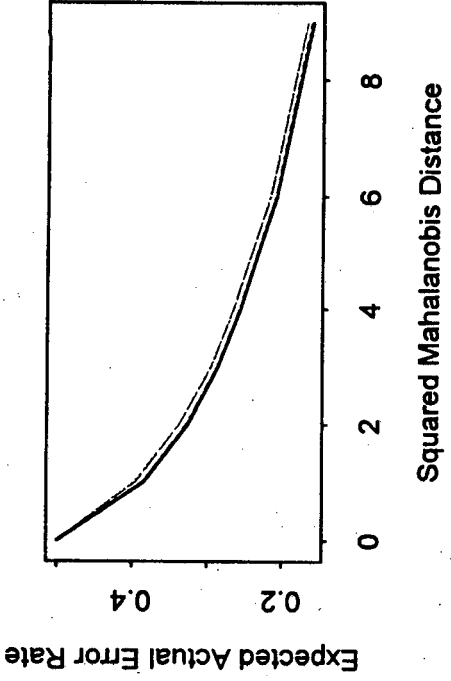
The results of the simulation study are displayed in Fig. 3.9, and will now be discussed.

The expected actual error rate associated with $W_5(\mathbf{x}; t(J_1(t)))$ (R^2 -based selection) is generally the lowest, while the error rate associated with $W_5(\mathbf{x}; t(J_4(t)))$ (selection by means of the posterior probability error rate estimator) is the same as that of $W_5(\mathbf{x}; t(J_1(t)))$ in case NL42, and only slightly higher in the other cases. Especially in cases NS22 and NL22 (corresponding to cases where the feature variables are uncorrelated), the error rates associated with $W_5(\mathbf{x}; t(J_i(t)))$, $i = 2, 3$ (selection by means of the apparent error rate and leave-one-out error rate respectively) are considerably higher than that of $W_5(\mathbf{x}; t(J_i(t)))$, $i = 1, 4$. An increase in the sample sizes and/or the introduction of correlation, reduce these differences.

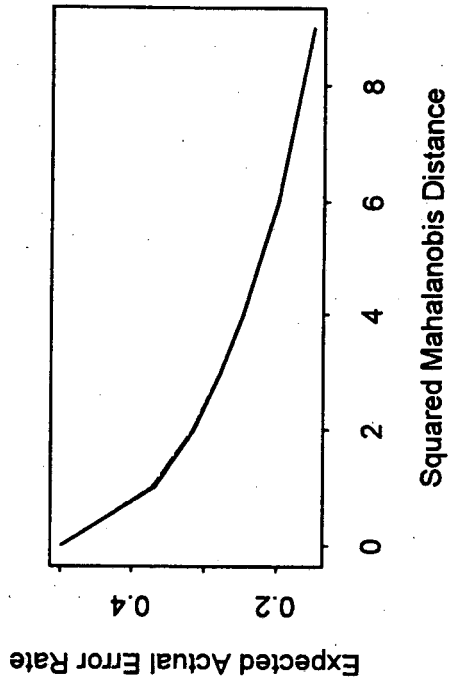
A problem that arises when applying the apparent error rate as selection criterion, is that it often happens that more than one subset of the prescribed size yield the same minimum apparent error rate, due to the 0-1 loss function employed when calculating this estimator. In such cases, a unique best subset cannot be identified. This is a serious problem, especially in small sample cases. The same is also true for selection based on the leave-one-out error rate estimator (or any other error rate estimator using a 0-1 loss function). This problem does not arise when using the posterior probability error rate estimator (or any other smoothed error rate estimator) as selection criterion. An added advantage of the posterior probability error rate estimator is that it utilises more information than estimators based on a 0-1 loss function (cf. Habbema and Hermans, 1977).

Based on the results of the simulation study as well as on the discussion above, it was decided to include R^2 and the posterior probability error rate estimator as selection criteria in the case of normal data in the more extensive study reported in Section 3.5.

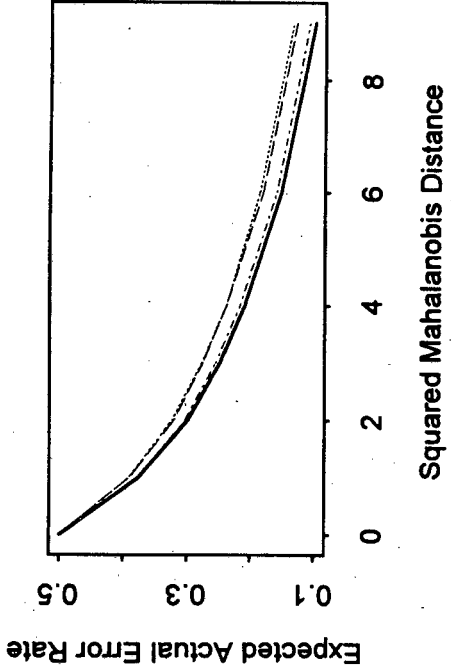
CASE NS42



CASE NL42



CASE NS22



CASE NL22

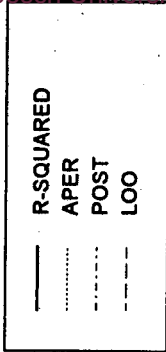
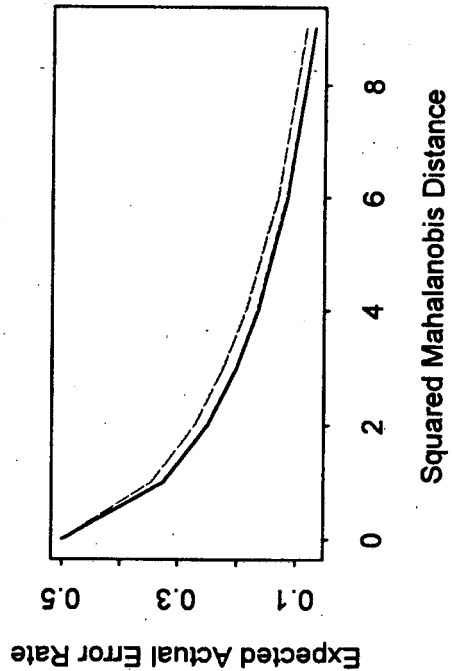


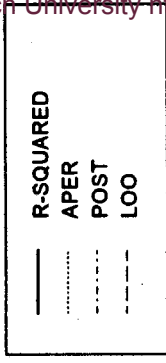
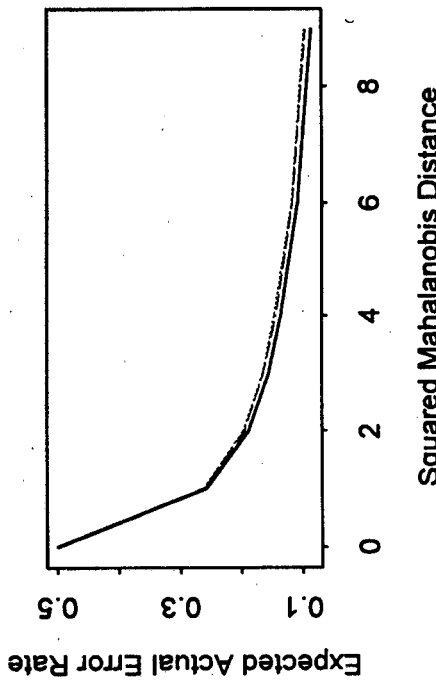
FIG. 3.9: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT SELECTION CRITERIA, NORMAL DATA, $r = 5$

3.4.2 THE LOGNORMAL CASE

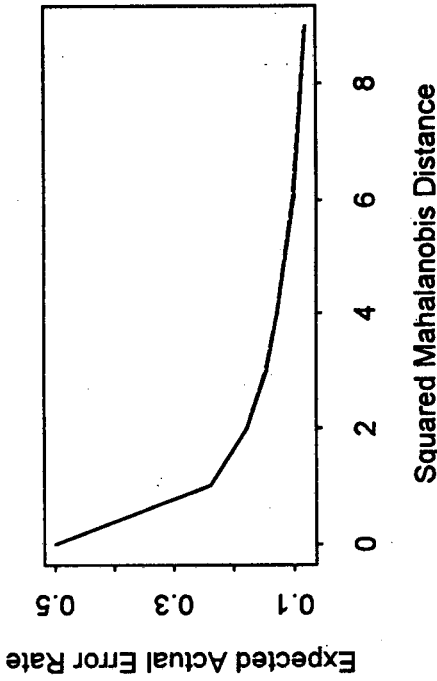
In the cases where the feature variables have a lognormal distribution, the actual error rates associated with $W_5(\mathbf{x}; t(J_i(\mathbf{t})))$, $i = 1, 2, 3, 4$, were estimated by means of Monte Carlo simulation. Five hundred repetitions were used to estimate the expected actual error rates. For each repetition, training data were generated from the two lognormal distributions. Each of the four selection criteria was used to identify the best subset of five feature variables. The actual error rate associated with each classification statistic $W_5(\mathbf{x}; t(J_i(\mathbf{t})))$, was estimated by generating a large number of cases (5000 per group) from the relevant distributions independently of the training data, and then classifying these cases using the classification statistic $W_5(\mathbf{x}; t(J_i(\mathbf{t})))$. The expected actual error rate associated with each $W_5(\mathbf{x}; t(J_i(\mathbf{t})))$, was then estimated by averaging the 500 actual error rates estimated in this way.

The results of this study are displayed in Fig. 3.10. The conclusions are largely the same as in the normal case, but the differences in the error rates associated with $W_5(\mathbf{x}; t(J_i(\mathbf{t})))$, $i = 2, 3$, and those yielded by $W_5(\mathbf{x}; t(J_i(\mathbf{t})))$, $i = 1, 4$, are smaller than in the corresponding normal cases. As in the normal case, the performance of only R^2 and the posterior probability error rate estimator as selection criteria, will be extensively studied in Section 3.5.

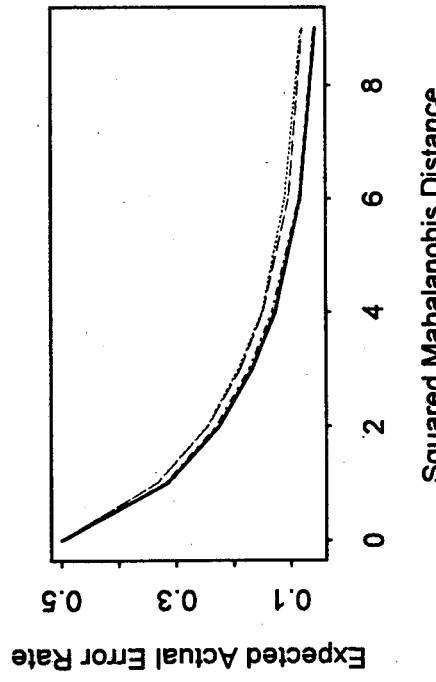
CASE LS42



CASE LL42



CASE LS22



CASE LL22

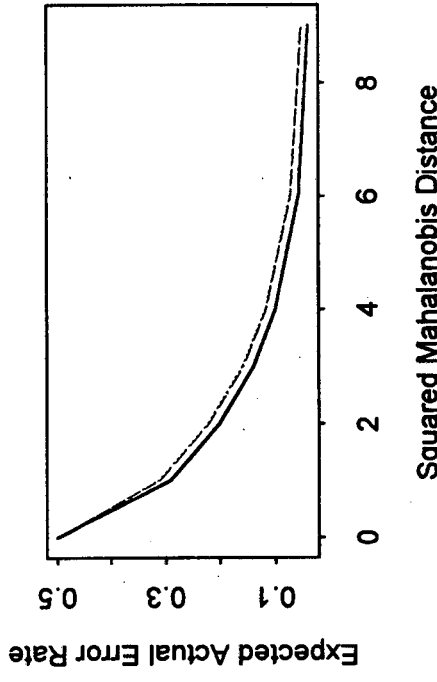


FIG. 3.10: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT SELECTION CRITERIA, LOGNORMAL DATA, $r = 5$

3.5 THE EFFECT OF MODEL DIMENSION ON THE PROPERTIES OF THE RESULTING CLASSIFICATION RULE (WITH SELECTION)

The simulation study described in Section 3.4 was carried out mainly to reduce the number of selection criteria to be included in a more extensive study. In this section, the performance of the two criteria identified in Section 3.4 as being the best in terms of yielding classification statistics with the lowest expected actual error rates, will be investigated in a much more extensive simulation study.

Consider once more two populations, Π_0 and Π_1 , and suppose that training samples of sizes n_0 and n_1 are available from the two populations respectively. A total of k feature variables have been observed on each of these entities. Assume that the two populations differ from each other with respect to r of the k feature variables. In this section each of the two selection criteria chosen in Section 3.4, viz. R^2 and the posterior probability error rate estimator, are applied to the training data to select the best subset of each possible size, $p = 1, \dots, k$. The actual error rates associated with the subsets identified by the two criteria will be investigated. The aim is twofold: firstly, to compare the error rate performance of the classification rules based on the subsets selected by the two criteria and secondly, to obtain insight into the manner in which the post-selection expected actual error rate varies with the number of selected variables, in the hope that this insight can be fruitfully employed in Chapter 4, where the construction of a new selection strategy for discriminant analysis is discussed.

3.5.1 COMPARISON OF POST-SELECTION ERROR RATES

The first aim now receives attention. The two selection criteria included in this study emphasise different aspects: R^2 -based selection concentrates on variables that best separate the two populations, while selection by means of the posterior probability error rate estimator focuses on variables that minimise this error rate estimator. The limited study discussed in Section 3.4 indicated that the expected actual error rates associated with $W_p(\mathbf{x}; \mathbf{t}(J_1(\mathbf{t})))$ (the Anderson classification statistic based on the best 5-dimensional subset selected by means of R^2) are slightly lower than that associated with $W_p(\mathbf{x}; \mathbf{t}(J_4(\mathbf{t})))$ (based on variables selected by means of the posterior probability error rate estimator). The aim is firstly to determine whether this is also the case for a wider range of situations. In this simulation study, $k = 10$ is used throughout, but $r = 1$, $r = 5$ and $r = 10$ are used. With respect to the correlation structure, $\Sigma = \mathbf{I}$ and Σ given by (2.4.1) are used, but a wider range of correlation is included, viz. $\rho = -0.1, 0, 0.4, 0.9$. In Section 3.4, the criteria were only required to select a best subset containing five variables, whereas subsets of each possible dimension $p = 1, \dots, k$, are selected by each criterion in this section. Once more, the normal and lognormal distributions are used as underlying distributions, and sample sizes

$n_0 = n_1 = 25$ (small samples) and $n_0 = n_1 = 100$ (large samples) are used. The same coding introduced in Section 3.3 is used to refer to the 48 cases identified by these factors. The same parameterisation as in Section 3.3 is used for the mean vectors of the two populations, viz. $\mu_0 = \mathbf{0}$, and for cases where $r = 1, 5$,

$$\mu_{1l} = \begin{cases} \Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r \sigma^{ij}} & , \quad l = 1, \dots, r \\ 0 & , \quad l = r+1, \dots, 10, \end{cases}$$

while for $r = 10$:

$$\mu_{1l} = \Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r \sigma^{ij}} , \quad l = 1, \dots, 10.$$

The Δ^2 - values 1, 2, 3, 4 are used for the squared Mahalanobis distance between the two populations, based on k variables.

3.5.1.1 THE NORMAL CASE

For the case where the feature variables are normally distributed, the actual error rates associated with $W_p(\mathbf{x}; t(\mathcal{J}_i(t)))$, $i = 1, 4$; $p = 1, \dots, k$ were obtained by means of simulation. A total of 1000 Monte Carlo repetitions were done. For each repetition, training data were generated from the relevant normal distributions. The two selection criteria were then applied to the training data to select the best subset containing $p = 1, \dots, k$ variables. For each size p , the selection is done by considering all possible subsets of that size, and selecting the subset that is best according to the criterion (i.e. the subset that maximises R^2 or the subset that minimises the posterior probability error rate estimator). The advantage of using an all possible subsets approach instead of a stepwise procedure, is that it ensures that the best subset in terms of the criterion is found, while in any stepwise procedure only some of the possible subsets are considered. At each value of Δ^2 , the actual error rates associated with the classification statistics $W_p(\mathbf{x}; t(\mathcal{J}_i(t)))$, $i = 1, 4$; $p = 1, \dots, k$, were calculated using (2.2.9). The expected actual error rates were estimated by averaging the 1000 actual error rates obtained for each p ($p = 1, \dots, k$) and each i ($i = 1, 4$). A selection of the results obtained for the small sample normal cases is displayed in Figs. 3.11 - 3.14. The results for case NS11 are displayed at $\Delta^2 = 1, 2, 3$ and 4 (see Fig 3.11). Since the relative performance of the two classification statistics is largely similar at all values of Δ^2 (as is evident from Fig. 3.11), only the results obtained at $\Delta^2 = 2$ are displayed for the other normal cases (see Figs. 3.12 - 3.14). Perusal of the graphs leads to the following conclusions.

1. For cases where $r = 1$ (see Figs. 3.11 and 3.12), there is very little difference in the expected actual error rates associated with $W_p(\mathbf{x}; t(J_1(t)))$ and that associated with $W_p(\mathbf{x}; t(J_4(t)))$ for case NS11 ($\rho = -0.1$) and NS41 ($\rho = 0.9$). In cases NS21 ($\rho = 0$) and NS31 ($\rho = 0.4$), $W_p(\mathbf{x}; t(J_4(t)))$ yields a slightly lower error rate than $W_p(\mathbf{x}; t(J_1(t)))$. In all cases, the minimum error rates associated with both statistics when p is varied, are approximately equal.
2. For $r = 5$ (see Fig. 3.13), the error rates of both statistics are largely the same, but $W_p(\mathbf{x}; t(J_1(t)))$ performs slightly better in cases NS42 and NS32. Once more, the minimum error rates over p are approximately the same for both rules.
3. In cases where $r = 10$ (see Fig. 3.14), there is very little difference in the error rates for case NS13, while $W_p(\mathbf{x}; t(J_1(t)))$ yields slightly lower error rates in case NS23. In case NS33, the minimum error rate associated with $W_p(\mathbf{x}; t(J_1(t)))$ is the lowest, while $W_p(\mathbf{x}; t(J_4(t)))$ yields lower error rates than $W_p(\mathbf{x}; t(J_1(t)))$ in case NS43, but the minimum error rates over p are approximately the same.

The differences between the error rates for large sample sizes are even smaller than in the small sample cases, and therefore graphs are not shown for the large sample cases. In general, none of the two criteria consistently outperforms the other, in terms of the expected actual error rates yielded by the classification functions based on the selected subsets. To select the best subset for a given dimension, there is very little difference in the expected actual error rates associated with the rules based on the variables selected by means of the two different selection criteria. Since selection using a criterion such as R^2 is much more readily available in standard statistical software packages, use of such criteria can be recommended to find the best subset of a given dimension.

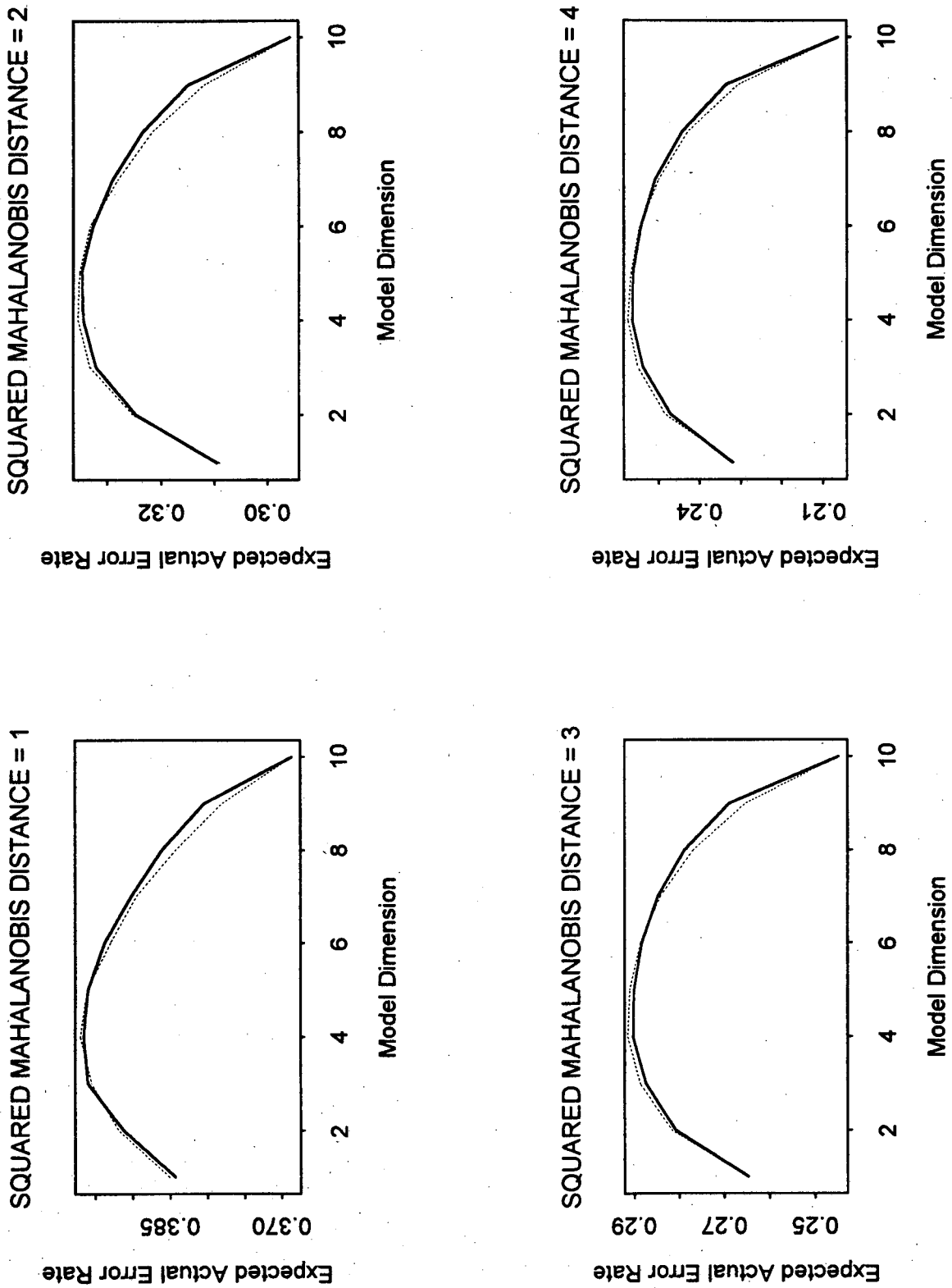


FIG. 3.11: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, CASE NS11

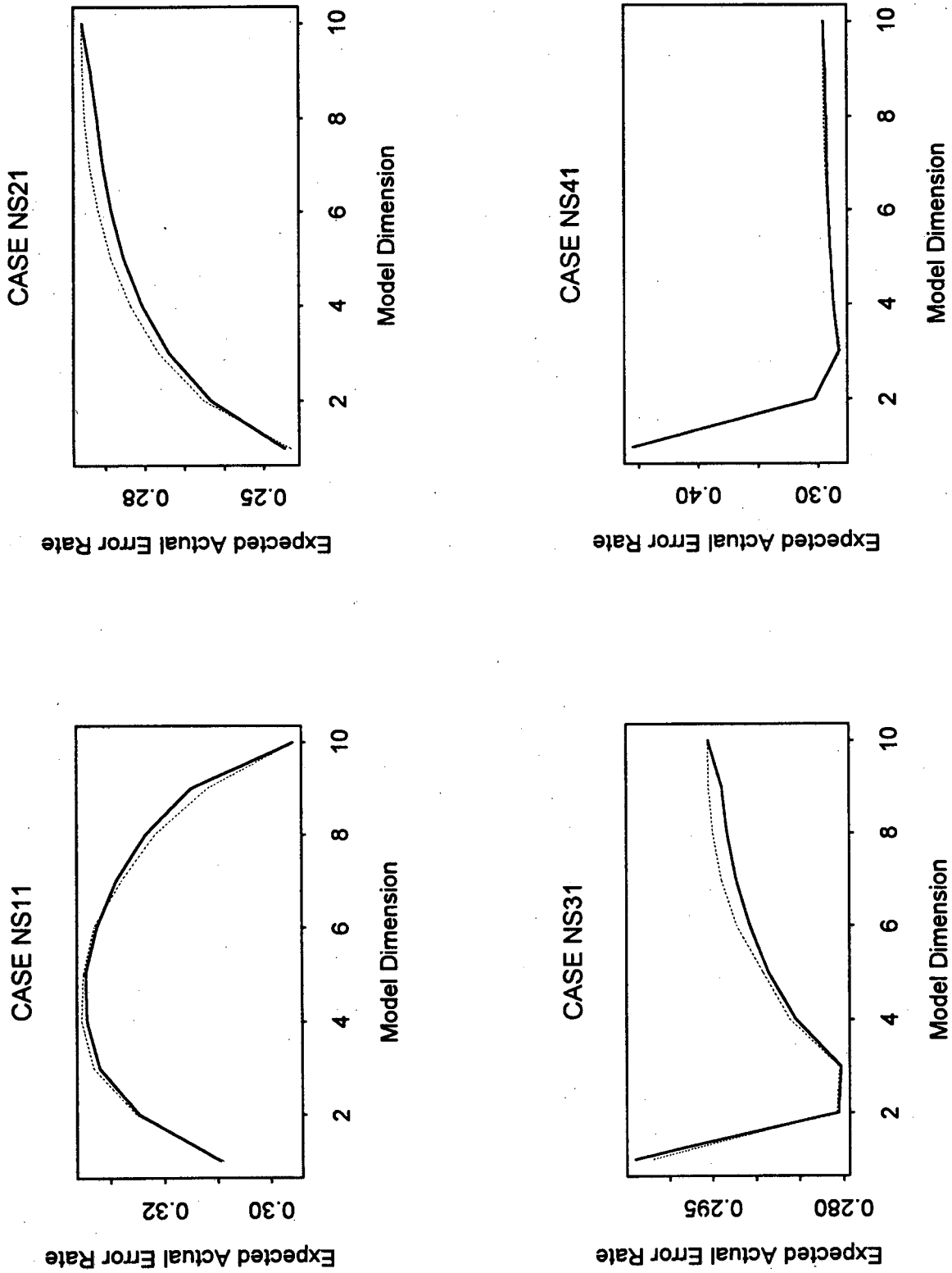
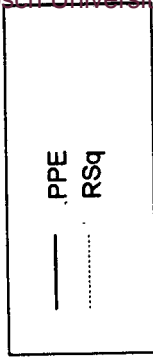
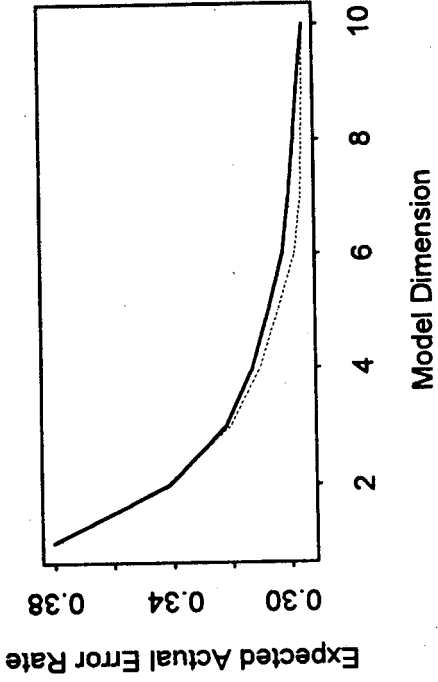
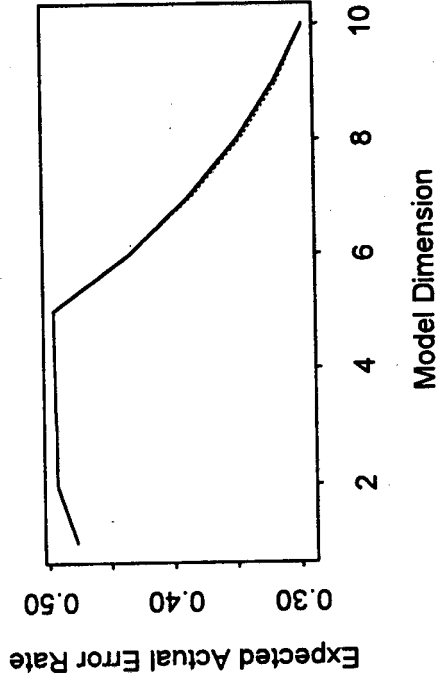


FIG. 3.12: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, NORMAL DATA, $r = 1$

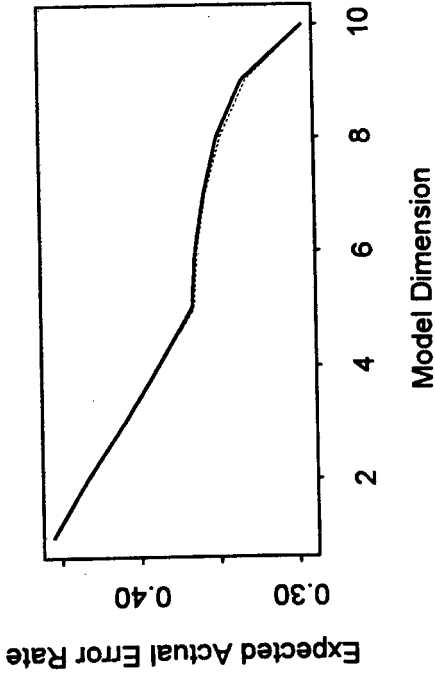
CASE NS22



CASE NS42



CASE NS12



CASE NS32

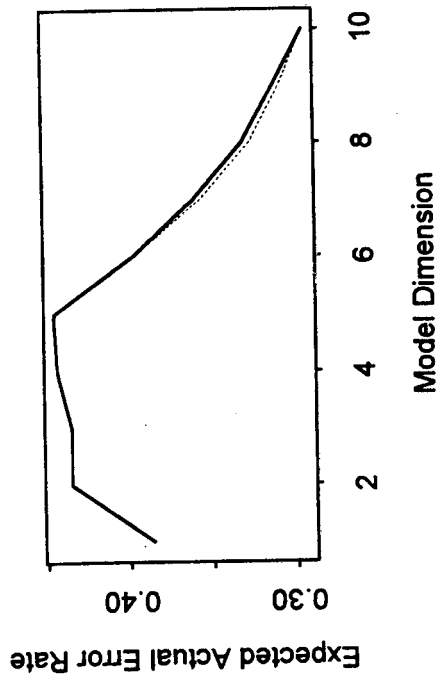
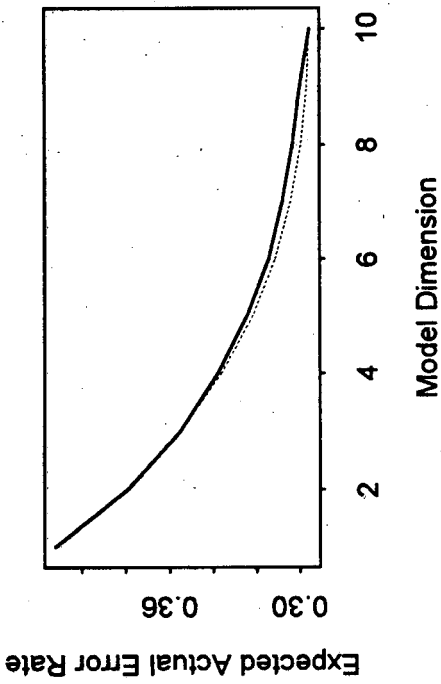
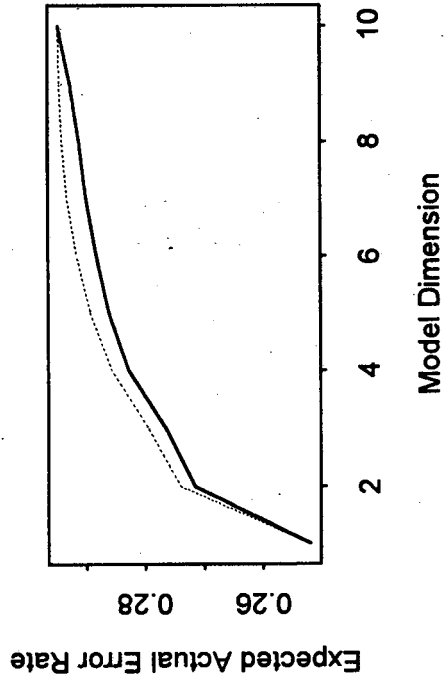


FIG. 3.13: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, NORMAL DATA, $r = 5$

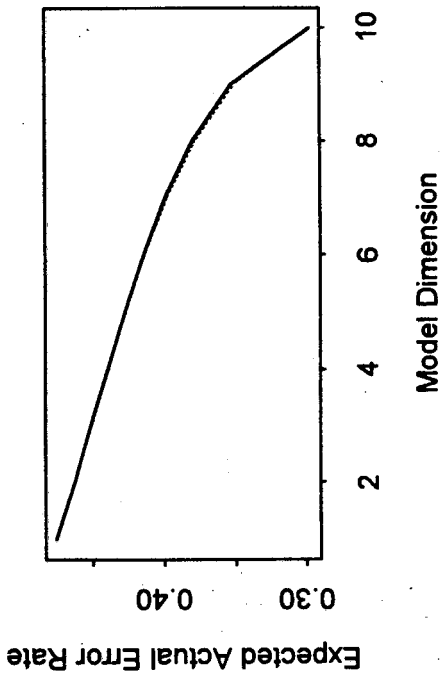
CASE NS23



CASE NS43



CASE NS13



CASE NS33

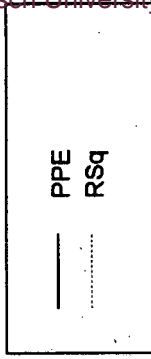
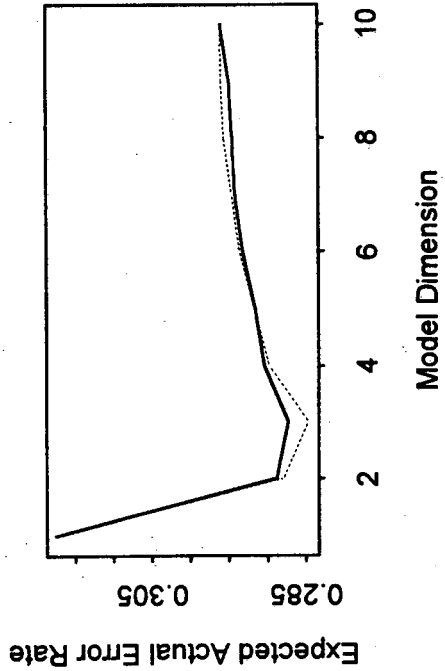


FIG. 3.14: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, NORMAL DATA, $r = 10$

3.5.1.2 THE LOGNORMAL CASE

If the feature variables have a lognormal distribution, the actual error rates associated with $W_p(\mathbf{x}; t(J_i(t)))$, $i = 1, 4$; $p = 1, \dots, k$ have to be estimated by means of simulation. A total of 500 Monte Carlo repetitions were done. For each repetition, training data were generated from the relevant lognormal distributions. The two selection criteria were then applied to the training data to select the best subset containing $p = 1, \dots, k$ variables. At each value of Δ^2 , the actual error rates associated with the classification statistics $W_p(\mathbf{x}; t(J_i(t)))$, $i = 1, 4$; $p = 1, \dots, k$, were estimated by means of simulation. To do this, a large number (2000 per group) of entities were generated from the relevant lognormal distributions, and classified using the classification statistics. The expected actual error rates were estimated by averaging the 500 actual error rates obtained for each p ($p = 1, \dots, k$) and each i ($i = 1, 4$). A representative selection of the results of the small sample lognormal cases is displayed in Figs. 3.15 - 3.18. The following conclusions can be made:

1. In the cases where $r = 1$, there is virtually no difference between the error rates associated with $W_p(\mathbf{x}; t(J_1(t)))$ and $W_p(\mathbf{x}; t(J_4(t)))$ in case LS41. In cases LS11 and LS21 the differences are small and the relative performance of the two statistics changes with dimension. However, the minimum error rate achieved by $W_p(\mathbf{x}; t(J_1(t)))$ is slightly lower than that of $W_p(\mathbf{x}; t(J_4(t)))$. The same is also true for case LS31, but the difference between the two minimum values is larger.
2. For cases with $r = 5$, the difference in the relative performance of the two classification functions is very small in cases LS12, LS32 and LS42, and both achieve approximately the same minimum error rates. In case LS22, $W_p(\mathbf{x}; t(J_1(t)))$ performs considerably better than $W_p(\mathbf{x}; t(J_4(t)))$ and also yields a lower minimum error rate.
3. If $r = 10$, the differences are again very small in cases LS13 and LS23. In cases LS33 and LS43, the error rates associated with $W_p(\mathbf{x}; t(J_4(t)))$ are slightly lower than those of $W_p(\mathbf{x}; t(J_1(t)))$.

As in the normal case, the differences between the error rates achieved by the statistics based on the subsets selected by the two criteria, are even smaller when large samples are taken. There is no criterion that performs best in all the cases considered. The differences in the expected actual error rates of the two statistics are generally small. Selection using a criterion that emphasises the separation between the groups, such as R^2 , can therefore be recommended when comparing different models of the same model dimension. Selection based on these criteria can be performed much more readily with available statistical software packages than selection based on error rate estimators.

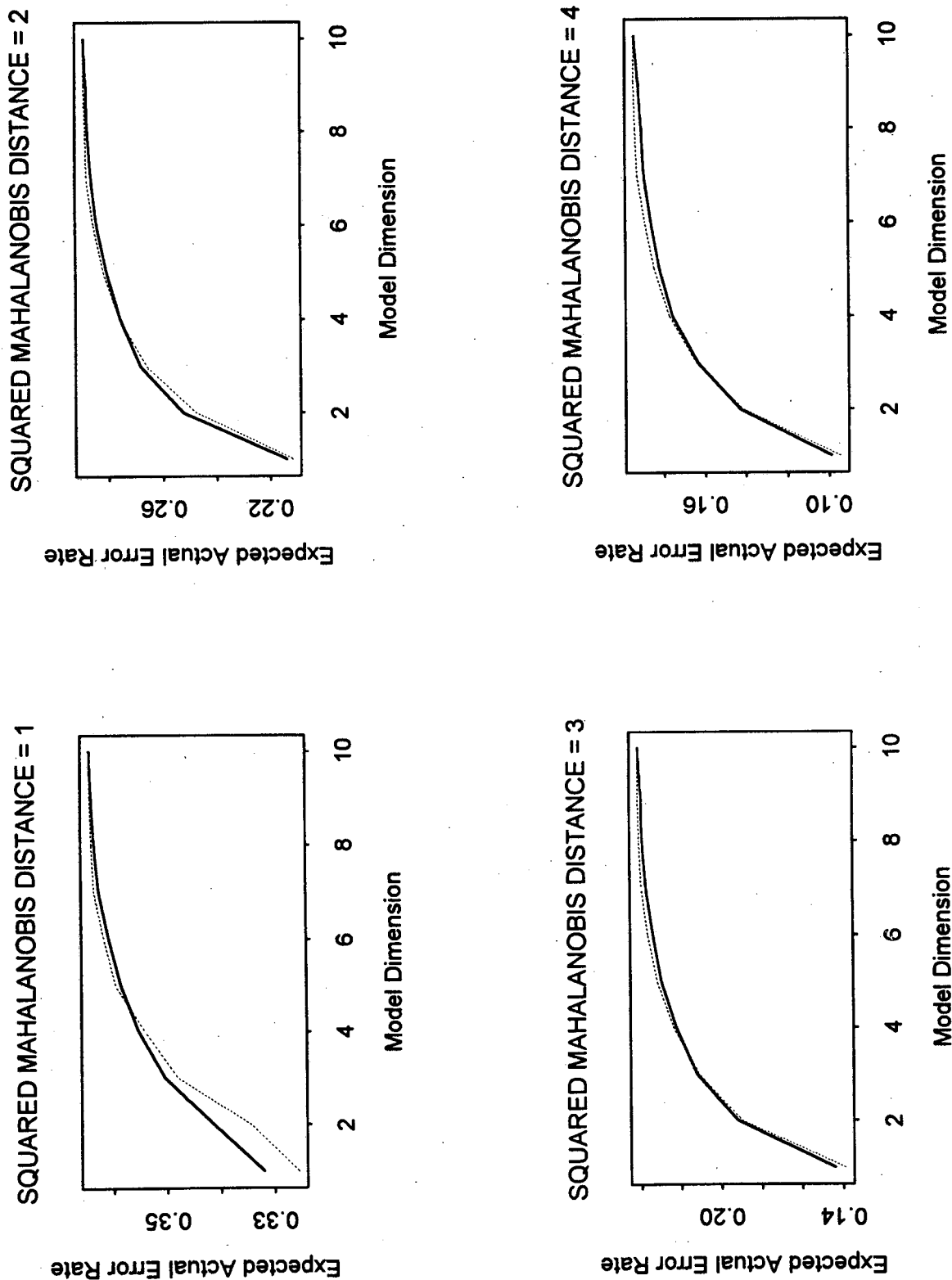
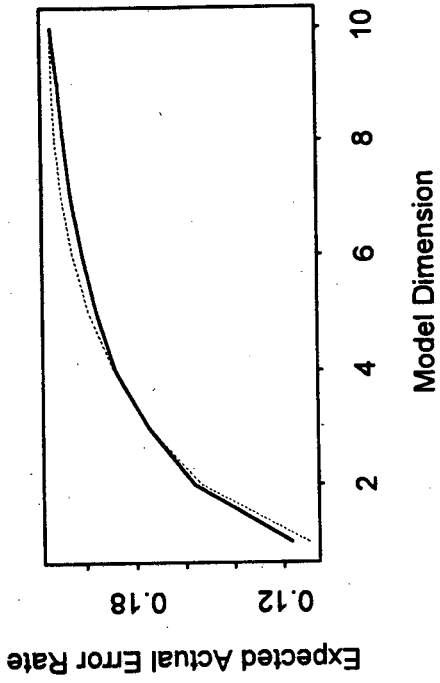
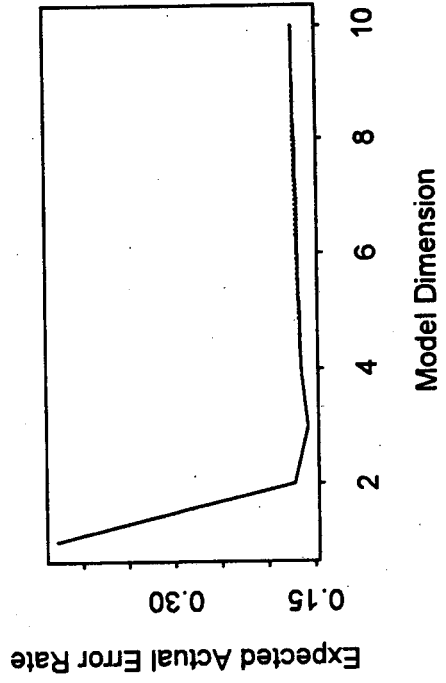


FIG. 3.15: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, CASE LS11

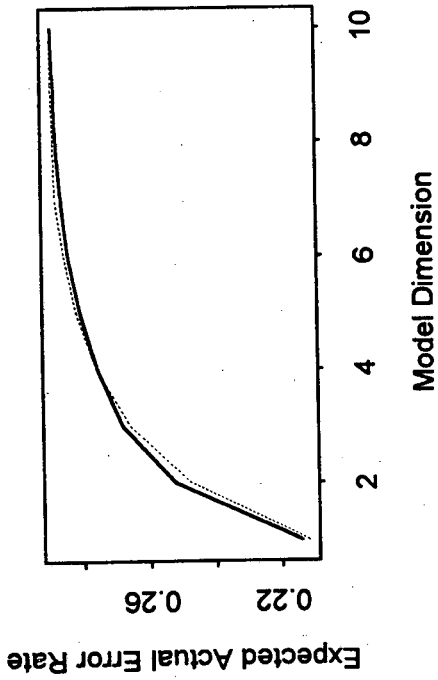
CASE LS21



CASE LS41



CASE LS11



CASE LS31

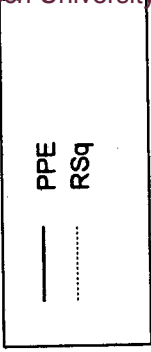
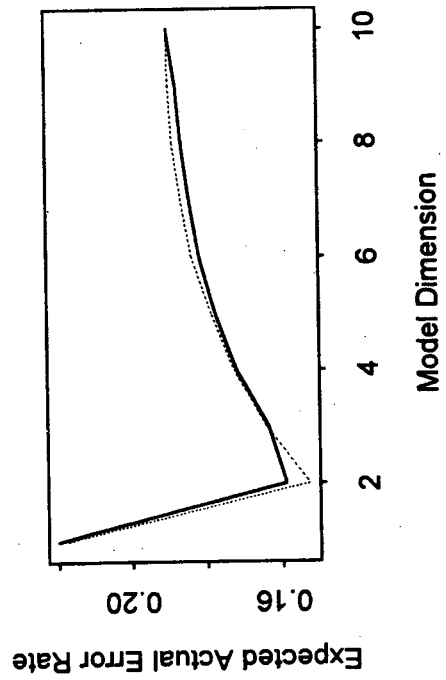
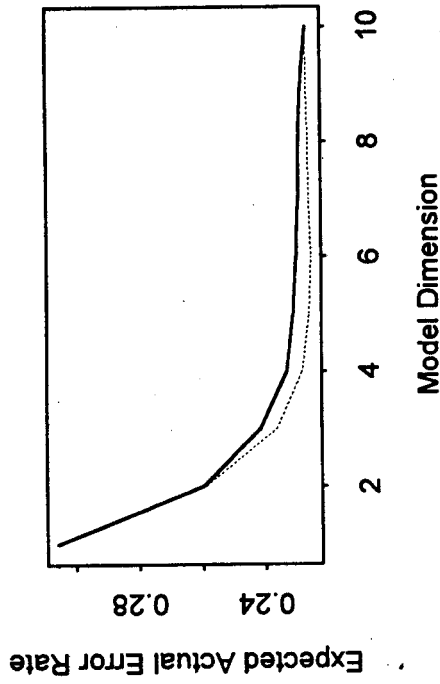
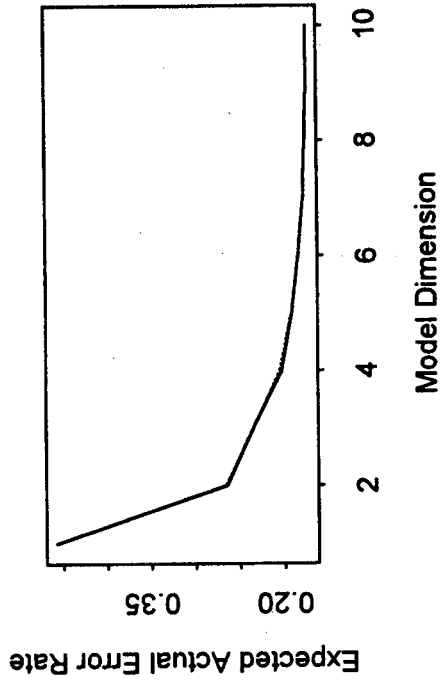


FIG. 3.16: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, LOGNORMAL DATA, $r = 1$

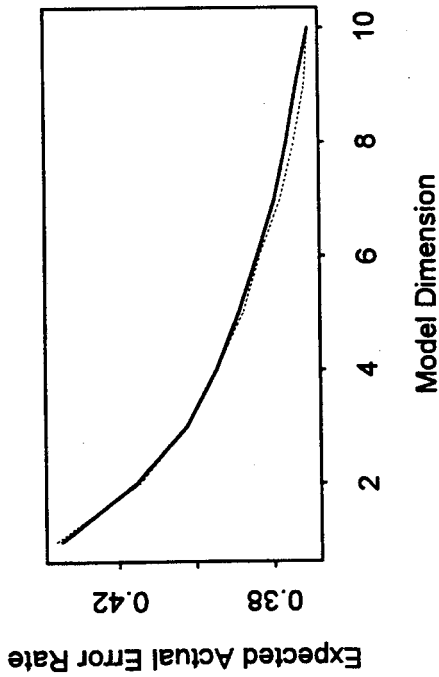
CASE LS22



CASE LS42



CASE LS12



CASE LS32

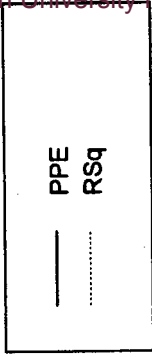
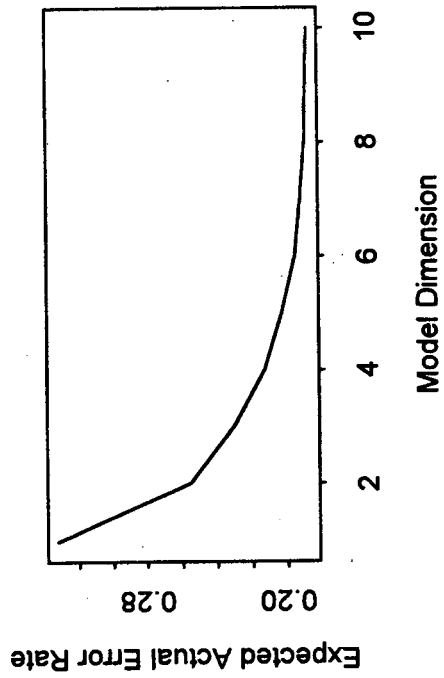
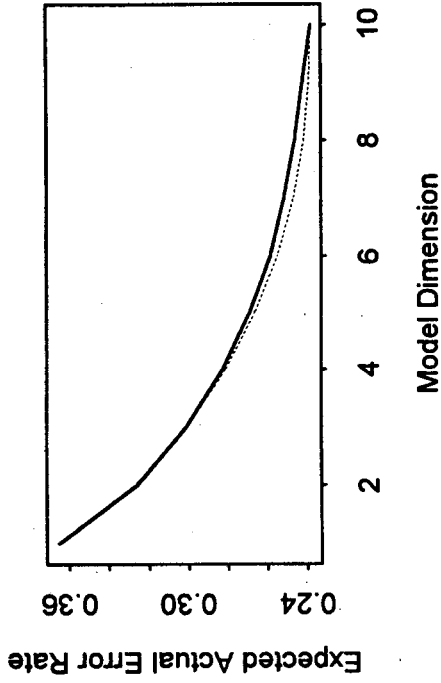
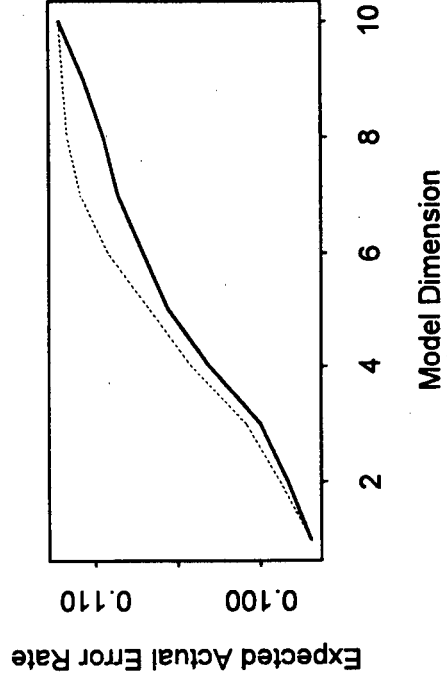


FIG. 3.17: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, LOGNORMAL DATA, $\tau = 5$

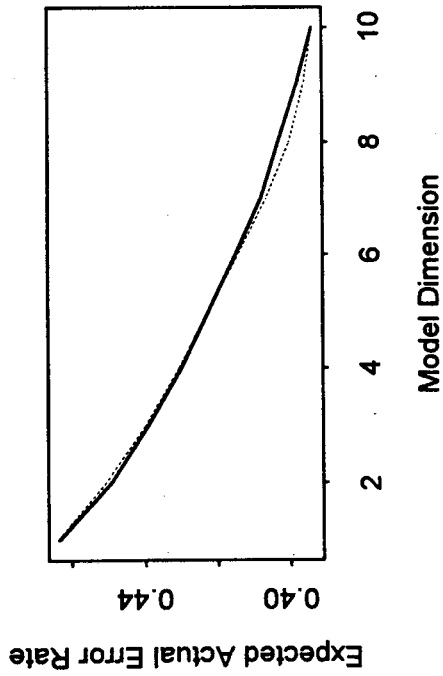
CASE LS23



CASE LS43



CASE LS13



CASE LS33

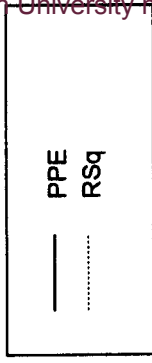
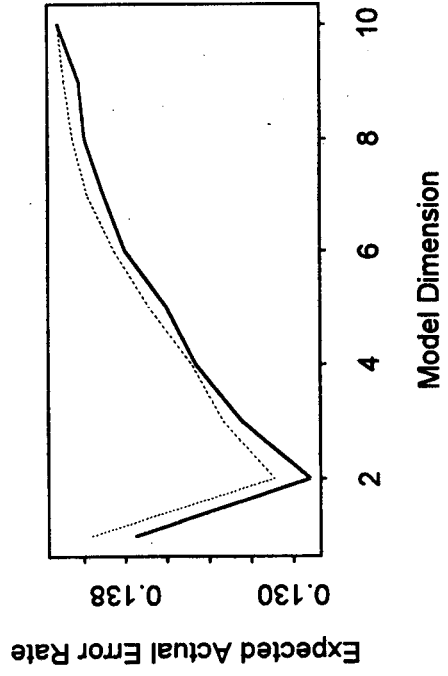


FIG. 3.18: EXPECTED ACTUAL ERROR RATE FOR DIFFERENT MODEL DIMENSIONS, SELECTION CRITERIA: R-SQUARED AND PPE, LOGNORMAL DATA, $r = 10$

3.5.2 THE EFFECT OF DIMENSION ON POST-SELECTION ERROR RATE

Regarding the second aim in this section, the cases $r=1$, $r=5$ and $r=10$ are considered separately.

1. Consider first Fig. 3.12 for the normal distribution and Fig. 3.16 for the lognormal distribution, being the graphs for the cases where $r=1$. From Fig. 3.12 (case NS21) and Fig. 3.16 (case LS21), it is clear that the optimal model dimension when the feature variables are uncorrelated, is $p=r=1$. For a small negative correlation between all the feature variables, the optimal model dimension in the lognormal case is once more $p=r=1$ (see Fig. 3.16 for case LS11), but this is no longer true for the normal case, where $p=10$ yields a lower error rate than $p=1$ (see Fig. 3.12 for case NS11). The difference in error rate at $p=1$ and $p=10$ is however not large in this case, and the question may arise whether it is worthwhile to use the much more complex model with $p=10$ instead of the simple model with $p=1$, which performs almost as well. For moderate and large positive correlation, the optimal model dimension for both the normal and the lognormal distribution is $p=2$ or $p=3$, with error rates at these values of p being appreciably lower than at $p=r=1$ (see Fig. 3.12 for cases NS31, NS41 and Fig. 3.16 for cases LS31 and LS41). Hence, in the case of positively correlated feature variables, inclusion of one or two seemingly irrelevant variables is definitely worthwhile.

2. Next, consider Fig. 3.13 for the normal distribution and Fig. 3.17 for the lognormal distribution, being the graphs for the cases where $r=5$. In the uncorrelated normal case, the error rate is merely a monotone decreasing function of p (see Fig. 3.13 for case NS22). The error rate at $p=r=5$ is however close to the global minimum at $p=10$, and it is once more questionable whether using the most complex model would really be worthwhile. In the lognormal case (see Fig. 3.17 for case LS22), the decrease in error rate beyond $p=r=5$, is very slight or non-existent, and the choice $p=r=5$ or even $p=r-1$, seems satisfactory. For small negative correlation, the optimal choice for both the normal and the lognormal distribution is $p=10$ (see Fig. 3.13 for case NS12 and Fig. 3.17 for case LS12). Especially in the normal case, there is a quite substantial decrease in the error rate when moving from $p=5$ to $p=10$. For moderate or large positive correlation and normal feature data, the choice $p=r=5$ is markedly inferior to a choice of $p>r$. A fairly large value of p (i.e. $p=8, 9$, or 10) would seem to be the optimal choice (see Fig. 3.13 for cases NS32 and NS42). For the corresponding lognormal cases, a much more parsimonious model would seem to be adequate (see Fig. 3.17 for cases LS32 and LS42).

3. Finally, consider Fig. 3.14 for the normal distribution and Fig. 3.18 for the lognormal distribution, being the graphs for the cases where $r=10$. In the uncorrelated cases, the choice $p=r=10$ yields the lowest error rates, but a choice $5 < p < 10$ would not pay too high a price in terms of increased error rate (see Fig.

3.14 for case NS23 and Fig. 3.18 for case LS23). For small negative correlation, the results are similar to those described above for $r = 5$ (see Fig. 3.14 for case NS13 and Fig. 3.18 for case LS13). For moderate positive correlation, the optimal choice in both the normal and lognormal cases is $p = 2$ or $p = 3$ (see Fig. 3.14 for case NS33 and Fig. 3.18 for case LS33). For large positive correlation, the optimal choice in both the normal and lognormal cases is $p = 1$ (see Fig. 3.14 for case NS43 and Fig. 3.18 for case LS43).

3.6 CONCLUSIONS AND RECOMMENDATIONS

Sections 3.3 - 3.5 of this chapter contain a report of an investigation into the influence of the number of variables in the linear discriminant function on its associated expected actual error rate. In Section 3.3 this was done without taking any variable selection into account. The expected actual error rate of the Anderson classification statistic $W_p(\mathbf{x}; t(\mathcal{J}))$ was calculated for $p = 1, \dots, k$, with variables entered in a pre-specified order. This error rate is given by

$$\alpha_{\text{act}}(p; t(\mathcal{J})) = \frac{1}{2} E \left\{ P \left[W_p(\mathbf{X}; t(\mathcal{J})) > 0 \mid \mathbf{X} \in \Pi_0 \right] + P \left[W_p(\mathbf{X}; t(\mathcal{J})) \leq 0 \mid \mathbf{X} \in \Pi_1 \right] \right\}, \quad (3.6.1)$$

where the expectation is taken with respect to the distribution of the training data t .

In Sections 3.4 and 3.5 a pre-specified number of variables was selected using different selection criteria, and the post selection expected actual error rate of the Anderson classification statistic $W_p(\mathbf{x}; t(\mathcal{J}(t)))$ was calculated (for $p = 5$ in Section 3.4 and for $p = 1, \dots, k$ in Section 3.5). This error rate is given by

$$\alpha_{\text{act}}(p; t(\mathcal{J}(t))) = \frac{1}{2} E \left\{ P \left[W_p(\mathbf{X}; t(\mathcal{J}(t))) > 0 \mid \mathbf{X} \in \Pi_0 \right] + P \left[W_p(\mathbf{X}; t(\mathcal{J}(t))) \leq 0 \mid \mathbf{X} \in \Pi_1 \right] \right\}, \quad (3.6.2)$$

where the expectation is once more taken with respect to the distribution of the training data t .

It should be noted that the full effect of selection is not taken into account when (3.6.2) is calculated, since the model dimension is pre-specified and not determined from the training data. The full post selection expected actual error rate of the Anderson classification statistic is given by

$$\alpha_{act}(p(\mathbf{t}); t(\mathcal{J}(\mathbf{t}))) = \frac{1}{2} \mathbf{E} \left\{ P[W_{p(\mathbf{t})}(\mathbf{X}; t(\mathcal{J}(\mathbf{t}))) > 0 | \mathbf{X} \in \Pi_0] + P[W_{p(\mathbf{t})}(\mathbf{X}; t(\mathcal{J}(\mathbf{t}))) \leq 0 | \mathbf{X} \in \Pi_1] \right\} \quad (3.6.3)$$

and this quantity receives attention in Chapter 4.

The conclusions arising from the investigations undertaken in this chapter, can be summarised as follows.

1. When considering whether a given variable should be included into the linear discriminant function, it is wrong to consider the variable on its own, since a variable that does not discriminate well between the two groups, may improve the classification performance of the linear discriminant function when it is added to the variables already in the linear discriminant function. Similarly, a variable that discriminates well when considered on its own, does not necessarily improve the classification performance of the linear discriminant function already containing other variables. These points are illustrated in Section 3.3.
2. Three allocatory criteria were investigated in Section 3.4 in terms of the expected actual error rate when these criteria are used to select a fixed number of variables for inclusion in the linear discriminant function. The expected actual error rate resulting when the posterior probability error rate estimator is used, was found to be lower than that resulting from use of the apparent error rate and the leave-one-out error rate. The weaker performance of the latter two criteria may be due to their use of a 0-1 loss function. The expected actual error rates resulting when R^2 is used as selection criterion, is in close agreement with those resulting when using the probability error rate estimator as selection criterion. Since selection using R^2 (or other equivalent separatory criteria) is easier to implement, the use of a criterion such as R^2 can be recommended when the aim is merely to identify an optimal subset of a given size. However, the use of separatory criteria can not in general be recommended to choose the final model dimension.
3. If the aim in forming the linear discriminant function is accurate classification of future cases, it seems sensible to base a decision regarding the number of variables that should be included in the linear discriminant function on an allocatory criterion. This idea will be developed fully in Chapter 4, where a new selection technique will be proposed and evaluated. This technique will comprise of two steps: firstly, a separatory criterion is used to identify optimal models of each possible dimension and secondly, the final model dimension is chosen by using an allocatory criterion.

CHAPTER 4

VARIABLE SELECTION AND ERROR RATE ESTIMATION IN DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION BY MEANS OF CROSS MODEL VALIDATION

4.1 INTRODUCTION

In Chapter 3, a preliminary investigation into various aspects regarding variable selection in discriminant analysis was reported. The following conclusions emanated from this investigation: the candidate variables should not be considered singly, since this may give a false impression regarding their discriminatory power when combined with other variables; use of a separatory criterion is acceptable when an optimal model of a pre-specified dimension has to be identified, but the choice of an optimal model dimension should be based on an allocatory criterion, especially if the classification performance of the rule being constructed is of primary interest. In this chapter, a selection technique that takes these considerations into account, is proposed. This technique is based on a procedure called *cross model validation* that was developed by Hjorth (1994) for selection of variables in regression analysis. After appropriate modification, this technique can be used for variable selection in discriminant analysis, as well as in logistic regression. This is one of the topics discussed in this chapter.

An important aspect that also needs to be addressed, is estimation of the error rate of a classification rule based on a selected subset of the available variables. This is a particular example of the more general and difficult problem of assessing the accuracy of a procedure using the same data that were employed in constructing the procedure. In Chapter 2, estimation of the actual error rate of a discriminant rule in a situation where variable selection did not take place, was discussed, and an overview of error rate estimators was given. As mentioned there, many of these estimators are biased and/or have large variances. In a situation where variable selection precedes the formation of the discriminant rule, additional bias is introduced by the selection step, and the variance of the estimators is inflated. A need therefore exists for the development of error rate estimators that can be used in a post-selection context. One of the attractive features of the cross model validation procedure is that application of this technique to identify a model, also yields an estimate of the accuracy of this model. The cross model validation technique therefore simultaneously addresses two important aspects of the selection problem: firstly, selecting a subset of the available feature variables to construct a classification rule and secondly, estimating the associated post-selection error rates accurately. This is in line with opinions expressed by Breiman (1992) and by Venter and Steel (1994) in a regression context.

At this stage it is useful to use the notation introduced in Chapter 3 to describe the quantities that will be investigated in this chapter. In a discriminant analysis context, the properties of classification statistics $W_{p(t)}(\mathbf{X}; \mathbf{t}(J(t)))$ will be studied. Here, both the model dimension $p(t)$ and the subset $J(t)$ of the indices $1, \dots, k$ corresponding to the selected variables, are determined from the training data \mathbf{t} . Various methods from the literature that can be used to find $p(t)$ and $J(t)$ will be compared to the proposed cross model validation method. This comparison will take place in terms of the expected actual error rates of the various rules, given by (3.6.3). These error rates give an indication of the classification performance of the different rules. In practice, these quantities are unknown and have to be estimated from the training data. The proposed cross model validation estimator of post-selection error rate will be compared to other estimators from the literature. This comparison will take place in terms of bias and unconditional mean squared error (UMSE), defined as follows. Let $\hat{\alpha} = \hat{\alpha}(p(t); \mathbf{t}(J(t)))$ denote an arbitrary post-selection error rate estimator of α_{act} as defined in (3.6.3). Then the (expected) bias of $\hat{\alpha}$ is defined by

$$B(\hat{\alpha}) = E[\hat{\alpha} - \alpha_{act}] \quad (4.1.1)$$

and the UMSE of $\hat{\alpha}$ by

$$U(\hat{\alpha}) = E[\hat{\alpha} - \alpha_{act}]^2 \quad (4.1.2)$$

where the expectation in these expressions are taken with respect to the training data.

In a logistic regression context, classification statistics

$$V_{p(t)}(\mathbf{X}; \mathbf{t}(J(t))) = \beta_0(\mathbf{t}(J(t))) + \beta_1'(\mathbf{t}(J(t)))\mathbf{X}$$

are considered. Once more, both the model dimension $p(t)$ and the subset $J(t)$ of the indices $1, \dots, k$ corresponding to the selected variables, are determined from the training data \mathbf{t} . The cross model validation method will be used to determine $p(t)$ and $J(t)$, and to estimate the post-selection actual error rate of the resulting logistic classification function. The performance of the cross model validation procedure will be compared to another procedure in the literature in terms of the criteria defined in (4.1.1) and (4.1.2).

In Section 4.2, an overview of the literature on post-selection error rate estimation is given. This is followed in Section 4.3 by an explanation of the general principles underlying the cross model validation technique, with specific reference to its application in multiple linear regression. In Section 4.4, a proposal regarding application of the cross model validation technique in linear discriminant analysis, is put forward. Special emphasis is given to the modifications to the technique required for its use in this context. A detailed Monte Carlo study, in which the performance of the proposed cross model validation technique is compared to existing procedures in

the literature, is discussed in Section 4.5. In Section 4.6, application of the cross model validation technique in logistic regression receives attention. The results of the simulation study undertaken to evaluate the performance of the proposal made in this regard, are reported in Section 4.7. Section 4.8 contains a comparison of the selection and classification performance of the cross model validation technique in discriminant analysis to that of the cross model validation technique in logistic regression. In Section 4.9, the proposed new techniques are applied to two example data sets.

4.2 OVERVIEW OF LITERATURE ON POST-SELECTION ERROR RATE ESTIMATION

Murray (1977) warned against the use of the observed apparent error rate of the discriminant rule based on a selected subset of variables as an estimator of the error rate for classification of new cases. As mentioned in Chapter 2, the apparent error rate has a severely optimistic bias, and since the selected variables will be those that perform best in terms of the training data, the optimism of the apparent error rate is increased even further by the selection process. The performance of the rule on new data, for which the same variables will not necessarily be optimal, will typically be much worse than suggested by the apparent error rate. Rencher and Larson (1980) examined the bias in stepwise selection procedures based on Wilks' Λ . They argued that in cases where none of the available variables are good discriminators, this bias may lead to selection of 'an entirely spurious subset' with artificially high correct classification rates. Ganeshanandam and Krzanowski (1989) also commented on the 'double helping of overoptimistic bias' in the custom of assessing the classification performance of a rule based on a selected subset by means of the apparent error rate. To reduce the bias of the error rate estimator, they suggested a leave-one-out approach, repeating the selection process (using an error rate estimator as selection criterion, cf. Section 3.2 where this is described in more detail) for each omitted case. The proportion of 'holdout' cases that are misclassified is then used to estimate the post-selection error rate. In a Monte Carlo study, they compared the performance of their proposal to that of two other error rate estimators, viz. the parametric estimator proposed by Lachenbruch (1968) and the leave-one-out error rate. Both these estimators were calculated following variable selection using error rate as criterion. They found both these estimators to have severe optimistic bias, while their proposed estimator had much lower bias. As mentioned in Chapter 3, they did not address the problem of choosing an optimal model dimension, but restricted their investigation to a pre-specified number of variables. Since Murray (1977) argued that the optimistic bias of post-selection error rate estimators is largest at around $p = \frac{1}{2}k$, Ganeshanandam and Krzanowski (1989) only studied cases where the selection rules were required to select five out of ten available feature variables.

Snapinn and Knoke (1989) also stated that 'error rate estimators that perform well in ordinary discriminant analysis may not perform well with variables selected by a preliminary analysis'. They compared the performance of various error rate estimators of the post-selection error rate, following variable selection by means of F-based

forward stepwise selection. They considered the NS - estimator and the NS*-estimator defined in Chapter 2, each being calculated in two different ways. They used the smoothed estimator defined by (2.2.17) with the smoothing constants defined by (2.2.19) and (2.2.20), giving the NS_k - and NS_k^* - estimators respectively (referred to as NS_p and NS_p^* in their paper, since they used the symbol p to indicate the total number of available variables). By replacing k , the total number of feature variables, in (2.2.19) and (2.2.20) with p , the number of variables that were selected (denoted in their paper by q), the NS_p - and NS_p^* - estimators (denoted in their paper by NS_q and NS_q^* respectively) were obtained. They also included the leave-one-out estimator, the bootstrap bias corrected apparent error rate and the bootstrap bias corrected NS-estimator, which were all defined in Chapter 2, in their study. These estimators were also calculated in two different ways, referred to as partial and full resampling respectively. For partial resampling, variable selection is applied only once to a given training data set, and the three error rate estimators are then calculated as described in Section 2.2, using only the selected variables. In the case of full resampling, a new set of variables is selected for each omitted case (for the leave-one-out estimator) or for each bootstrap replication (for the two bootstrap estimators). In a Monte Carlo simulation study the performance of these estimators was evaluated for a number of different distributions (the normal distribution, the exponential distribution and the double exponential distribution) and parameter configurations. The assessment was done by comparing the expected bias and unconditional mean squared errors of the estimators when estimating the actual error rate. They concluded that the NS_k^* - estimator performed best in the case of normal distributions, but mentioned that this estimator is not robust, its performance being influenced by skewness of the parent distribution (as in the case of the exponential distribution).

Rutter et al. (1991) performed a study similar to that done by Snapinn and Knoke (1989). They included in their study the resubstitution (apparent) error rate, two versions of a plug-in error rate estimator, the bias corrected plug-in estimator suggested by McLachlan (1980a), the NS_k^* - and NS_p^* - estimators of Snapinn and Knoke (1989) as well as a 'holdout' estimator calculated by holding out a percentage (20% and 40% were used) of the data, performing stepwise selection on the remaining data, and classifying the 'holdout' cases. They recommended using the 'holdout' estimator, based on its very small bias in estimating the actual error rate. However, they did not consider the variance of the estimators. As will be shown later, the holdout estimator has a large variance, resulting in its unconditional mean squared error being much larger than for example that of the NS_k^* - estimator.

Rencher (1992) carried out an extensive Monte Carlo simulation study to investigate the bias of the apparent error rate of a discriminant rule based on a subset of variables selected by means of forward stepwise selection. He considered the null case of no difference between the groups, having an expected error rate of $G/(G+1)$, where $G+1$ is the number of groups. He calculated the apparent error rate of the rule based on the variables selected by means of forward selection, and also the apparent error

rate of a rule based on a randomly selected subset of the same size. The difference between these two error rates is considered to be the bias due to the stepwise selection, while the bias due to the resubstitution is obtained by calculating the difference between the expected error rate ($G/(G+1)$) and the apparent error rate of the rule based on the randomly selected variables. A large number of configurations were obtained by varying the number of groups (2, 4, 6 and 8), the number of potential variables before selection (10, 20, 30 and 40) and the sample size per group (5 and 10). The case where all variables were uncorrelated, was studied, as well as the correlated case with different values of the index of correlation between the variables (defined as $\left(\sum_{i=1}^k 1/\lambda_i\right)/k$, where $\lambda_1, \dots, \lambda_k$ are the eigenvalues of the correlation matrix). The values 1, 10, 100 and 1000 were used for this index. For each of the configurations, four different threshold F-values for the forward selection were used. Based on analyses of two data sets containing large numbers of variables and relatively small sample sizes, Rencher expected the bias to be largest in these types of situations. He therefore deliberately included many configurations where the number of variables exceeded the degrees of freedom for error, to obtain an indication of the extent of the bias under these circumstances. He found that the bias due to the resubstitution varied between 0.06 and 0.77, and the selection induced bias varied between 0.01 and 0.23. It must be noted that the selection bias of 0.01 was obtained in a situation where the apparent error rate of the rule based on a randomly selected subset was 0.01, while the apparent error rate of the rule based on the variables selected by means of forward selection was 0. In general, the total bias increased with a decrease in the ratio of cases to variables, approaching $G/(G+1)$ (the maximum possible total bias for an expected error rate of $G/(G+1)$) in cases where the number of variables was very large (40) and the sample sizes small (5). The total bias also increased with decreasing threshold F-value, and with decreasing correlation between the variables.

The papers discussed above all considered estimation of the error rate of the linear discriminant rule based on a selected subset. In a logistic regression context, Efron and Gong (1983) and Gong (1986) investigated the estimation of excess error, defined as the difference between the true error rate and the apparent error rate of a logistic discriminant rule based on a subset obtained by means of forward selection. Efron and Gong (1983) suggested the following bootstrap procedure to estimate the excess error. For each bootstrap sample generated from the training data, the variable selection process is repeated, and the logistic classification function based on the selected variables is used to classify the entities in the bootstrap sample as well as the entities in the original training data set. The difference between the error rates obtained when classifying the original training data and the bootstrap sample, is calculated. These differences are averaged over all bootstrap replications, and this is used as estimator of the excess error. The excess error can be used to correct the apparent error rate for bias.

Gong (1986) compared the performance of the excess error estimator described above to that of estimators obtained by means of cross validation and the jackknife. The

results of her Monte Carlo simulations indicated that although the cross validation and jackknife estimators are nearly unbiased, they do not perform much better than the apparent error rate in terms of mean squared error. The bootstrap estimator has a small optimistic bias, but shows a considerable improvement on the apparent error rate in terms of mean squared error. This estimator is therefore recommended for estimation of excess error and to correct the apparent error rate for bias.

4.3 CROSS MODEL VALIDATION

4.3.1 GENERAL PRINCIPLES

In this section the general principles underlying the cross model validation (CMV) approach are discussed. This can best be done by contrasting the cross model validation approach with the ordinary cross validation (CV) approach in a general variable selection context, highlighting the important differences between the two approaches.

Consider k variables X_1, \dots, X_k , and suppose n independent measurements are available on each of these variables. Denote the complete data set by \mathbf{X} , an $n \times k$ matrix, and let $\mathbf{X}_{(j)}$ denote the data with the j -th observation (row) deleted. Let $\mathcal{K} = \{1, \dots, k\}$. The problem is to select a subset of variables $J \subset \mathcal{K}$ such that the variables with indices in J define a model that is optimal in some sense. To be more specific, let $M_p(J)$ denote the model defined by the variables with indices in J , where $\#(J) = p$. Also, let $H(\mathbf{X}; M_p(J))$ denote a data-dependent criterion of the inaccuracy of the model, that has to be minimised with respect to model dimension p and model $M_p(J)$. Denote the optimising model by $\tilde{M}_{\tilde{p}(\mathbf{X})}(\tilde{J}(\mathbf{X}))$, i.e.

$$H(\mathbf{X}; \tilde{M}_{\tilde{p}(\mathbf{X})}(\tilde{J}(\mathbf{X}))) = \min\{H(\mathbf{X}; M_p(J)), p \in \mathcal{K}; J \subset \mathcal{K}\}.$$

When model selection is done by means of cross validation, all possible models of each dimension $p = 1, \dots, k$ are considered. For each of these $2^k - 1$ models, a measure of prediction error is obtained by means of cross validation. To calculate this measure, each of the n cases is omitted in turn, and the model is fitted to the remaining $n - 1$ cases. This model is then used to predict the omitted case, and some measure of loss associated with this prediction is obtained. The cross validation criterion for each $p \in \mathcal{K}$ and $J \subset \mathcal{K}$ is obtained by averaging the loss for all omitted cases, i.e.

$$H^{\text{CV}}(\mathbf{X}; M_p(J)) = \frac{1}{n} \sum_{i=1}^n H(\mathbf{X}_{(i)}; M_p(J)).$$

The optimal model is identified by minimising the cross validation criterion over all possible models, i.e.

$$\tilde{H}^{CV} = \frac{1}{n} \sum_{i=1}^n H(\mathbf{X}_{(i)}; \tilde{M}_{\tilde{p}(\mathbf{X})}^{CV}(\tilde{\mathcal{J}}(\mathbf{X}))) = \min \left\{ \frac{1}{n} \sum_{i=1}^n H(\mathbf{X}_{(i)}; M_p(\mathcal{J})), p \in \mathcal{K}; \mathcal{J} \subset \mathcal{K} \right\}.$$

The model $\tilde{M}_{\tilde{p}(\mathbf{X})}^{CV}(\tilde{\mathcal{J}}(\mathbf{X}))$ yielding this minimum is chosen as the optimal model, and the minimum value of the criterion is used to estimate the prediction error of this model. However, as argued by Hjorth (1994, p. 34-37), the cross validation estimator of prediction error is optimistically biased. Hjorth stated that 'the very selection of such a model (to minimise a measure of loss) introduces bias error in the measure'.

According to Hjorth (1994), cross validation can be performed in such a way that model selection effects are measured, and a less biased estimator of the prediction error is obtained. To achieve this, it is important that a fixed model should not be used to predict each omitted case (as is done in the cross validation procedure described above) but that selection should be repeated at each case being omitted, so that potentially different models of dimension p could be considered as the different cases are omitted. When this is done, model selection effects can be measured, since selection errors come into play during the leave-one-out process.

Hjorth developed a procedure, called cross model validation (CMV) along these lines. To calculate the cross model validation variable selection criterion, each of the n data cases is once more omitted in turn. For each omitted case, a so-called *inner criterion* is applied to the remaining $n-1$ data cases to identify an optimal model of each possible dimension, $p = 1, \dots, k$. Denote these models by $M_p(\mathcal{J}(\mathbf{X}_{(i)}))$, for $p = 1, \dots, k; i = 1, \dots, n$. It is important to note that for each fixed value of p , the models $M_p(\mathcal{J}(\mathbf{X}_{(i)}))$ can differ for each value of i . Each of the models $M_p(\mathcal{J}(\mathbf{X}_{(i)}))$ is used to predict the omitted case, and some measure of loss associated with this prediction is obtained. The CMV criterion for each $p \in \mathcal{K}$ is calculated by averaging these losses over all the omitted cases, i.e.

$$H^{CMV}(\mathbf{X}; p) = \frac{1}{n} \sum_{i=1}^n H(\mathbf{X}_{(i)}; M_p(\mathcal{J}(\mathbf{X}_{(i)}))), \quad p = 1, \dots, k.$$

An optimal model dimension $\tilde{p}(\mathbf{X})$ is identified by minimising this criterion over p , i.e.

$$H^{CMV}(\mathbf{X}; \tilde{p}(\mathbf{X})) = \min \{ H^{CMV}(\mathbf{X}; p); p = 1, \dots, k \}.$$

To complete the variable selection process, the inner criterion is once more applied to all n data cases, but only models of dimension $\tilde{p}(\mathbf{X})$ are considered. In this way a

final subset $\tilde{\mathcal{J}}(\mathbf{X})$ containing $\tilde{p}(\mathbf{X})$ indices, is identified. The minimum value of $H^{\text{CMV}}(\mathbf{X}; p)$, i.e. $H^{\text{CMV}}(\mathbf{X}; \tilde{p}(\mathbf{X}))$, is used as an estimate of the prediction error of the finally selected model, $\tilde{M}_{\tilde{p}(\mathbf{X})}^{\text{CMV}}(\tilde{\mathcal{J}}(\mathbf{X}))$. Hjorth (1994) claims that $H^{\text{CMV}}(\mathbf{X}; \tilde{p}(\mathbf{X}))$ is less biased than $H^{\text{CV}}(\mathbf{X}; \tilde{p}(\mathbf{X}))$ as an estimator of the prediction error of the finally selected model.

In cross validation therefore, a measure of inaccuracy is calculated for each of the $2^k - 1$ possible models. A single model is selected by minimising this measure over all $2^k - 1$ candidate models, and the minimum value thus obtained is also used to estimate the prediction error of the selected model. In cross model validation however, only the k possible model dimensions are in effect considered, and a measure of inaccuracy is calculated for each value of $p = 1, \dots, k$. The selected model dimension $\tilde{p}(\mathbf{X})$ minimises this criterion, and this minimum value is used to estimate the prediction error of the $\tilde{p}(\mathbf{X})$ -dimensional model selected by application of the inner criterion to the full data set.

4.3.2 CROSS MODEL VALIDATION IN A REGRESSION CONTEXT

An important application of cross model validation occurs when variable selection has to be done in the well known multiple regression set-up. The general description given in the previous section, specialises as follows. Let \mathbf{X} be the $n \times k$ matrix of observations on the covariates and let $\mathbf{X}_{(j)}$ denote the data with the j -th observation (row) deleted. Denote the n -dimensional vector of observations on the response variable by \mathbf{y} and let $\mathbf{y}_{(j)}$ denote the response vector with observation j deleted.

For each j ($j = 1, \dots, n$) the best regression model of $\mathbf{y}_{(j)}$ on $\mathbf{X}_{(j)}$ is selected for each model size p ($p = 1, \dots, k$). To achieve this, the inner criterion is compared for a set of candidate models of the same size p , and the 'best' model for size p is selected. As mentioned before, it is important to note that different models of a given size p may be selected for each different j . Measures that can be used as inner criterion include the residual sum of squares, the multiple correlation coefficient, the average predicted loss or even the cross validation estimator of prediction error. If there is a small number of potential variables, all possible subsets of a given size can be considered at each step, but if the number of candidate variables is large, the selection for each specified model size can be done in a stepwise manner, such as forward selection or backward elimination.

Denote the best model of size p when observation j is excluded by

$$M_p(\mathbf{X}_{(j)}, \mathbf{y}_{(j)})$$

and the prediction based on this model by

$$\hat{y}_j(p) = \hat{y}(\mathbf{x}_j, M_p(\mathbf{X}_{(j)}, \mathbf{y}_{(j)})).$$

Define the cross model validation criterion for model size p as

$$\text{CMV}(p) = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j(p))^2 \quad (4.3.2.1)$$

or more generally

$$\text{CMV}(p) = \frac{1}{n} \sum_{j=1}^n L(\hat{y}_j(p), y_j) \quad (4.3.2.2)$$

where $L(\hat{y}_j(p), y_j)$ is an appropriate loss function. The optimal model size p_0 is chosen to minimise $\text{CMV}(p)$, i.e.

$$\text{CMV}(p_0) = \min\{\text{CMV}(p) : p = 1, \dots, k\}. \quad (4.3.2.3)$$

In a final step only models of size p_0 are considered. Using all the data, the 'best' model of this dimension is identified according to the inner selection criterion, either by considering all possible subsets or using a stepwise procedure.

Hjorth (1994, p. 30-45) compared cross validation and cross model validation by applying both techniques to the well known data set of Hald (1952). He used an all possible subsets approach and firstly identified a best model of each possible dimension 2, ..., 5 (all models also included an intercept) by minimising the CV criterion over the sets of models of different dimensions. The minimum CV value for dimensions 2, ..., 5 are estimates of the prediction error of the corresponding optimal model. He then repeated this process, applying cross model validation as described above, once more finding a best model of each dimension and estimates of the prediction error of these models. For each model size, the CMV-based estimate is larger than the CV-based estimate, except for the model including all the variables, in which case the two estimates are equal. The difference between the two sets of estimates can be ascribed to the repeated model selection being done in the CMV-procedure.

4.4 CROSS MODEL VALIDATION IN DISCRIMINANT ANALYSIS

The cross model validation method described in the previous section, can also be applied to the problem of variable selection in discriminant analysis. However, in order to do this, the procedure as described by Hjorth in a regression context, has to be modified considerably. In this section the case of two groups is considered, i.e. $G = 1$.

The first important aspect that has to receive attention, is the choice of an inner criterion to select the best model of each possible size p if case j is deleted from the training data set ($j = 1, \dots, n$; $p = 1, \dots, k$). Different inner criteria can be considered for this purpose. Possibilities that come to mind are forward selection, backward elimination, fully stepwise selection and an all possible subsets approach. At this stage it should be pointed out that a special form of forward selection (or backward elimination) has to be used if it is implemented as inner criterion in cross model validation. The reason for this is that a best model of *each* possible size $p = 1, \dots, k$ is required from the inner criterion. Ordinarily, if forward selection is applied in for example regression analysis, the practitioner specifies a so-called F-to-enter value. Then at any stage of the selection process only the variables that have not yet been included in the model and that have F-test values exceeding the F-to-enter value, are candidates for inclusion at this particular stage. If none of the variables that have not yet been selected pass the F-to-enter criterion, selection terminates. Hence, by specifying an F-to-enter value, the practitioner is by implication also determining the size of the final selected model. The only way to ensure that a model of every possible size is identified by means of forward selection, is to use an F-to-enter value equal to zero. This point is also emphasised by Hjorth (1994, p. 41) when he states: "*We think of the basic forward selection, without testing for inclusion or deletion of variables*". In this connection it should be borne in mind that the later cross model validation step is used to decide on the dimension of the final model.

The above remarks are equally valid if backward elimination is considered as inner criterion. To ensure that a best model of every possible model size is identified, an F-to-leave value that is very large has to be specified. A problem arises when considering a fully stepwise approach as inner criterion. Now both F-to-enter and F-to-leave values have to be specified, and the arguments above would suggest F-to-enter = 0 and F-to-leave = ∞ . But such a specification is not suitable for a fully stepwise procedure, since any variable that is included at a given stage will automatically also qualify for deletion at a later stage, causing the procedure to continue indefinitely. Hence, as far as the stepwise procedures are concerned, it seems that only the special forms of forward selection and backward elimination described above are suitable as inner criteria.

It is a well known fact that application of forward selection (or backward elimination) does not guarantee that the best model of any given size p will be selected, since only a relatively small number of the potential models of size p are actually considered. A solution to this problem would be to use an all possible subsets approach as inner criterion. Although this is computationally more expensive than forward selection or

backward elimination, especially in cases where there is a large number of feature variables, the growing availability of powerful computers reduces the importance of this aspect. It should also be remembered that a practitioner typically applies such a procedure to a single data set.

To investigate the performance of cross model validation as variable selection technique in discriminant analysis, an extensive Monte Carlo simulation study was undertaken. In the first part of this simulation study, the performance of cross model validation selection and error rate estimation is compared to the proposals of Rutter et al. (1991) and Snapinn and Knoke (1989), which were discussed in Section 4.2. Since F-based forward selection was used in both these papers, it was decided to use F-based forward selection as inner criterion in the cross-model validation procedure. Any observed differences in the performance of cross model validation and the other two procedures can therefore be ascribed to the effect of the cross model validation step. In the second part of the Monte Carlo study, an all possible subsets approach based on R^2 , was used as inner criterion, to investigate the effect of using this approach instead of a forward selection approach.

The cross model validation procedure used in the first study, is now described. Consider $n = n_0 + n_1$ observations on k variables, of which a subset has to be selected for inclusion in a discriminant function. Denote the $n \times k$ data matrix by X , and the data matrix with the j -th observation (row) deleted, by $X_{(j)}$. In the two-group discriminant analysis context the n -dimensional response vector y will contain observations y_j indicating group membership, viz.

$$y_j = \begin{cases} 0 & \text{for an observation from } \Pi_0 \\ 1 & \text{for an observation from } \Pi_1 \end{cases}$$

Let $y_{(j)}$ denote the response vector with observation j deleted from the training data set.

Using F-based forward stepwise selection as inner criterion when case j is deleted, entails the following. Firstly, the single variable that discriminates best between the two populations (in terms of F-values) is identified. To find the best two-dimensional model, only models that contain the best single variable identified at the previous step, with one of the previously omitted variables added, are considered. The variable that, in combination with the variable that has already been entered, yields the largest F-value, is included in the model. This procedure is repeated for $p = 3, \dots, k$, where at any stage the variables already selected at the previous stage are retained, and only the best remaining variable is added. Denote this model for each j and p by $M_p(X_{(j)}, y_{(j)})$ and denote the prediction based on this model by

$$\hat{y}_j(p) = \hat{y}(x_j, M_p(X_{(j)}, y_{(j)})).$$

The value of $\hat{y}_j(p)$ will be the predicted group membership of the deleted observation \mathbf{x}_j , i.e.

$$\hat{y}_j(p) = \begin{cases} 0 & \text{if } W(\mathbf{x}_j) > 0 \\ 1 & \text{if } W(\mathbf{x}_j) \leq 0, \end{cases}$$

where $W(\cdot)$ is the Anderson classification statistic defined in (2.1.7), based only on the p variables selected at this stage, $p = 1, \dots, k$. The squared error loss function

$$L(\hat{y}_j(p), y_j) = [y_j - \hat{y}_j(p)]^2$$

is in this context equivalent to

$$L(\hat{y}_j(p), y_j) = \begin{cases} 0 & \text{for correct classification} \\ 1 & \text{for misclassification.} \end{cases}$$

If this dichotomous loss function is used, not all the information contained in the value of $W(\mathbf{x})$ is utilised (see Habbema and Hermans (1977)). Another disadvantage in the present context is that it can quite easily happen that some of the $CMV(p)$ values are equal, especially in small sample cases. In these cases, a unique p_0 can not be identified. To avoid this difficulty, a normally smoothed version of this loss function, similar to the function defined by Snapinn and Knoke (1985), is proposed :

$$\tilde{L}(p, j) = \begin{cases} \Phi[-W(\mathbf{x}_j)/(b_1 D)] & \text{if } \mathbf{x}_j \in \Pi_0 \\ \Phi[W(\mathbf{x}_j)/(b_2 D)] & \text{if } \mathbf{x}_j \in \Pi_1. \end{cases} \quad (4.4.1)$$

In this definition b_1 and b_2 are smoothing constants given by

$$b_1 = \left\{ \frac{[(p+2)(n_0-2) + n_1 - 1]}{[(n_0-1)(n_0+n_1-p-4)]} \right\}^{1/2},$$

$$b_2 = \left\{ \frac{[(p+2)(n_1-2) + n_0 - 1]}{[(n_1-1)(n_0+n_1-p-4)]} \right\}^{1/2}.$$

The cross model validation criterion for model size p is then defined as

$$CMV(p) = \frac{1}{n} \sum_{j=1}^n \tilde{L}(p, j). \quad (4.4.2)$$

In a preliminary simulation study it was found that Hjorth's suggestion of choosing p_0 to minimise $CMV(p)$ as in (4.3.2.3), often lead to overfitting in the sense that seemingly irrelevant variables were included in the discriminant function. This was caused by the fact that $CMV(p)$ often tended to decrease very slightly with the addition of seemingly irrelevant variables to the model. In an attempt to address this problem, the following procedure that takes the *magnitude* of the reduction in the criterion with increasing model size into account, is proposed:

Consider the successive values of $CMV(p)$, $p = 1, \dots, k$.

Define an initial value, $CMV^* = CMV(1)$.

For $p = 2, \dots, k$, perform the following steps:

Calculate the difference $d_p = CMV^* - CMV(p)$.

If $d_p \geq \phi CMV^*$, then $CMV^* = CMV(p)$.

The final value of CMV^* is used as the cross model validation based error rate estimator, and the dimension p_0 for which $CMV(p) = CMV^*$ is taken as the estimated optimal model size.

This procedure implies that a more complex model will be selected only if such a model yields a fairly considerable reduction in CMV . The parameter ϕ ($0 < \phi < 1$) can be used to control the amount of reduction in CMV required before such a more complex model is preferred. Using a small value of ϕ favours selection of a more complex model and vice versa. After experimenting with a number of different ϕ -values, it became evident that no value exists that is ideal for all data configurations. The criteria (such as UMSE) used to evaluate the proposed method were however fairly robust with respect to changes in ϕ in the neighbourhood of 0.025. Therefore this compromise value was used.

Another strategy that may be employed in practice, is to plot the values of $CMV(p)$ against p , and to use this graph as an aid in finding the final model dimension. This is similar to the use of a scree plot in determining the number of factors in a factor analysis (cf. Cattell, 1966). The effect of using this plot is similar to what is achieved by using ϕ , as described in the previous paragraph. This type of plot can be used when applying the cross model validation technique to a data set (see Section 4.9), but is not feasible in a simulation study. The strategy involving ϕ is therefore used in the simulation study described in Section 4.5.1. In the practical examples discussed in Section 4.9, the use of a plot of $CMV(p)$ against p , will be illustrated.

4.5 MONTE CARLO SIMULATION STUDY FOR DISCRIMINANT ANALYSIS

An extensive Monte Carlo simulation study was undertaken to compare the performance of the cross model validation technique to that of the procedures proposed by Snapinn and Knoke (1989) and Rutter et al. (1991), both described in Section 4.2. The behaviour of the three methods was evaluated for populations with different underlying distributions: the normal distribution, the double exponential distribution and the lognormal distribution. In each of these cases, three different sample sizes were considered: $n_0 = n_1 = 25$ (small samples), $n_0 = 75$; $n_1 = 25$ (mixed samples) and $n_0 = n_1 = 100$ (large samples). The following coding will be used to denote the different cases: the codes NS, NM and NL will be used to denote the small sample, mixed sample and large sample normal cases respectively, with DS, DM and DL being used similarly for the double exponential cases, and LS, LM and LL for the lognormal cases. Regarding the covariance structure, $\Sigma = I$ was used for all the distributions. In the normal case, Σ given by (2.4.1) with $\rho = 0.9$ was also used. The value $k = 10$ was used throughout. It is assumed that the feature vector \mathbf{X} has mean vector $\mu_0 = \mathbf{0}$ in Π_0 , and that the first r elements of μ_1 , the mean vector of \mathbf{X} in Π_1 , differ from zero. The values $r = 1, 5$ and 10 were used. For $r = 1$ and 5 , the elements of μ_1 were chosen as

$$\mu_{1l} = \begin{cases} \Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r \sigma^{ij}} & , \quad l = 1, \dots, r \\ 0 & , \quad l = r + 1, \dots, 10. \end{cases} \quad (4.5.1)$$

For $r = 10$, two different choices for the elements of μ_1 were considered. Firstly the case where all the elements of μ_1 are equal, viz.

$$\mu_{1l} = \Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r \sigma^{ij}} , \quad l = 1, \dots, 10, \quad (4.5.2)$$

was considered. A second choice in which the components of μ_1 are equi-spaced, was also considered, viz.

$$\mu_{1l} = l\Delta / \sqrt{\sum_{i=1}^r \sum_{j=1}^r ij\sigma^{ij}} , \quad l = 1, \dots, 10. \quad (4.5.3)$$

For each of these cases, the performance of the three methods was studied at the following values of Δ^2 : 0, 1, 2, 3, 4, 6 and 9.

The procedures included in this study were evaluated in terms of a number of aspects of their performance. Two main aspects were considered, viz. the selection performance and the accuracy of estimation of the resulting actual error rates. The post-selection expected actual error rates of the techniques were compared as measure of allocatory performance. The separatory performance of the techniques was investigated in terms of the probability of correct selection (PCS), i.e. the probability of including all the seemingly relevant variables and no seemingly irrelevant variables. To evaluate the accuracy of estimation of the resulting post-selection actual error rates, the bias and the unconditional mean squared error (UMSE) of each of the three estimators were compared.

All of the above quantities were estimated by means of simulation using 500 repetitions. Cases where a selection procedure did not select any variables, were excluded from further analyses. Additional simulation repetitions were then performed until 500 cases were obtained where each of the procedures selected one or more variables. An example of the Fortran program that was used in this simulation study, appears as Program 2 in the Appendix.

4.5.1 INNER CRITERION : FORWARD STEPWISE SELECTION

4.5.1.1 THE NORMAL CASE

In the normal case, a Monte Carlo simulation study was done to compare the selection and estimation performance of the three procedures in terms of the criteria defined above. To estimate the required quantities, 500 Monte Carlo repetitions were used at each value of Δ^2 . For each repetition, a training data set was generated from the two relevant normal distributions. For the procedures proposed by Snapinn and Knoke (1989) and Rutter et al. (1991), F-based forward selection with α -to-enter = 0.15, was performed on the training data, and variable selection by means of the cross model validation procedure was also done. Since the same selection strategy is used for the procedures of Snapinn and Knoke (1989) and Rutter et al. (1991), the same subset will of course be selected by these procedures. All aspects of the selection performance of these two procedures, viz. the post-selection actual error rates and the PCS, will therefore be identical. Different error rate estimators are however proposed by Snapinn and Knoke (1989) and Rutter et al. (1991), resulting in a difference in estimation performance.

For each of the selected subsets, the post-selection actual error rate was calculated using (2.2.9). To calculate the post-selection actual error rate associated with a specific selection technique, the quantities in (2.2.9) were calculated using only the variables selected by that technique. The three different post-selection error rate estimators, viz. the NS_t^* -estimator, the holdout estimator and the CMV-estimator, were also calculated. With a view to estimating the bias and unconditional mean

squared error of each of the error rate estimators, the difference and squared difference between the value of each error rate estimator and the post-selection actual error rate, were also calculated. To obtain the expected post-selection actual error rates, the 500 actual error rates obtained for each technique, were averaged. To estimate the probability of correct selection associated with each technique, the fraction of repetitions in which all the seemingly relevant variables and no seemingly irrelevant variables were selected, was calculated. The bias associated with each technique was estimated by averaging the differences between the value of each error rate estimator and post-selection actual error rate over the 500 repetitions, i.e.

$$\hat{B}_j = \frac{1}{500} \sum_{i=1}^{500} (\hat{\alpha}_{ij} - \alpha_{ij}^{\text{act}}), j = 1, 2, 3, \text{ where } \hat{\alpha}_{ij} \text{ denotes a value of an error rate estimator}$$

obtained by means of technique j , $j = 1, 2, 3$ for the i -th Monte Carlo repetition and α_{ij}^{act} denotes the actual error rate calculated for technique j for the i -th Monte Carlo repetition. To estimate the unconditional mean squared error of the j -th error rate estimator, the squared differences between the relevant error rate estimator and post-

selection actual error rate, were averaged, i.e.
$$\hat{U}_j = \frac{1}{500} \sum_{i=1}^{500} (\hat{\alpha}_{ij} - \alpha_{ij}^{\text{act}})^2.$$

The results of the simulation study were summarised by means of graphs. A representative selection of these graphs is given in Figs. 4.1 - 4.7. In Figs. 4.1 - 4.2, graphs of the post-selection expected actual error rates are given, while Fig. 4.3 displays the PCS associated with the procedures. Figs. 4.4 - 4.5 contain graphs of the bias of the three error rate estimators, and graphs of the unconditional mean squared errors of the error rate estimators are given in Figs. 4.6 - 4.7.

The factors mentioned at the beginning of Section 4.5, identify a total of 24 different normal cases. In the small sample cases, the coding NS11, NS21, NS31 and NS41 is used to denote the cases where $\Sigma = I$ and $r = 1$, $r = 5$, $r = 10$ (with μ_{lt} , $l = 1, \dots, 10$ given by (4.5.2)) and $r = 10$ (with μ_{lt} , $l = 1, \dots, 10$ given by (4.5.3)), in that order. For the equi-correlated cases, the coding NS12, NS22, NS32 and NS42 is used similarly. For the mixed and large sample cases, similar coding with NM and NL instead of NS, is used.

SELECTION PERFORMANCE

The selection performance of the techniques is firstly evaluated. Two aspects are considered, viz. the post-selection expected actual error rate and the probability of correct selection associated with the techniques. Since the procedures of Snapinn and Knoke (1989) and Rutter et al. (1991) use the same selection strategy, the selection performance of these two methods is identical, and therefore indistinguishable on the graphs displaying the post-selection actual error rates and probabilities of correct selection. This section is therefore a comparison of F-based forward selection with α -to-enter = 0.15 and selection by means of cross model validation. As described in

Section 4.4, cross model validation is a two-stage procedure in which the optimal model size p_0 is firstly determined. The optimal subset containing p_0 variables, is then obtained. In this simulation study, this was done by means of F-based forward selection. When applying F-based forward selection in the usual way, the size of the selected subset is implicitly determined by specifying an α -to-enter value. Any difference in the selection performance of cross model validation and F-based forward selection with α -to-enter = 0.15, can therefore only be due to the fact that subsets of different sizes are selected.

Expected Actual Error Rate

In the case of normal data, the post-selection expected actual error rate of the cross model validation procedure is very slightly larger at some values of Δ^2 than that of the other procedures in cases NS12 and NS32 (see Fig. 4.2 where case NS32 is displayed). In cases NS31, NM31, NM41, NS22, NM22 and NM42 the cross model validation procedure is appreciably better, especially for large separation between the populations (see Fig. 4.1 for cases NM31 and NS31 and Fig. 4.2 for case NM22). In cases NS21, NS41, NM21, NL31, NS42 and NL22 the expected actual error rates associated with the cross model validation procedure are only slightly lower than that of the other procedures (see Fig. 4.1 for case NS41 and Fig. 4.2 for case NL22). In the remaining cases, the expected actual error rates are practically identical (see Fig 4.1 for case NL31 and Fig. 4.2 for case NM12). In general, the differences described above seem to be largest for the mixed sample case, and smallest for large samples. The relative performance of the selection strategies are not influenced by the introduction of correlation between the feature variables, although the error rates are generally higher in the presence of correlation. The cross model validation technique never performed appreciably worse in terms of post-selection actual error rate than F-based forward selection, and performed considerably better in a number of the cases considered. This is an indication that a classification function based on variables selected by means of cross model validation, will in general perform better in terms of accurate classification of future cases.

Probability of Correct Selection (PCS)

In the cases where the feature variables were independent (NS11 - NS41, NM11 - NM41 and NL11 - NL41), the cross model validation based selection procedure consistently outperformed the ordinary forward selection procedure with respect to the PCS. Especially in the cases where $r = 1$ (cases NS11, NM11 and NL11) cross model validation dominated, achieving PCS between 0.4 and 0.6, opposed to PCS of approximately 0.2 achieved by the other procedure (see Fig 4.3 for cases NS11 and NL11). In the cases where $r = 5$ (cases NS21, NM21 and NL21) cross model validation also achieved higher PCS than the other procedure, but the difference is not as large as in the cases mentioned above (see Fig. 4.3 for case NM21). In the cases

where $r = 10$ and the elements of μ_1 are given by (4.5.2) (cases NS31, NM31 and NL31), cross model validation yielded higher PCS than the other procedure, the difference between the two procedures increasing with Δ^2 (see Fig. 4.3 for case NL31). In the cases where $r = 10$ and the elements of μ_1 are given by (4.5.3) (cases NS41, NM41 and NL41), both procedures achieved very low PCS. For uncorrelated cases, variable selection using a cross model validation based procedure seems to outperform ordinary forward stepwise selection with respect to the probability of selecting the seemingly relevant variables.

In cases where the feature variables were correlated (NS12 - NS42, NM12 - NM42 and NL12 - NL42), all the procedures yielded very low PCS values. It should however be noted that the PCS is defined as the probability of selecting all the seemingly relevant variables, and no seemingly irrelevant variables. As discussed in Chapter 3, inclusion of one or more seemingly irrelevant variables that are highly correlated with the seemingly relevant variables in the classification function, increases the separation between the populations and leads to a reduction in the error rate. The fact that the procedures achieved very low PCS-values, is therefore not an indication of poor performance, but rather a reflection of the fact that the techniques often selected one or more seemingly irrelevant variables, due to the increase in separation or decrease in error rate resulting from inclusion of such variables.

ESTIMATION PERFORMANCE

To evaluate the estimation accuracy of the three procedures, the bias and unconditional mean squared errors (UMSE) of the error rate estimators are compared.

Bias

When the bias of the three error rate estimators is compared, it is clear that the holdout estimator proposed by Rutter et al. (1991), consistently outperforms the other two estimators. In large sample cases (NL11 - NL41 and NL12 - NL42) the holdout estimator is nearly unbiased (see Fig. 4.4 for cases NL11 and NL41 and Fig. 4.5 for cases NL12 and NL22). In the small sample cases (NS11 - NS41 and NS12 - NS42) the holdout estimator is slightly biased at small to moderate values of Δ^2 , but the bias decreases with increasing Δ^2 (see Fig. 4.4 for case NS11 and Fig. 4.5 for case NS42). The same holds for the mixed sample cases (NM11 - NM41 and NM12 - NM42) where the decrease in bias occurs at smaller values of Δ^2 than in the small sample cases (see Fig. 4.4 for case NM21 and Fig. 4.5 for case NM32).

The NS_k^* -estimator is generally more biased than the holdout estimator, outperforming it in some cases only at a few values of Δ^2 (see Fig. 4.5 for case NS42 where the NS_k^* -estimator is less biased than the holdout estimator at $\Delta^2 = 1$). The NS_k^* -estimator is also in most cases less biased than the CMV-estimator at small values of

Δ^2 ($\Delta^2 = 0,1$). At moderate values of Δ^2 ($\Delta^2 = 2,3$) the NS_k^* -estimator is less biased than the CMV-estimator only in a few cases (see Fig. 4.4 for case NM21 and Fig. 4.5 for cases NS42 and NL12), while the CMV-estimator has smaller bias at moderate separation in other cases (see Fig. 4.4 for cases NS11, NL11 and NL41 and Fig. 4.5 for cases NM32 and NL22). At large values of Δ^2 ($\Delta^2 > 3$) the CMV-estimator consistently outperforms the NS_k^* -estimator with respect to bias (see all cases in Figs. 4.4 and 4.5). The CMV-estimator also outperforms the holdout estimator at large values of Δ^2 in some cases (see Fig. 4.4 for cases NS11, NL11 and NL41 and Fig. 4.5 for case NL22).

In general, the holdout estimator performs best with respect to bias. Regarding the NS_k^* - and CMV-estimators, the NS_k^* - estimator performs better at small separations, while the CMV-estimator performs better at large separations.

Unconditional Mean Squared Error

When considering the graphs displaying the unconditional mean squared errors of the three error rate estimators (Figs. 4.6 and 4.7), it is clear that the holdout estimator performs very badly in terms of this criterion. Despite being nearly unbiased, the large variance of the holdout estimator causes its unconditional mean squared error to be much larger than that of the NS_k^* - estimator and the CMV-estimator, although these estimators were more biased. The large UMSE-values cast doubt over the suitability of the holdout estimator as post-selection error rate estimator.

An interesting point is revealed when perusing the graphs in Figs. 4.6 and 4.7, viz. the extremely small UMSE of the NS_k^* - estimator when there is no separation between the two groups. This is a result of the way in which the smoothing constant b is defined in (2.2.22): when $\Delta^2 = 0$, the choice $b = \infty$ is often made, resulting in the estimator being very close to 0.5, which is of course the correct value when $\Delta^2 = 0$. The fact that b in (2.2.22) is a discontinuous function of D^2 , explains the interesting behaviour of the UMSE of the NS_k^* - estimator when $\Delta^2 = 0$. For this reason the discussion concentrates on cases where $\Delta^2 \geq 1$.

For normal data the UMSE of the CMV-estimator is appreciably lower than that of the NS_k^* - estimator in cases NS11, NM11 (see Fig. 4.6 for these cases), NS32 and NM32 (see Fig. 4.7 for these cases). In cases NM31, NL31 and NM22 the UMSE of the CMV estimator is slightly higher than that of the NS_k^* procedure for $\Delta^2 = 1$, but the opposite is true when Δ^2 increases (see Fig. 4.6 for case NM31 and Fig. 4.7 for case NM22). In all other cases, the differences in UMSE are very small (see Fig. 4.6 for case NL11 and Fig. 4.7 for case NL42). The UMSE of the CMV-estimator is appreciably higher than that of the other procedures only at $\Delta^2 = 0$.

In general, the CMV-estimator performs best in terms of estimation accuracy, as reflected in the values of the unconditional mean squared error. Since the UMSE takes bias as well as variance into account, a procedure that performs well with respect to this criterion should be preferred.

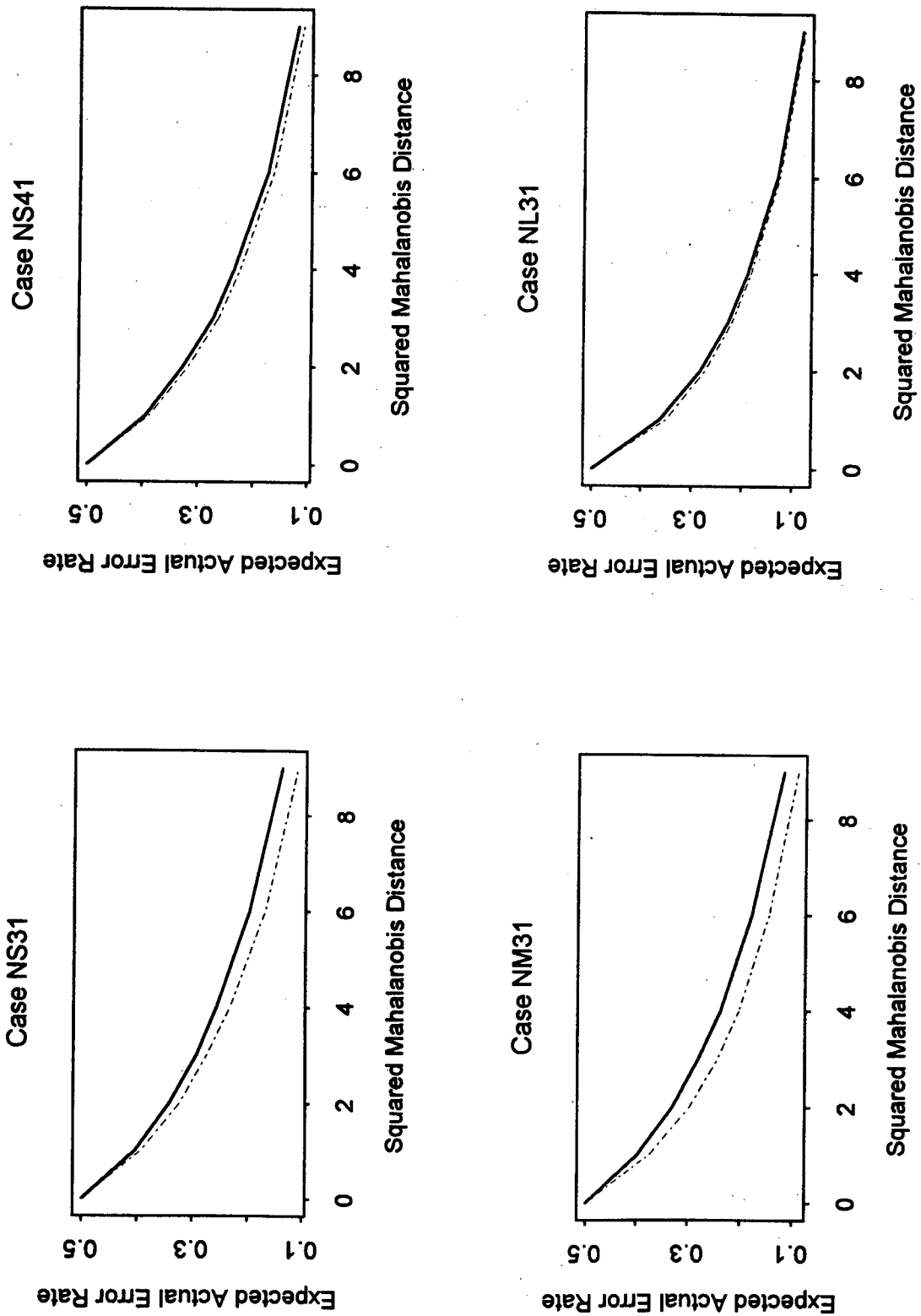
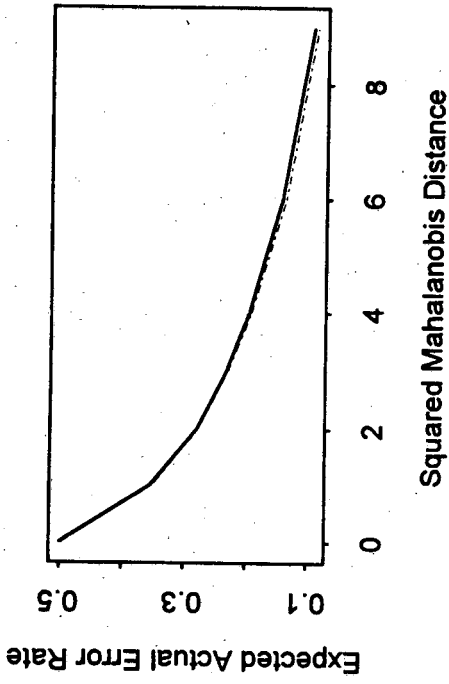
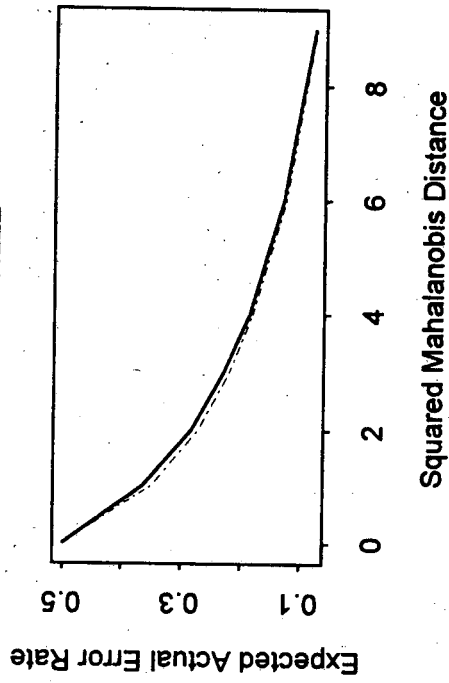


FIG. 4.1: EXPECTED ACTUAL ERROR RATE, UNCORRELATED NORMAL DATA

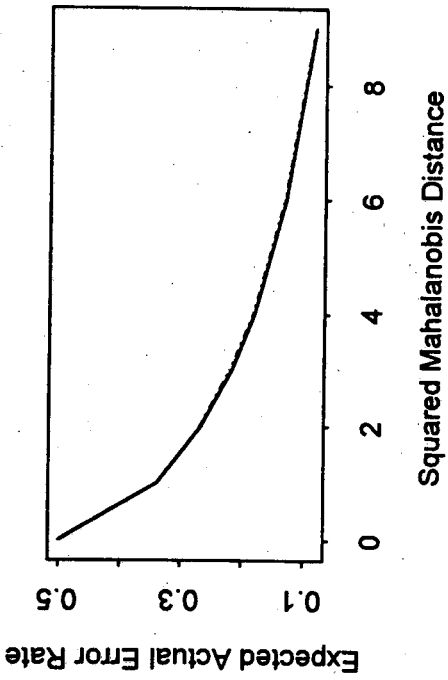
Case NM12



Case NL22



Case NS32



Case NM22

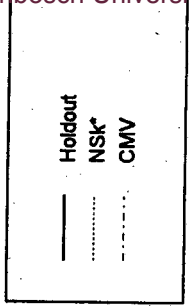
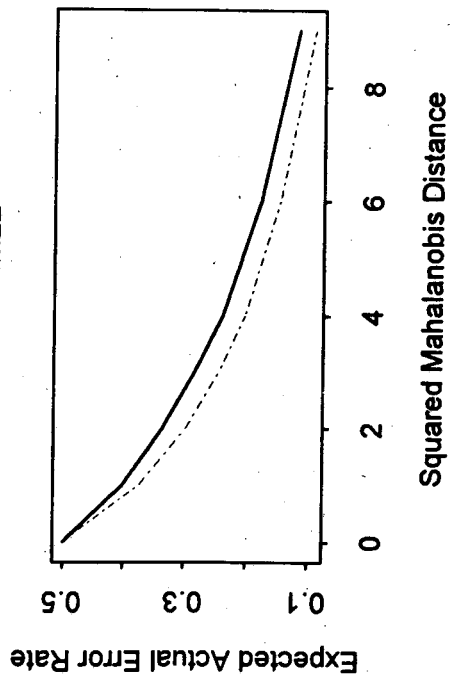
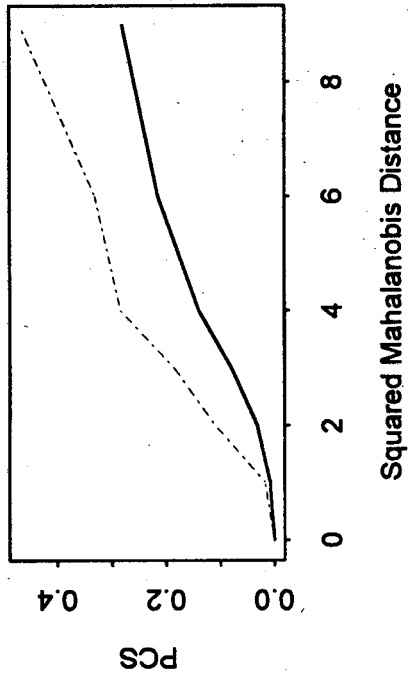
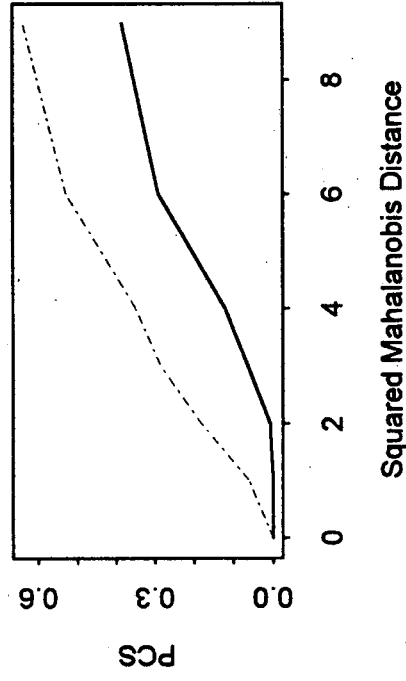


FIG. 4.2: EXPECTED ACTUAL ERROR RATE, CORRELATED NORMAL DATA

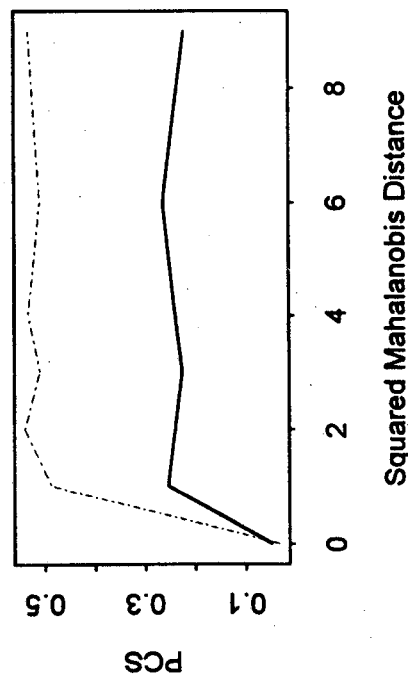
Case NM21



Case NL31



Case NS11



Case NL11

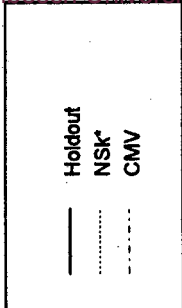
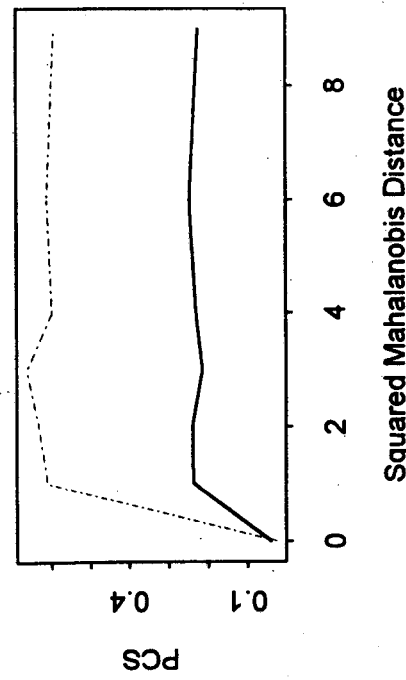
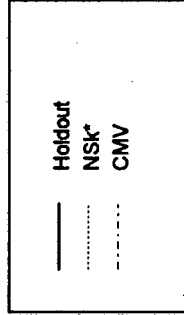
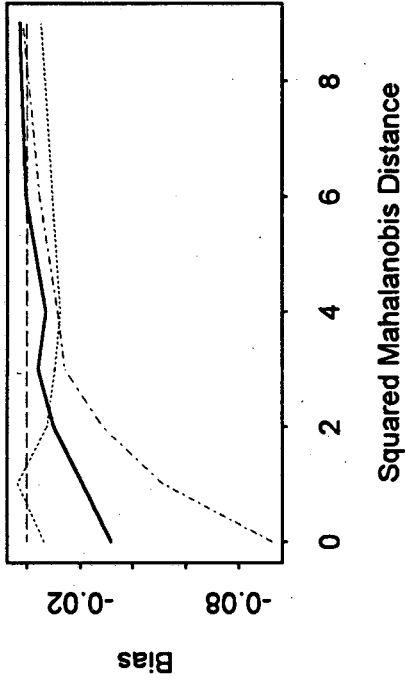
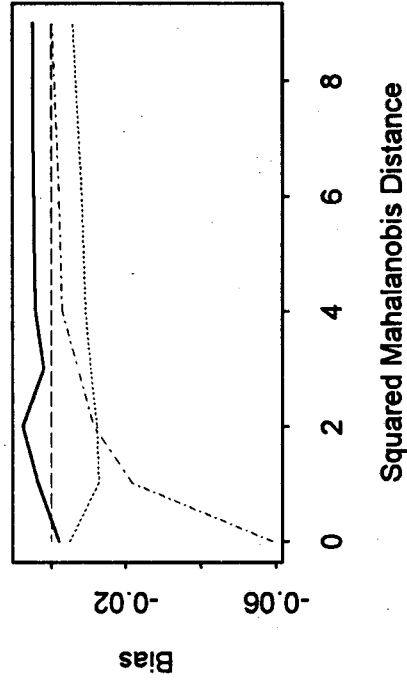


FIG. 4.3: PROBABILITY OF CORRECT SELECTION, UNCORRELATED NORMAL DATA

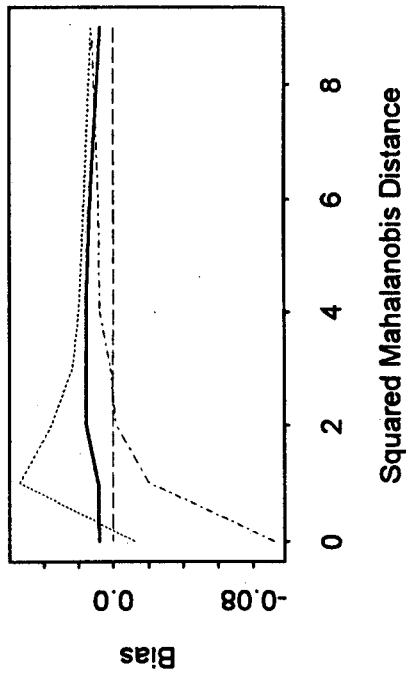
Case NM21



Case NL41



Case NS11



Case NL11

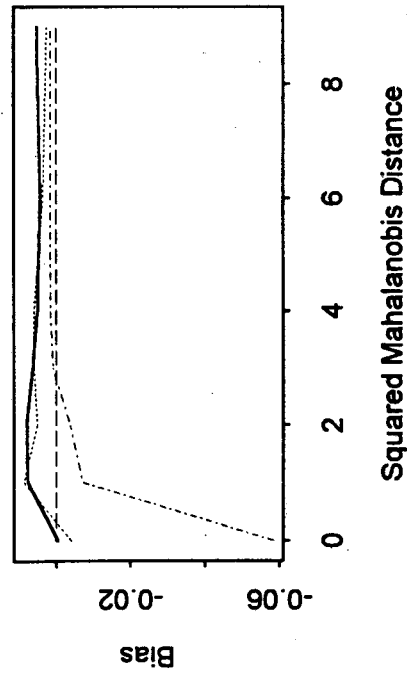
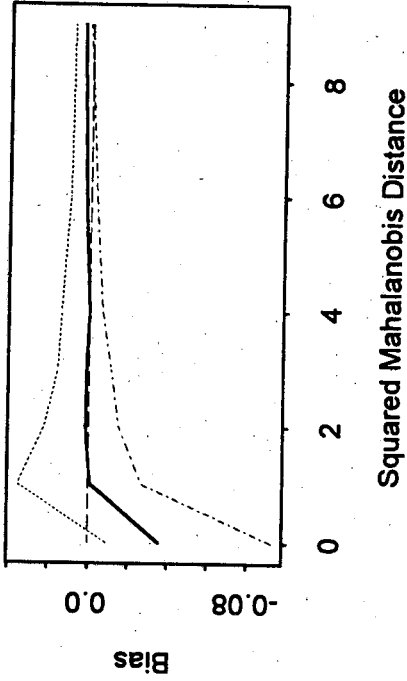
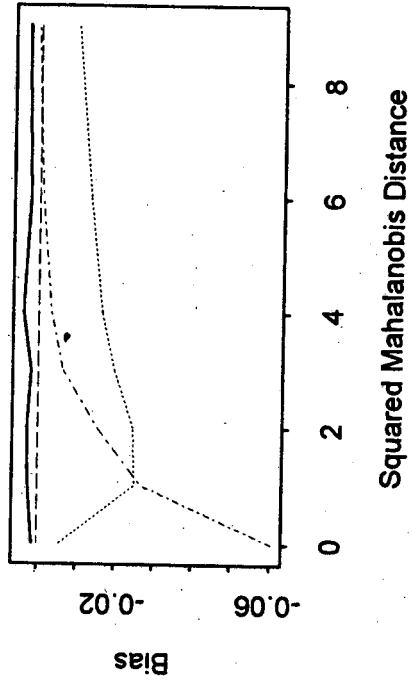


FIG. 4.4: BIAS OF ERROR RATE ESTIMATORS, UNCORRELATED NORMAL DATA

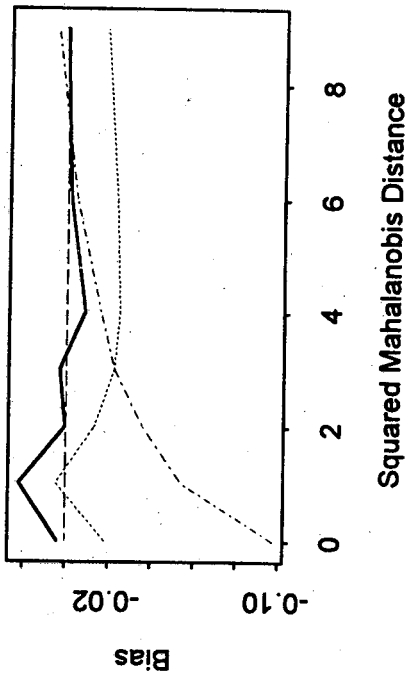
Case NM32



Case NL22



Case NS42



Case NL12

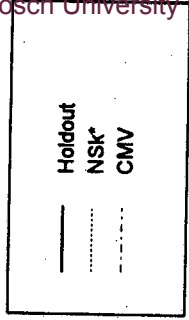
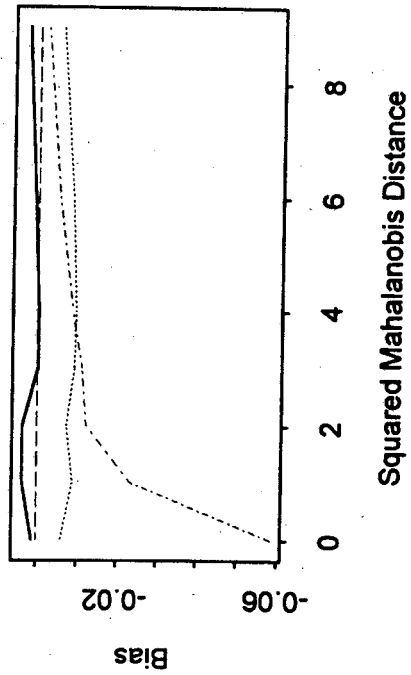
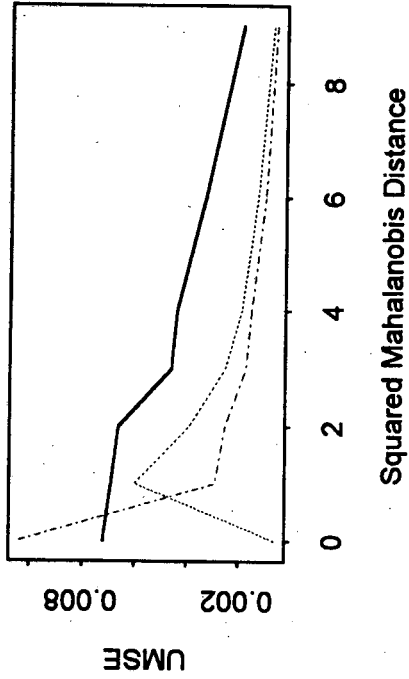
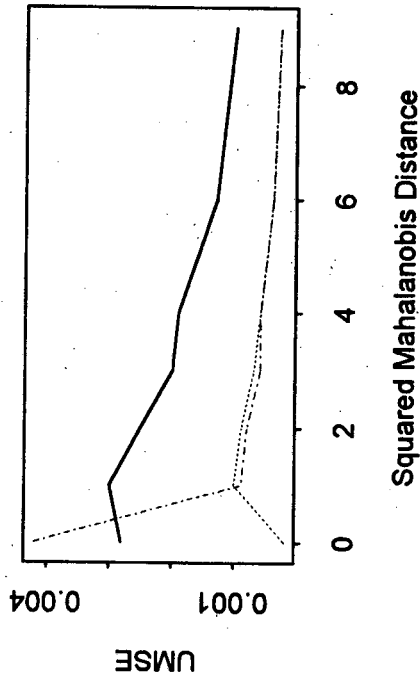


FIG. 4.5: BIAS OF ERROR RATE ESTIMATORS, CORRELATED NORMAL DATA

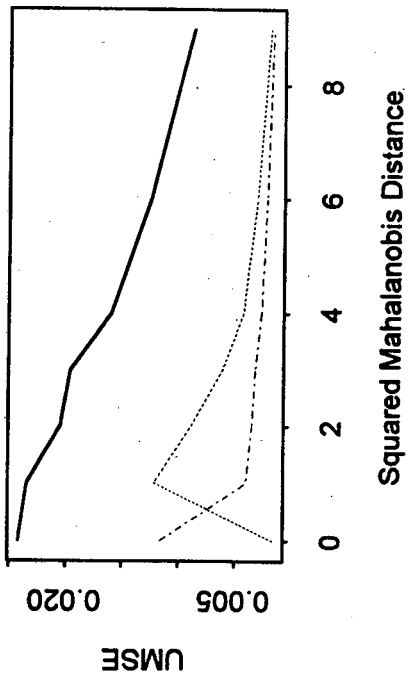
Case NM11



Case NL11



Case NS11



Case NM31

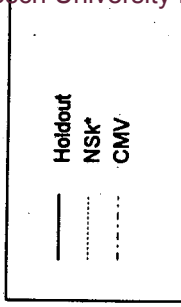
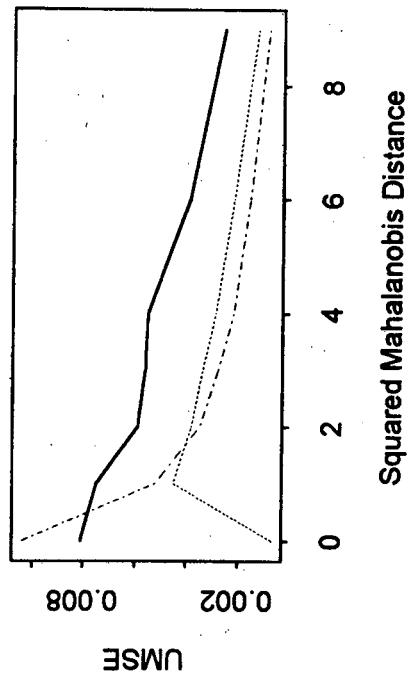
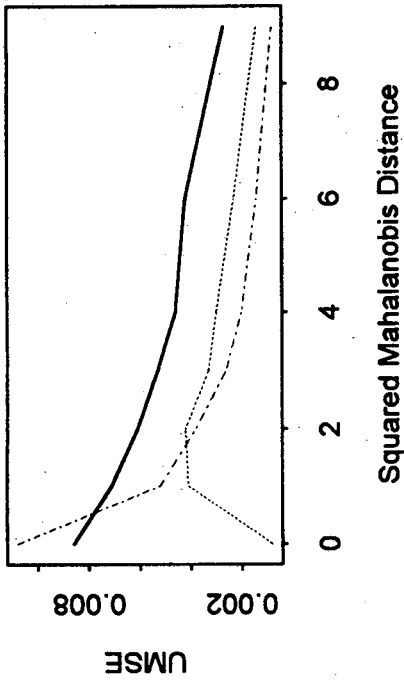
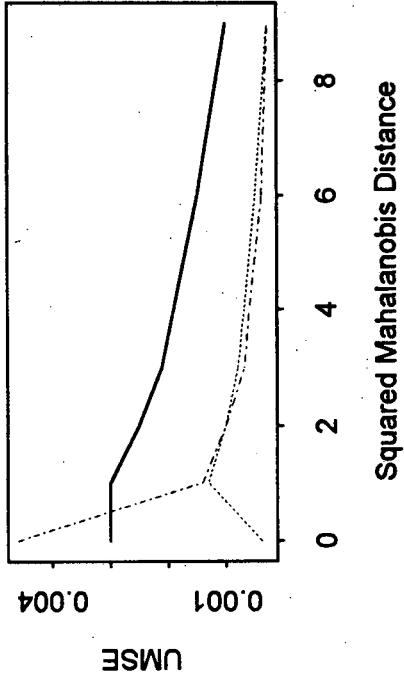


FIG. 4.6: UNCONDITIONAL MEAN SQUARED ERROR OF ERROR RATE ESTIMATORS, UNCORRELATED NORMAL DATA

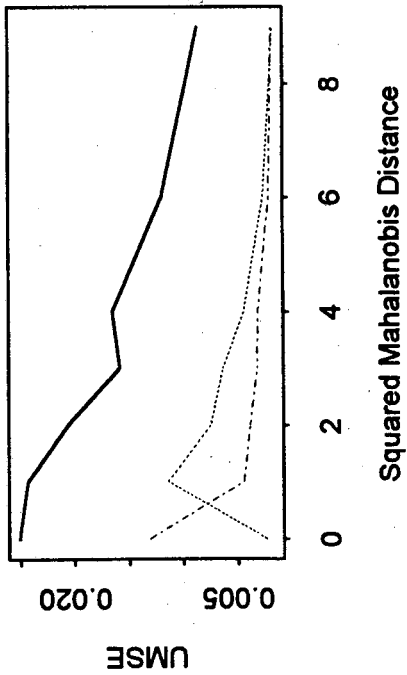
Case NM22



Case NL42



Case NS32



Case NM32

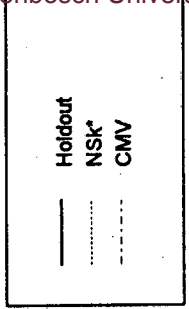
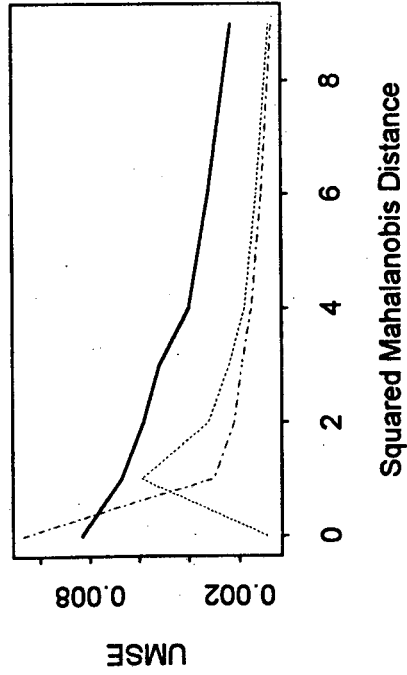


FIG. 4.7: UNCONDITIONAL MEAN SQUARED ERROR OF ERROR RATE ESTIMATORS, CORRELATED NORMAL DATA

4.5.1.2 THE DOUBLE EXPONENTIAL CASE

In the double exponential case, the Monte Carlo simulation study was limited to cases where the feature variables were uncorrelated. To estimate the quantities of interest, 500 Monte Carlo repetitions were used at each value of Δ^2 . For each repetition, a training data set was generated from the two relevant double exponential distributions. The different techniques were applied to the training data to select a subset. For each of the selected subsets, the post-selection actual error rates were estimated using simulation. To do this, a large number (500 per group) of entities were generated, and classified using the Anderson classification statistic based on each of the selected subsets. The three different post-selection error rate estimators, viz. the NS_k^* -estimator, the holdout estimator and the CMV-estimator, were also calculated. The expected post-selection actual error rates were obtained by averaging the 500 actual error rates obtained for each technique. Estimates of the PCS, bias and unconditional mean squared errors were obtained in the same way as in the normal case.

The results of the simulation study are summarised by means of graphs. A representative selection of these graphs is given in Figs. 4.8 - 4.11. Graphs of the post-selection expected actual error rates appear in Fig. 4.8, while the PCS associated with the procedures is displayed in Fig. 4.9. Fig. 4.10 contains graphs of the bias of the three error rate estimators, and graphs of the unconditional mean squared errors of the error rate estimators are given in Fig. 4.11.

The factors mentioned at the beginning of Section 4.5, identify a total of 12 double exponential cases. The coding DS1, DS2, DS3 and DS4 is used to denote the small sample cases with $r=1$, $r=5$, $r=10$ (with $\mu_{il}, l=1, \dots, 10$ given by (4.5.2)) and $r=10$ (with $\mu_{il}, l=1, \dots, 10$ given by (4.5.3)), in that order. Similar coding, with DM and DL instead of DS, is used for the mixed and large sample cases respectively.

SELECTION PERFORMANCE

Expected Actual Error Rate

In the double exponential case, cross model validation generally performs better than the other procedures. Although the actual error rates were often approximately equal (see Fig. 4.8 for cases DM4 and DL1), there were a number of cases where cross model validation performed appreciably better (see Fig. 4.8 for cases DS3 and DM3). The expected actual error rate associated with the cross model validation procedure is never larger than that of the other procedures.

Probability of Correct Selection (PCS)

For double exponential data behaviour similar to that in the normal case is displayed. In the cases DS1, DM1, DL1 the cross model validation selection performed very well, achieving PCS between 0.5 and 0.8, compared to PCS of between 0.2 and 0.25 achieved by the other procedures (see Fig. 4.9 for case DS1). In cases DS2, DM2 and DL2, cross model validation also outperformed the other procedures, but the difference in PCS is not as large as in the previous cases (see Fig 4.9 for case DM2). In cases DS3, DM3 and DL3, cross model validation also performed best, the difference between the procedures increasing with Δ^2 (see Fig. 4.9 for case DM3). In cases DS4, DM4 and DL4 none of the procedures performed well with respect to PCS. (see Fig. 4.9 for case DL4).

ESTIMATION PERFORMANCE

Bias

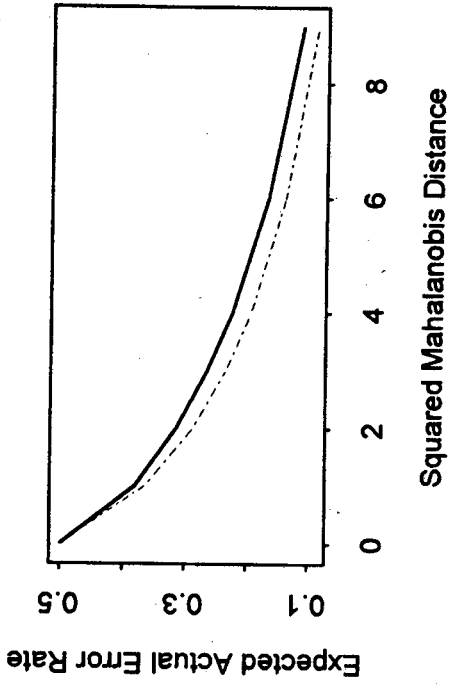
The performance of the error rate estimators in terms of bias is largely the same as in the normal case. The holdout estimator is nearly unbiased, except in small and mixed sample cases at small values of Δ^2 (see Fig. 4.10 for cases DS1 and DM4). The NS_k^* -estimator is less biased than the CMV-estimator at small values of Δ^2 , while the opposite is true for moderate to large values of Δ^2 (see Fig. 4.10 for cases DS1, DM4, DL2 and DL3).

Unconditional Mean Squared Error

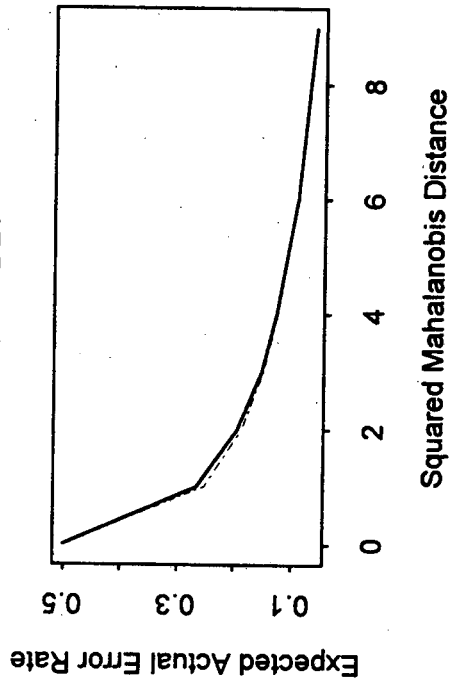
In most of the double exponential cases, the unconditional mean squared error of the holdout estimator is much larger than that of the other two error rate estimators. An exception to this is case DM1 (see Fig. 4.11) where the UMSE of the NS_k^* -estimator is larger than that of the holdout estimator at $\Delta^2 = 1$. In cases DS1, DM1 and DL1 the UMSE of the cross model validation error rate estimator is much smaller than that of the NS_k^* -estimator, especially for small values of Δ^2 (see Fig. 4.11 for cases DS1 and DM1). In cases DS2, DS3, DS4, DM2, DM3 and DM4 the UMSE of the CMV-estimator is also smaller than that of the NS_k^* -estimator, but the difference is smaller than in the previous cases (see Fig. 4.11 for case DM3). In cases DL2, DL3 and DL4, the difference between the UMSE-values of these two estimators is very small (see Fig. 4.11 for case DL2). Only in cases DL1 and DL2 is the UMSE of the cross model validation error rate estimator slightly higher than that of the NS_k^* -estimator, at a few of the Δ^2 -values considered (see Fig 4.11 for case DL2).

The CMV-estimator generally performs best in terms of UMSE. It consistently outperforms the holdout procedure (except at $\Delta^2 = 0$) and also outperforms the NS_k^* -estimator in almost all cases. Except at $\Delta^2 = 0$, it never performs appreciably worse than any of the other two estimators.

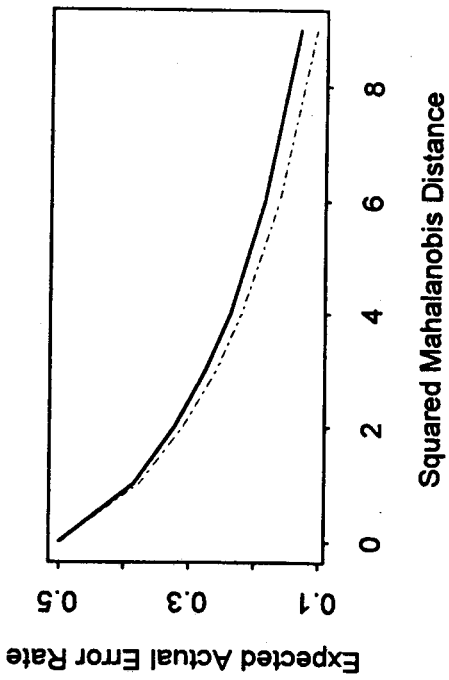
Case DM3



Case DL1



Case DS3



Case DM4

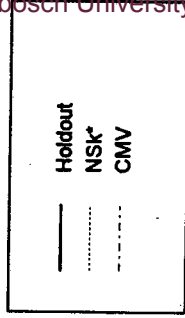
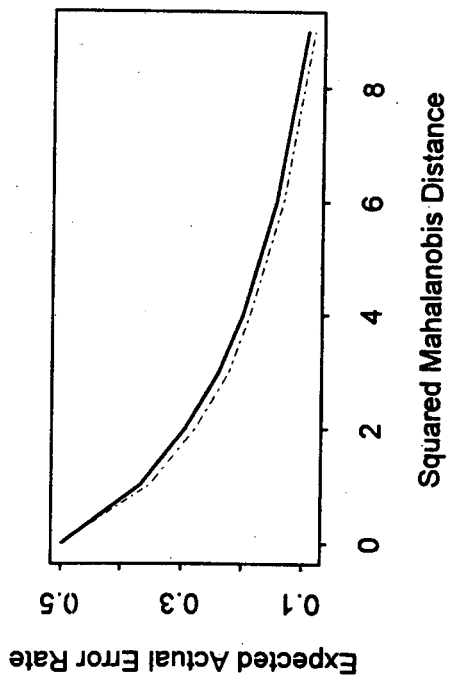
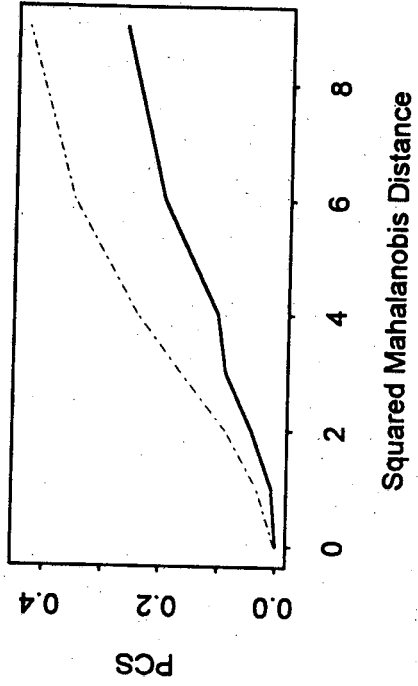
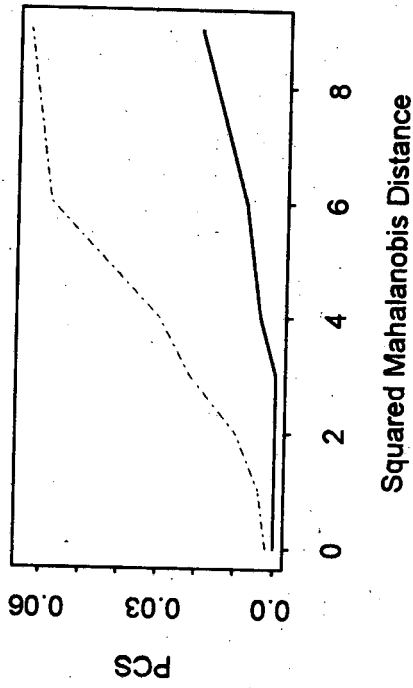


FIG. 4.8: EXPECTED ACTUAL ERROR RATE, DOUBLE EXPONENTIAL DATA

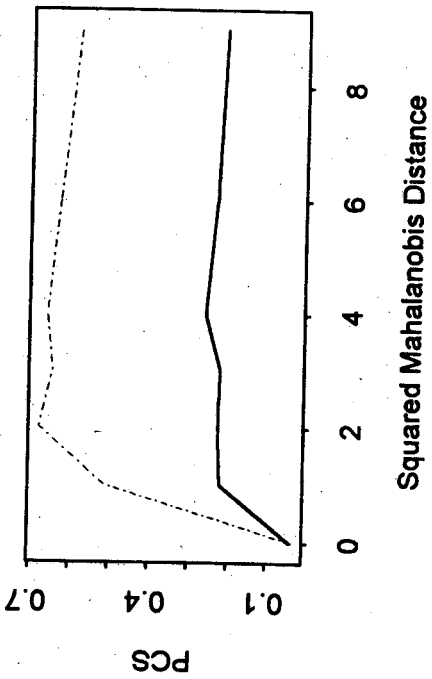
Case DM2



Case DL4



Case DS1



Case DM3

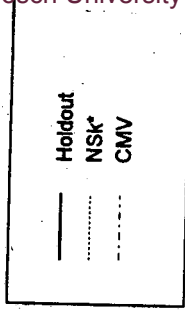
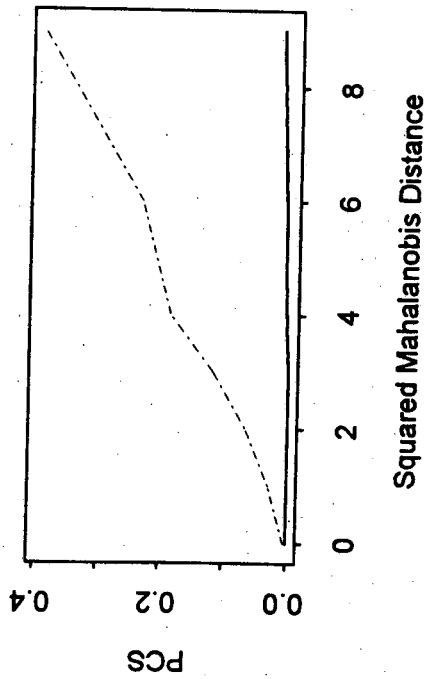
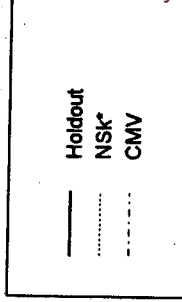
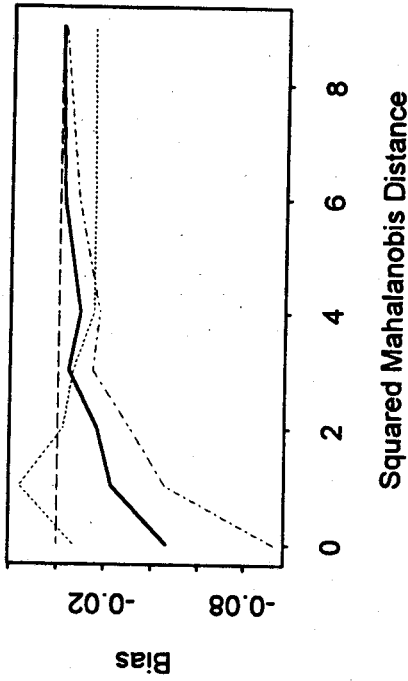
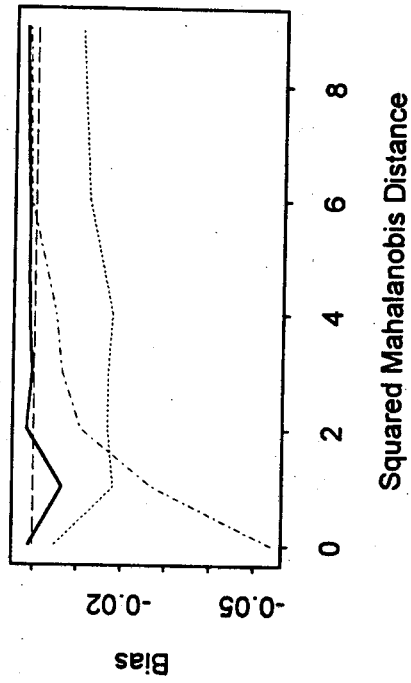


FIG. 4.9: PROBABILITY OF CORRECT SELECTION, DOUBLE EXPONENTIAL DATA

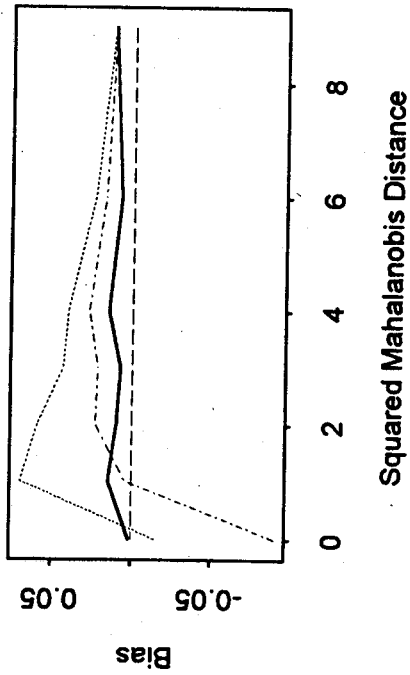
Case DM4



Case DL3



Case DS1



Case DL2

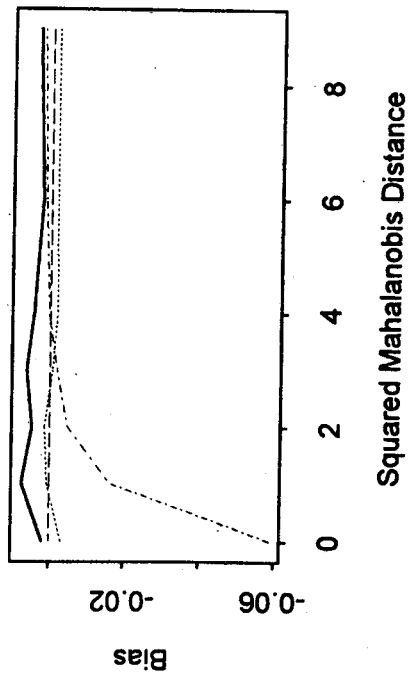
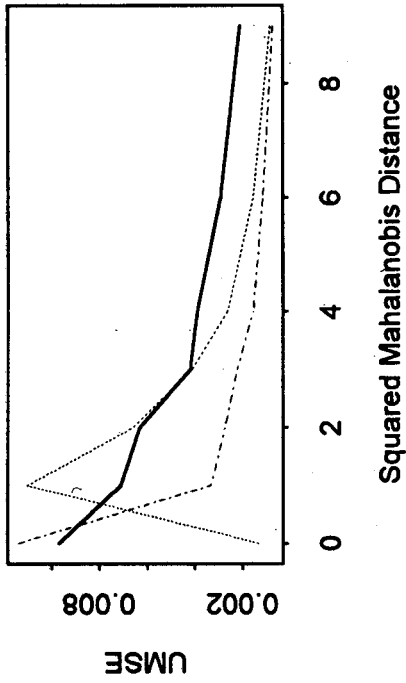
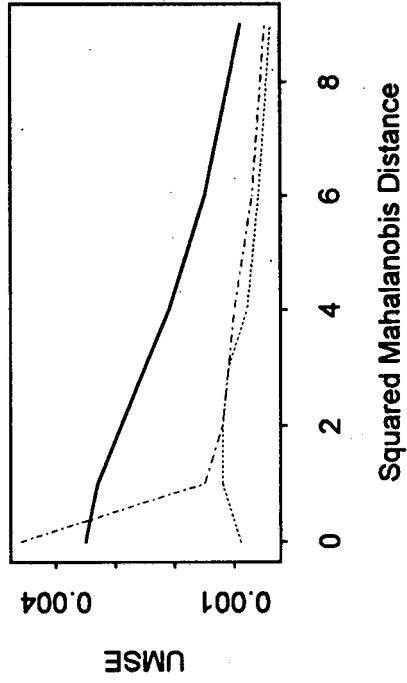


FIG. 4.10: BIAS OF ERROR RATE ESTIMATORS, DOUBLE EXPONENTIAL DATA

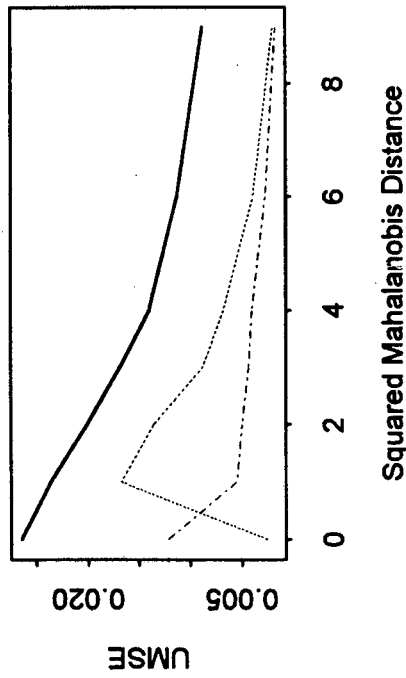
Case DM1



Case DL2



Case DS1



Case DM3

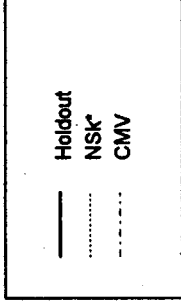
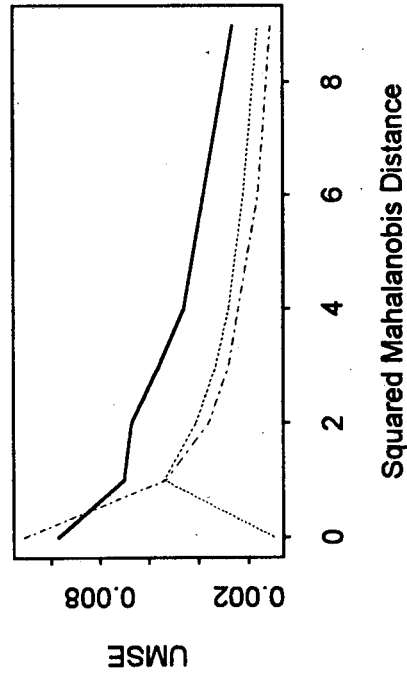


FIG. 4.11: UNCONDITIONAL MEAN SQUARED ERROR OF ERROR RATE ESTIMATORS, DOUBLE EXPONENTIAL DATA

4.5.1.3 THE LOGNORMAL CASE

In the lognormal case, a Monte Carlo simulation study similar to that in the double exponential case was done to compare the selection and estimation performance of the three procedures. To estimate the required quantities, 500 Monte Carlo repetitions were done at each value of Δ^2 . For each repetition, a training data set was generated from the two relevant lognormal distributions. The different techniques were applied to the training data to select a subset of variables. For each of the selected subsets, the post-selection actual error rate was estimated using simulation. To do this, a large number (500 per group) of entities were generated from the relevant lognormal distributions, and classified using the Anderson classification statistic based on each of the selected subsets. The three different post-selection error rate estimators, were also calculated. The expected post-selection actual error rates were obtained by averaging the 500 actual error rates obtained for each technique. Estimates of the PCS, bias and unconditional mean squared errors were obtained in the same way as in the normal and double exponential cases.

The results of the simulation study are summarised by means of graphs. A representative selection of these graphs appears in Figs. 4.12 - 4.15. In Fig. 4.12 graphs of the post-selection expected actual error rates are shown, while graphs of the PCS associated with the procedures appear in Fig. 4.13. Graphs of the bias of the three error rate estimators are given in Fig. 4.14, and Fig. 4.15 contains graphs of the unconditional mean squared errors of the error rate estimators.

The factors mentioned at the beginning of Section 4.5 identify a total of 12 different lognormal cases. For small samples, the cases $r=1$, $r=5$, $r=10$ (with $\mu_{1l}, l=1, \dots, 10$ equal) and $r=10$ (with $\mu_{1l}, l=1, \dots, 10$ equi-spaced), are denoted by the coding LS1, LS2, LS3 and LS4, in that order. For the mixed and large sample cases, similar coding with LM and LL instead of LS, is used.

SELECTION PERFORMANCE

Expected Actual Error Rate

In the case of lognormal data, the CMV procedure generally performed better than the other procedures. Although the expected actual error rates were often approximately equal (see Fig 4.12 for case LL4), there were a number of cases where the CMV procedure performed appreciably better, namely LS1, LS3, LM1, LM3 and LL1 (see Fig 4.12 where cases LS1, LS3 and LM3 appear).

Probability of Correct Selection (PCS)

The behaviour in the lognormal case is similar to that in the normal case. In cases LS1, LM1, and LL1, the CMV procedure performed very well, achieving PCS between 0.5 and 0.7, compared to PCS of approximately 0.2 for the other procedures (see Fig. 4.13 for cases LS1 and LL1). In cases LS2, LM2, LL2, LS3, LM3 and LL3 the CMV procedure also yielded higher PCS-values than the other techniques, but the difference is not as large as before (see Fig. 4.13 for case LM2). In cases LS4, LM4 and LL4, none of the procedures achieved high PCS (see Fig. 4.13 for case LS4).

ESTIMATION PERFORMANCE

Bias

As in the normal and double exponential cases, the holdout estimator has very small bias in the lognormal case, except in some small and mixed sample cases at small values of Δ^2 (see Fig. 4.14 for case LM2). The NS_k^* -estimator seems to perform worse than in the normal and double exponential cases, often being more biased than the CMV-estimator even at small values of Δ^2 (see Fig. 4.14 for cases LS1 and LL1, where the NS_k^* -estimator has larger bias than the CMV-estimator at all values of Δ^2 , except $\Delta^2 = 0$). This is in agreement with the findings of Snapinn and Knoke (1989), that the performance of the NS_k^* -estimator is adversely influenced by skewness of the distribution of the feature data. The CMV-estimator once more has fairly large bias at $\Delta^2 = 0$, but the bias decreases with increasing Δ^2 (see Fig. 4.14 for cases LS1, LM2 and LL3).

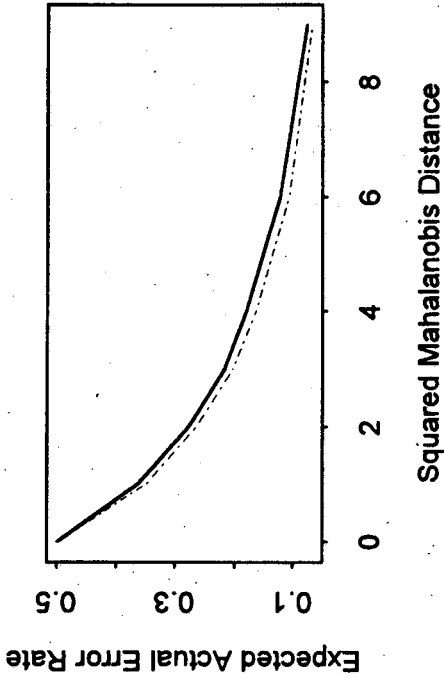
Unconditional Mean Squared Error

In most of the lognormal cases, the unconditional mean squared error of the holdout estimator is much larger than that of the other two error rate estimators. Exceptions to this are cases LS1, LM1 and LL1 where the UMSE of the NS_k^* -estimator is larger than that of the holdout estimator at $\Delta^2 = 1, 2$ and 4 (see Fig. 4.15 for cases LS1 and LM1). In cases LS1, LM1 and LL1 the UMSE of the cross model validation error rate estimator is much smaller than that of the NS_k^* -estimator, especially for small values of Δ^2 (see Fig. 4.11 for cases LS1 and LM1). In cases LS2, LS3, and LS4 the UMSE of the CMV-estimator is also smaller than that of the NS_k^* -estimator, but the difference is not as large as in the previous cases (see Fig. 4.15 for case LS2). In cases LM2, LM3, LM4, LL2, LL3 and LL4, the difference between the UMSE-values of these two estimators is very small (see Fig. 4.15 for case LL3). Only in cases LM3 and LL3 is

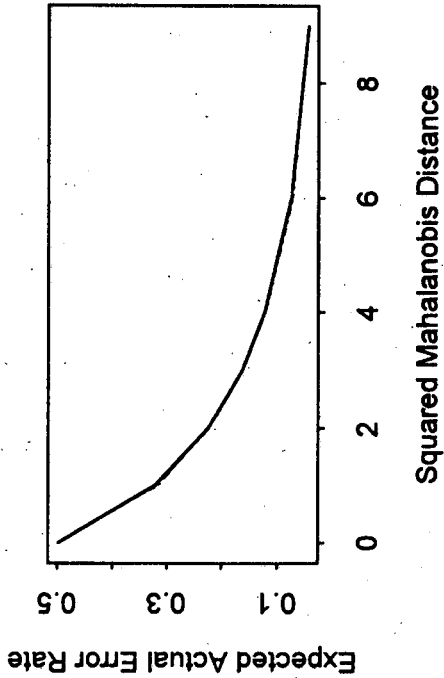
the UMSE of the cross model validation error rate estimator slightly higher than that of the NS_k^* - estimator, at a few of the Δ^2 -values considered (see Fig. 4.15 for case LL3).

As for the normal and double exponential cases, the CMV-estimator generally performs best in terms of UMSE. It consistently outperforms the holdout procedure (except at $\Delta^2 = 0$) and also outperforms the NS_k^* - estimator in almost all cases. Except at $\Delta^2 = 0$, it never performs appreciably worse than any of the other two estimators.

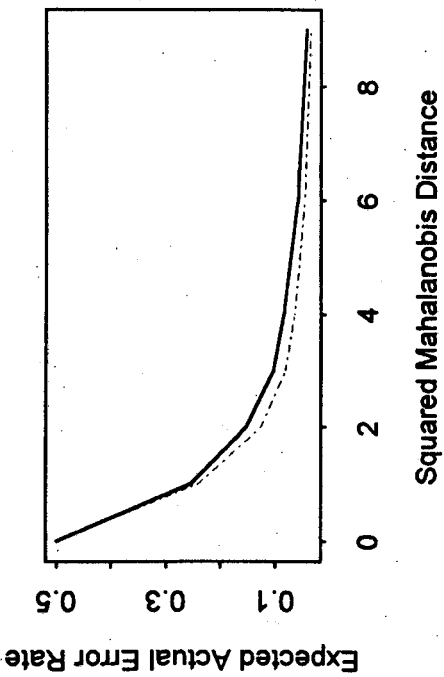
Case LS3



Case LL4



Case LS1



Case LM3

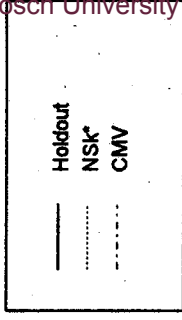
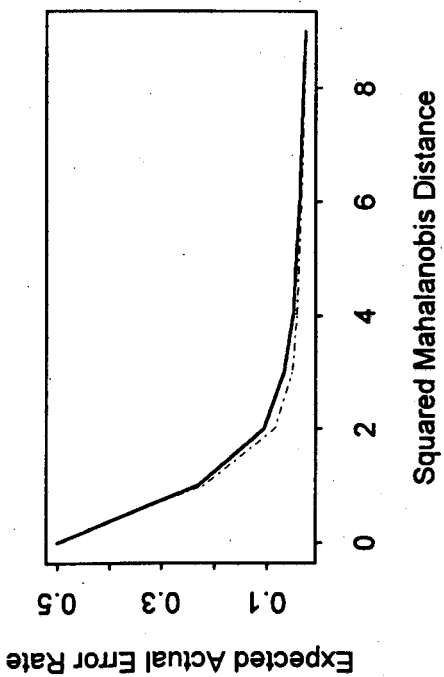


FIG. 4.12: EXPECTED ACTUAL ERROR RATE, LOGNORMAL DATA

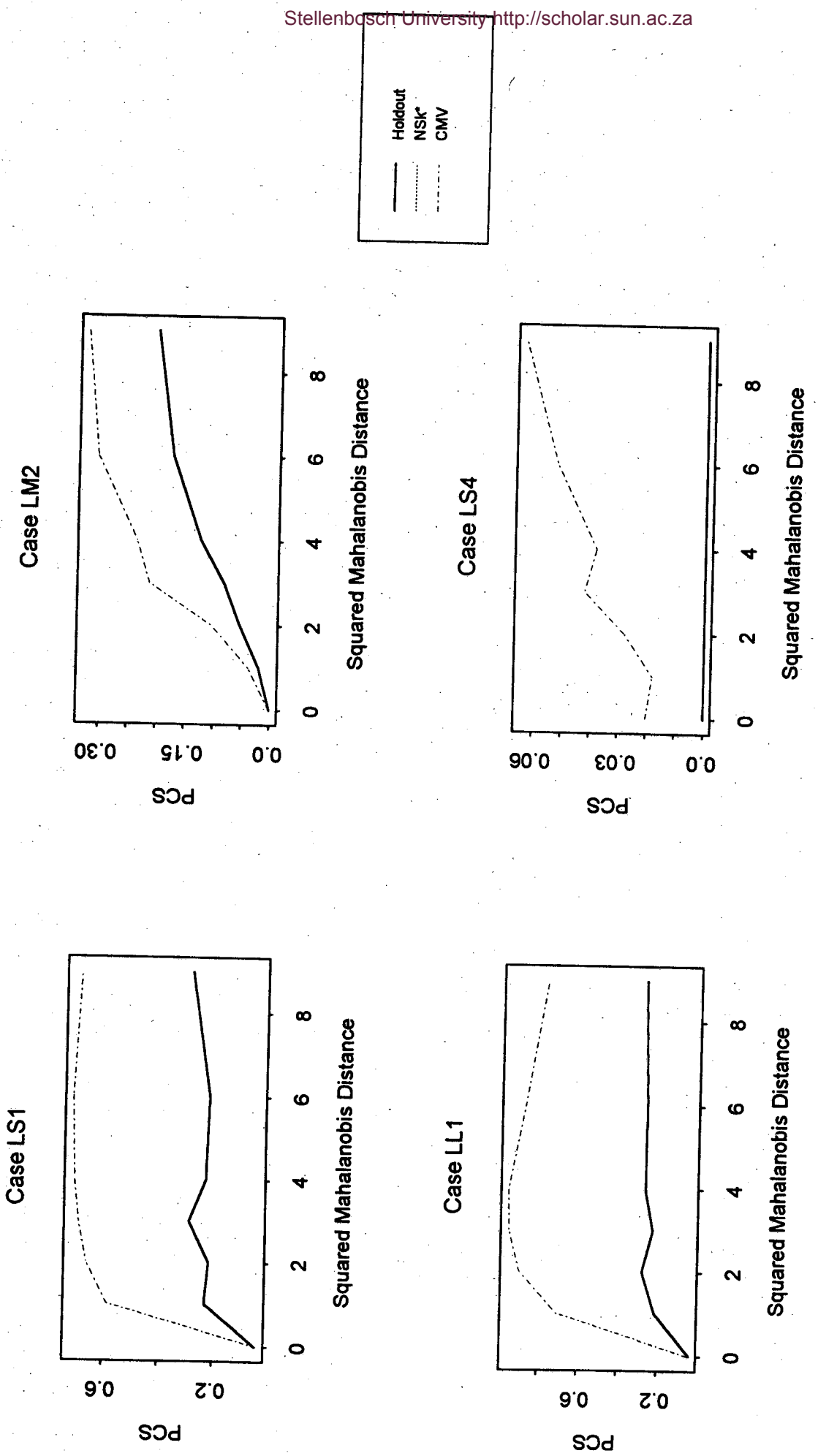
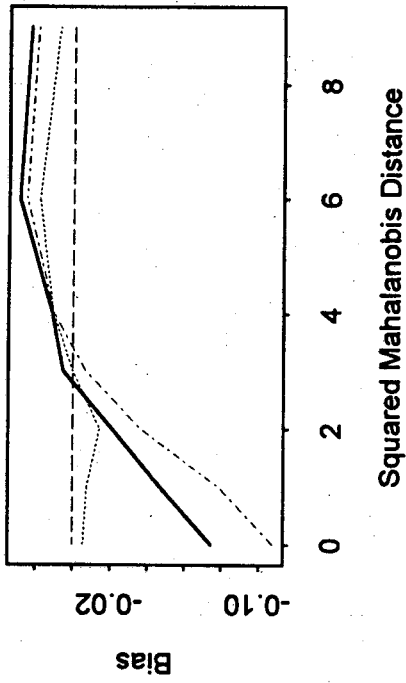
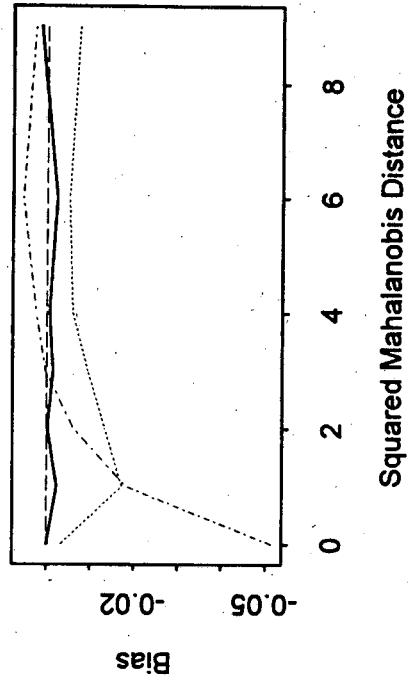


FIG. 4.13: PROBABILITY OF CORRECT SELECTION, LOGNORMAL DATA

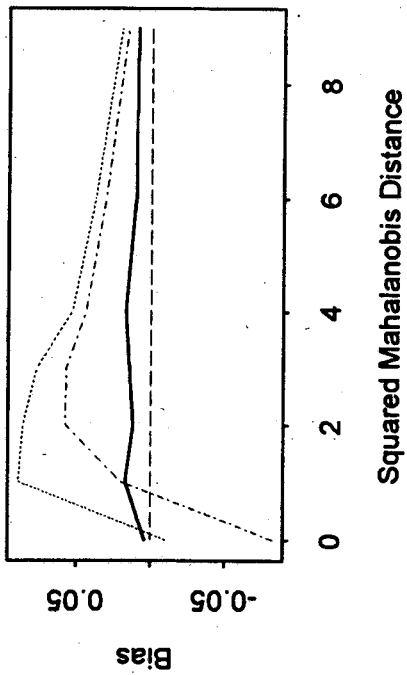
Case LM2



Case LL3



Case LS1



Case LL1

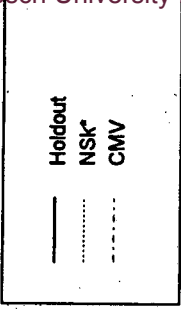
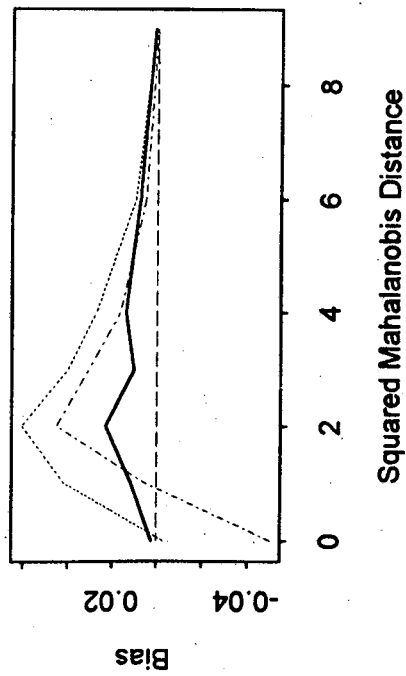
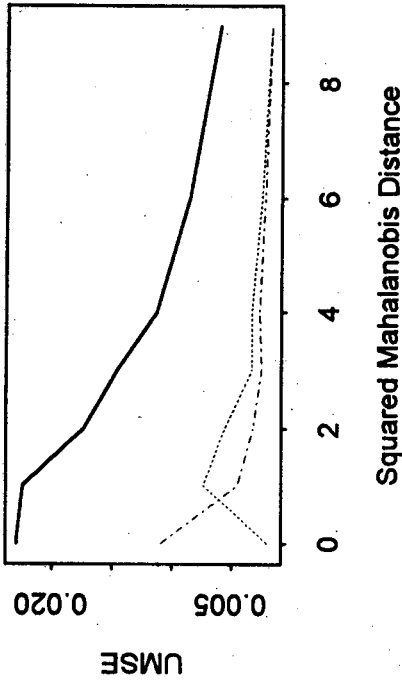
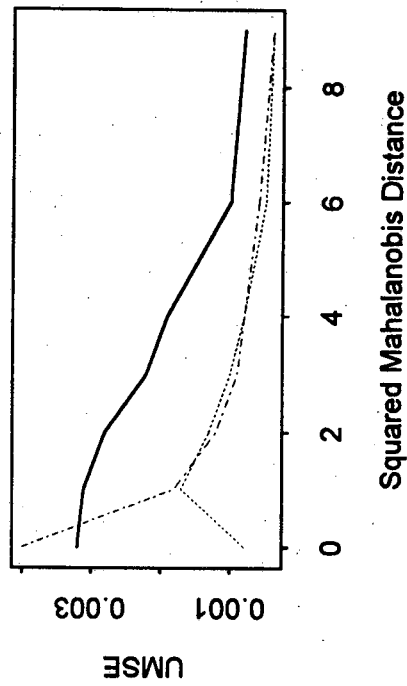


FIG. 4.14: BIAS OF ERROR RATE ESTIMATORS, LOGNORMAL DATA

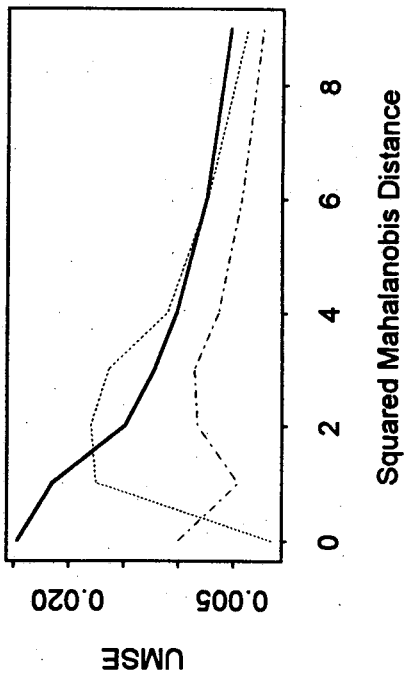
Case LS2



Case LL3



Case LS1



Case LM1

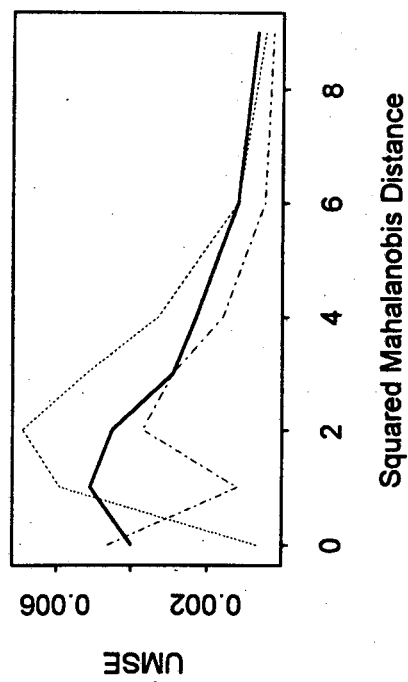


FIG. 4.15: UNCONDITIONAL MEAN SQUARED ERROR OF ERROR RATE ESTIMATORS, LOGNORMAL DATA

4.5.2 INNER CRITERION : ALL POSSIBLE SUBSETS SELECTION BASED ON R^2

As mentioned in Section 4.4, the main reason for using forward F-based selection as inner criterion in the cross model validation procedure investigated in Section 4.5.1, is to facilitate a comparison with the procedures of Snapinn and Knoke (1989) and Rutter et al. (1991). Using the same selection procedure employed by these authors made it possible to investigate the effect of the cross model validation step without involving other factors which could possibly lead to differences in performance. In the second part of the simulation study, an all possible subsets approach based on R^2 , was used as inner criterion in the cross model validation procedure, to investigate the effect of this on the performance of the technique. Preliminary simulation studies suggested that overfitting which occurred when using forward selection as inner criterion (cf. Section 4.4), is less prevalent when using an all possible subsets approach based on R^2 as inner criterion. It was therefore decided to follow the recommendation of Hjorth (1994) to choose the model dimension p_0 to minimise $CMV(p)$, rather than using the strategy involving ϕ outlined in Section 4.4. In practice, a plot of $CMV(p)$ against p may again be used in deciding on the final model dimension (see Section 4.9). Exactly the same cases included in the first part of the study, were investigated. The results of this study, will now be compared to that of the study described in Section 4.5.1.

The same 24 normal cases, 12 double exponential cases and 12 lognormal cases included in the simulation study described in Section 4.5.1, were included in this part of the simulation study. The aim of this study is to investigate the effect of using a different inner criterion, and of identifying the optimal dimension in a different way, as outlined above. To compare the results of the two studies, graphs of the post-selection expected actual error rates, probabilities of correct selection, bias and unconditional mean squared errors of the error rate estimators were constructed. Each of the graphs contains the results of the two different ways in which the cross model validation procedure was performed, viz. using F-based forward selection as inner criterion together with the strategy involving ϕ to identify the optimal model dimension (henceforth referred to as the CMV-1 procedure), and using an all possible subsets approach based on R^2 as inner criterion combined with identifying the optimal model dimension by minimising the CMV-criterion (henceforth referred to as the CMV-2 procedure). Since the relative performance of the two techniques for the three distributions considered are largely similar, only the normal case will be discussed. The same conclusions are also valid for the double exponential and lognormal cases. In Fig. 4.16 a selection of graphs of the post-selection actual error rates is given, while a selection of graphs showing the probability of correct selection (PCS) appears in Fig. 4.17. As discussed in Section 4.5, these quantities reflect the allocatory and separatory performance of the techniques. A selection of graphs of the bias and unconditional mean squared errors of the two error rate estimators is given in Figs. 4.18 and 4.19 respectively. These quantities give an indication of the estimation performance of the techniques.

4.5.2.1 SELECTION PERFORMANCE

Expected Actual Error Rate

The post-selection expected actual error rates achieved by the classification functions resulting from the CMV-1 and CMV-2 procedures, are virtually identical. This is not only true for the selection of cases (NS12, NM41, NL21 and NL32) shown in Fig. 4.16, but also for the other 20 normal cases not shown here. With respect to allocatory performance, the change of inner criterion and way of identifying of the optimal model dimension, appear to have no effect.

Probability of Correct Selection

The PCS of a procedure is defined as the probability to select all the seemingly relevant variables (defined as variables with respect to which the two populations differ) and no seemingly irrelevant variables. In cases NS11, NM11 and NL11 there is only one seemingly relevant variable. The PCS behaviour of CMV-1 and CMV-2 for these three cases are similar, and case NS11 is given as a representative example of this (see Fig. 4.17). The CMV-1 procedure achieves higher PCS in these cases than the CMV-2 procedure. The reason for this is the slight tendency of the CMV-2 procedure to overfit, resulting in more than one variable being selected, which will of course decrease the PCS. In cases NS21, NM21, NL21 there are five seemingly relevant variables. The tendency of the CMV-2 procedure to overselect, again led to its PCS being slightly lower than that of the CMV-1 procedure (see Fig. 4.17 for case NM21). In cases NS31, NM31, NL31, NS41, NM41 and NL41, there are ten seemingly relevant variables, and the tendency of the CMV-2 procedure to select less parsimonious models, leads to it having higher PCS than the CMV-1 procedure in these cases (see Fig. 4.17 for case NM31 and NL41). The CMV-1 procedure seems to select more parsimonious models, while still achieving the same post-selection expected actual error rates as the CMV-2 procedure.

4.5.2.2 ESTIMATION PERFORMANCE

Bias

Fig. 4.18 contains a selection of graphs of the bias of the error rate estimators yielded by the two CMV techniques. Perusal of these graphs shows that the differences in the bias are very slight. In some cases, the bias of the CMV-1-estimator and that of the CMV-2-estimator are virtually identical (see Fig. 4.18 for case NS12 and NM41), while there are very small differences at some values of Δ^2 in other cases (see Fig.

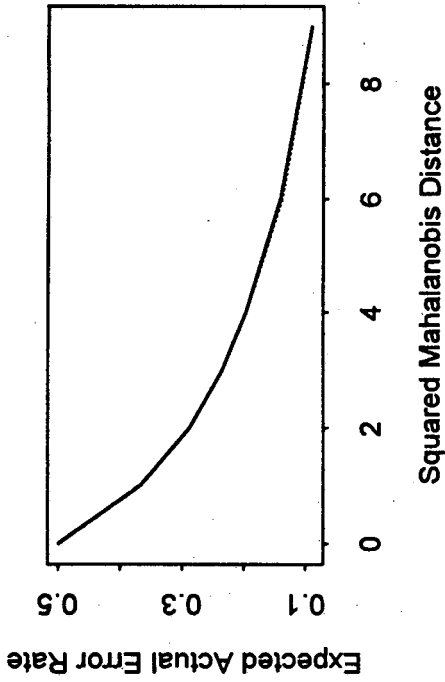
4.18 for case NL21 and NL32). The differences in the two ways of implementing the CMV procedure, do not seem to have appreciable influence on the bias of the resulting error rate estimators.

Unconditional Mean Squared Error

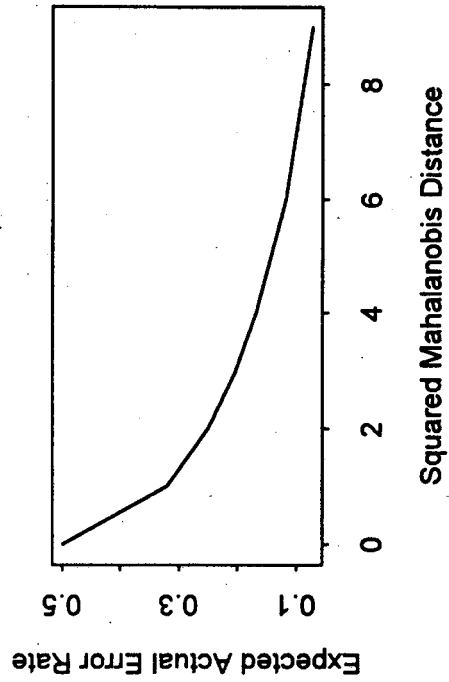
A representative selection of graphs of the unconditional mean squared errors of the CMV-1 and CMV-2 error rate estimators, appears in Fig. 4.19. As is the case with bias, the unconditional mean squared errors of the two estimators are virtually identical. At moderate to large values of Δ^2 ($\Delta^2 \geq 3$), the differences are almost non-existent, while at smaller values of Δ^2 ($\Delta^2 \leq 2$), very slight differences occur in some cases (see Fig. 4.19 for cases NM21 and NL32).

Overall, there seems to be very little difference between using an all possible subsets approach based on R^2 as inner criterion and using F-based forward selection as inner criterion. The strategy used to identify the optimal model dimension seems to influence only the PCS.

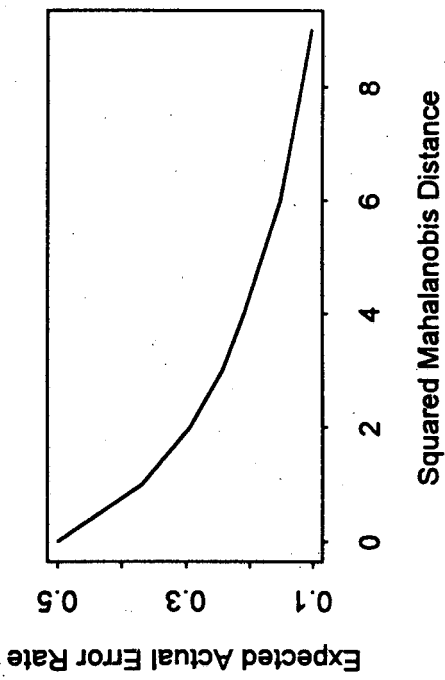
Case NM41



Case NL32



Case NS12



Case NL21

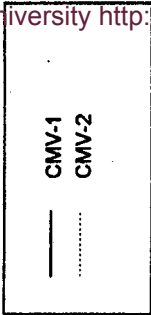
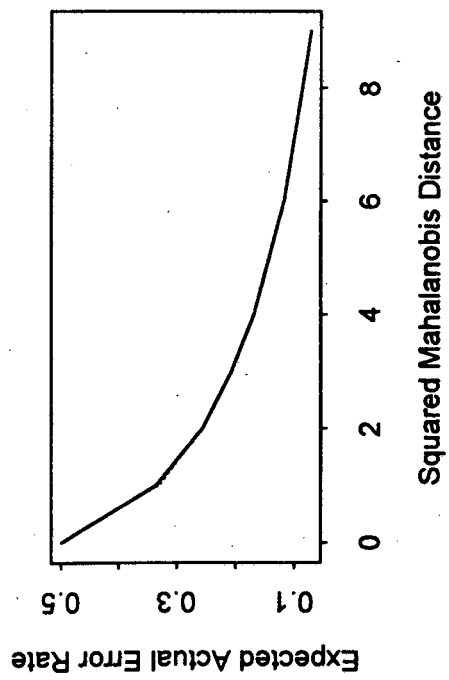
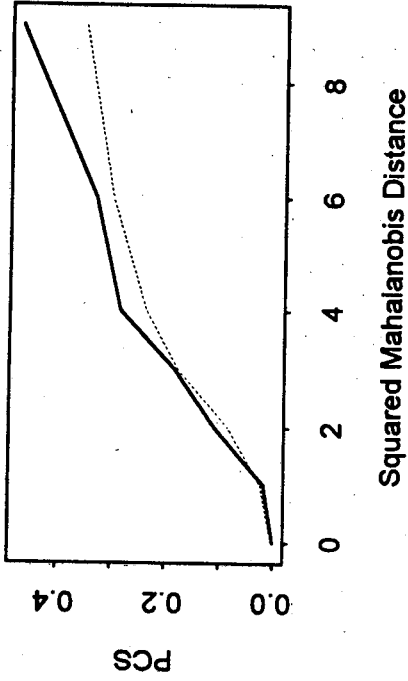
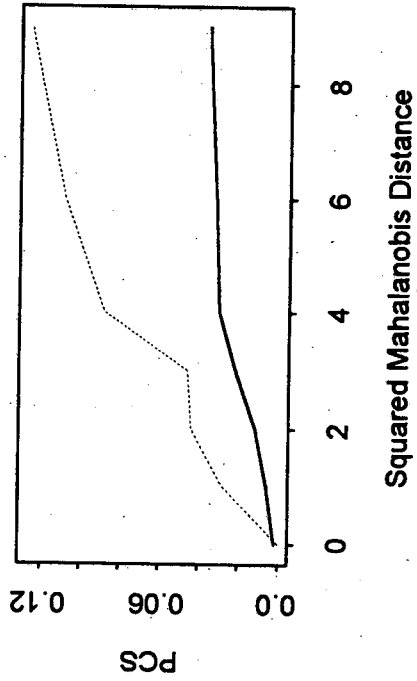


FIG. 4.16: EXPECTED ACTUAL ERROR RATE, NORMAL DATA

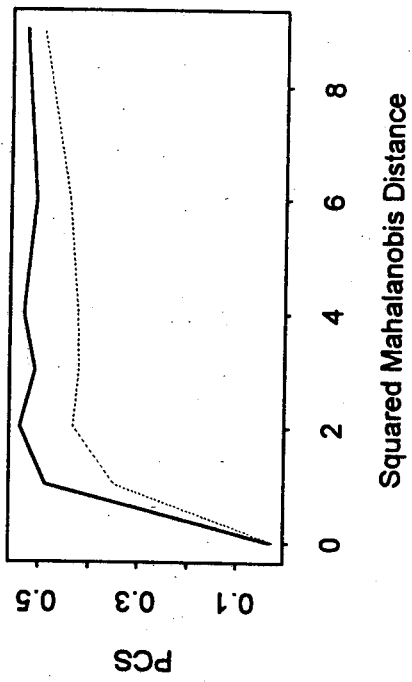
Case NM21



Case NL41



Case NS11



Case NM31

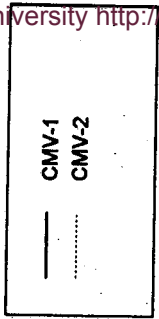
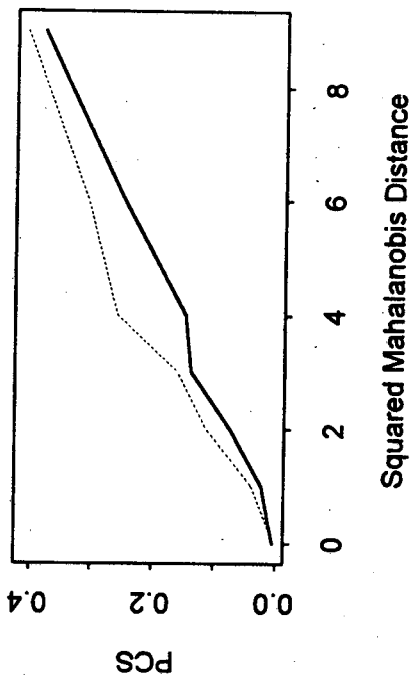


FIG. 4.17: PROBABILITY OF CORRECT SELECTION, NORMAL DATA

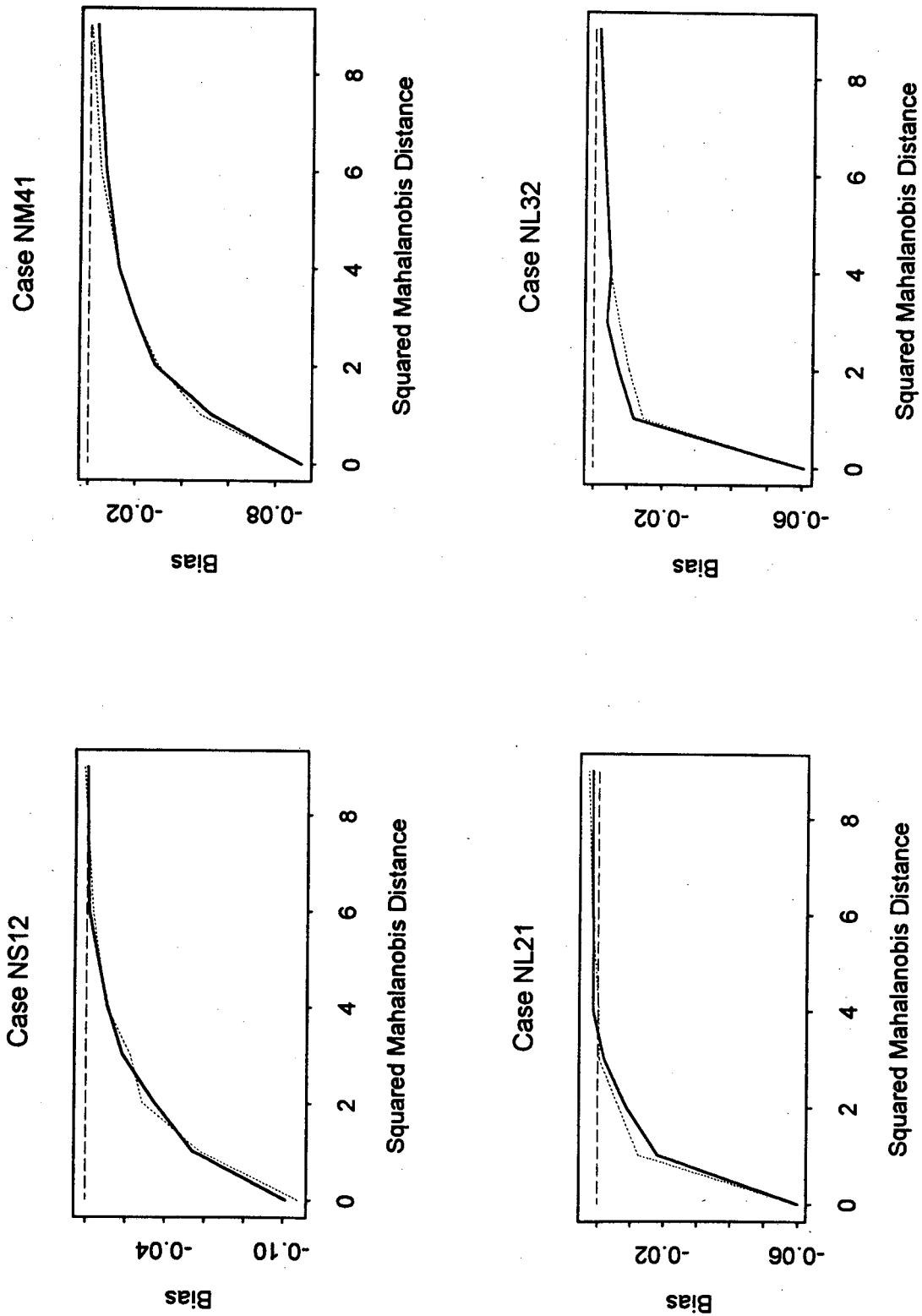
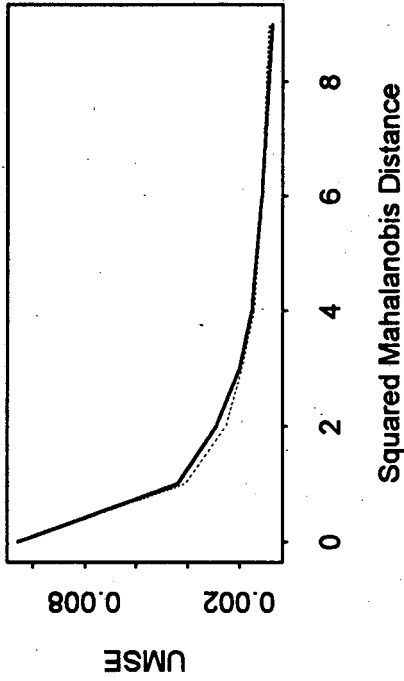
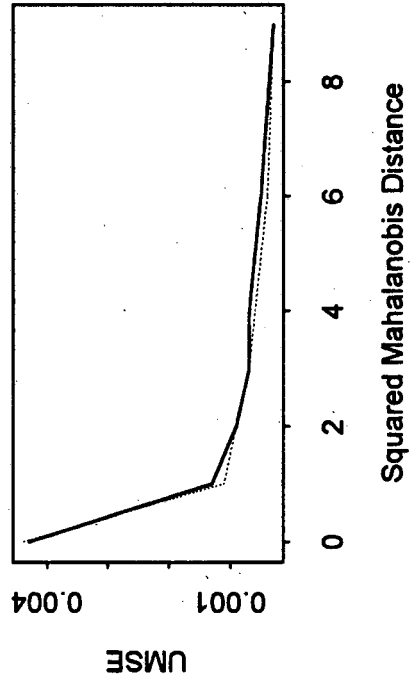


FIG. 4.18: BIAS OF ERROR RATE ESTIMATORS, NORMAL DATA

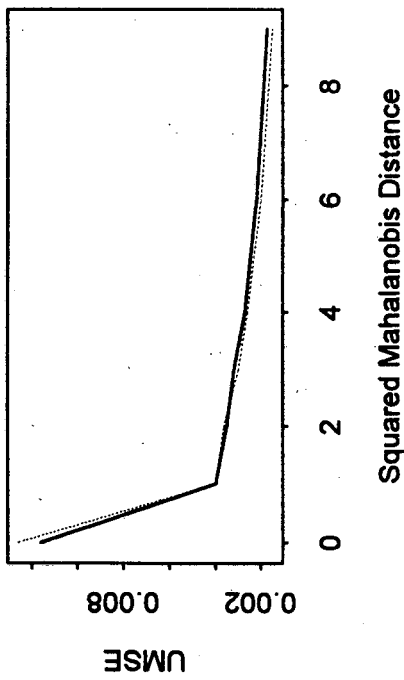
Case NM21



Case NL41



Case NS11



Case NL32

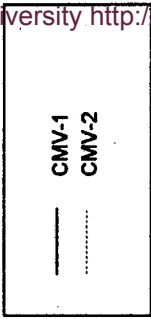
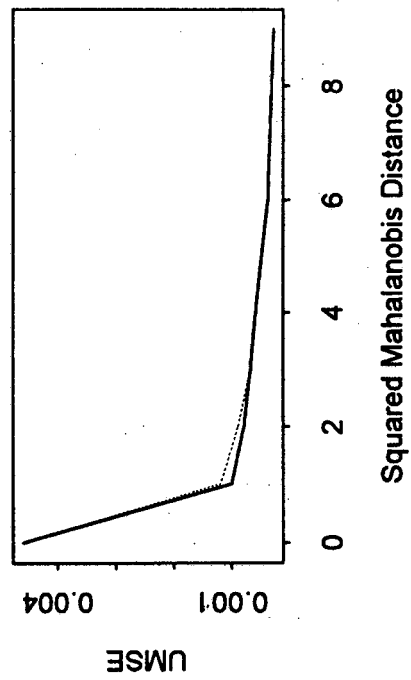


FIG. 4.19: UNCONDITIONAL MEAN SQUARED ERROR OF ERROR RATE ESTIMATORS, NORMAL DATA

4.6 CROSS MODEL VALIDATION IN LOGISTIC REGRESSION

Application of cross model validation in logistic regression proceeds analogously to application of the technique in discriminant analysis. Consider the logistic discriminant rule in the case of $G + 1 = 2$ groups, viz.

$$V(\mathbf{x}) \equiv \hat{c}_{10}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}'_1 \mathbf{x}. \quad (4.6.1)$$

This rule is obtained by replacing the unknown parameters in (2.1.8) by their maximum likelihood estimates. In (4.6.1), $\mathbf{x}: k \times 1$ is a vector of measurements obtained from an entity of unknown origin that has to be classified into one of the two groups, Π_0 and Π_1 . If $V(\mathbf{x}) \leq 0$ this entity is classified into Π_0 , and into Π_1 otherwise.

If one contemplates using cross model validation to select a subset of the available feature variables for use in a logistic classification function, the choice of inner criterion should receive attention. As in the case of ordinary multiple linear regression and discriminant analysis, a stepwise approach is a possibility. However, implementing a stepwise approach as inner criterion in logistic regression, entails replacing the F-test used in ordinary regression and discriminant analysis by a likelihood ratio chi-square test (cf. Hosmer and Lemeshow, 1989, p. 106-118). At each selection step of a forward selection procedure, the variable resulting in the largest increase in the likelihood ratio statistic when added to the variables already in the model, is selected. For backward elimination, the variable resulting in the smallest decrease in the likelihood ratio statistic, will be excluded at each step.

As alternative to a stepwise approach, an all possible subsets approach, using R^2 or C_p as criterion, can also be employed. As explained by Hosmer and Lemeshow (1989, p. 118-126), best subsets logistic regression can be performed using any program for best subsets linear regression in the following way.

Let $\tilde{\mathbf{X}} = [\mathbf{1}, \mathbf{X}]$ denote the $n \times (k+1)$ matrix, containing the observed values of the k feature variables, with the first column $\mathbf{1}$ representing the constant term in the logistic regression equation. Let $\hat{\tau}_i$ be the estimated posterior probability of the i -th case belonging to group Π_1 , i.e. $\hat{\tau}_i = e^{\hat{\beta}'_1 \tilde{\mathbf{x}}_i} / (1 + e^{\hat{\beta}'_1 \tilde{\mathbf{x}}_i})$, where $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ and $\tilde{\mathbf{x}}_i = [1, \mathbf{x}'_i]$.

Let \mathbf{P} be the $n \times n$ diagonal matrix with elements $\hat{\tau}_i(1 - \hat{\tau}_i)$, $i = 1, \dots, n$. Then

$$\hat{\beta} = (\tilde{\mathbf{X}}' \mathbf{P} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \mathbf{P} \mathbf{z}, \quad (4.6.2)$$

(cf. Pregibon, 1981), where $\mathbf{z} = \tilde{\mathbf{X}}'\hat{\boldsymbol{\beta}} + \mathbf{P}^{-1}(\mathbf{y} - \hat{\boldsymbol{\tau}})$, \mathbf{y} is the response vector of 0–1 entries indicating group membership and $\hat{\boldsymbol{\tau}}:n \times 1$ has elements $\hat{\tau}_i$, $i = 1, \dots, n$. It is clear from (4.6.2) that $\hat{\boldsymbol{\beta}}$ can be obtained from a weighted linear regression analysis using \mathbf{z} as dependent variable and the diagonal elements of \mathbf{P} as weights.

Using an all possible subsets approach as inner criterion in cross model validation therefore entails the following. For each omitted case, the logistic regression equation of $y_{(i)}$ on $\tilde{\mathbf{X}}_{(i)}$ is determined, using the data on all k variables. This equation is used to obtain the estimates $\hat{\tau}_i$, $i = 1, \dots, n$ needed to calculate \mathbf{P} and \mathbf{z} . An all possible subsets linear regression program is then used to identify the best model (according to a criterion such as R^2 or C_p) of each possible dimension $1, \dots, k$, and the logistic classification function based on each of these subsets is then calculated. The group membership of the omitted case is then predicted using the logistic classification function associated with each model dimension, and a measure of loss is calculated.

In the simulation study that was undertaken to evaluate the performance of cross model validation in logistic regression, the IMSL subroutine DRBEST was used for this purpose, with R^2 as criterion to find the best model of each dimension. As mentioned in Section 4.4, use of C_p in place of R^2 would give identical results, since only models of the same dimension are compared at this stage of the cross model validation process.

A number of different possibilities regarding the measure of loss to be used for each omitted case at each model dimension, were investigated. Each of these possibilities is now discussed.

1. The most natural choice is to use a 0-1 loss function, i.e. to take

$$H(y_i, \mathbf{X}_{(i)}; \mathbf{M}_p(J(\mathbf{X}_{(i)}))) = (y_i - \hat{y}_i)^2, \quad i = 1, \dots, n; \quad p = 1, \dots, k, \quad (4.6.3)$$

where $y_i \in \{0,1\}$ denotes the actual group membership of the i -th case, and \hat{y}_i is the predicted group membership. The disadvantages of the 0-1 loss function that were discussed in Section 4.4, are also relevant here. In particular, it may be impossible to identify a unique optimal model dimension, especially in the case of small samples. The 0-1 loss function was investigated in a preliminary simulation study. In cases where more than one value of p correspond to the minimum value of (4.3.2.1), the smallest such p -value was used as optimal model dimension, i.e. the most parsimonious choice was made. It was found that the resulting estimator $H^{\text{CMV}}(\mathbf{X}; \tilde{\mathbf{p}}(\mathbf{X}))$ of the post-selection actual error rate of the selected model, under-estimates this quantity and that the estimator generally also has a large variance, leading to unacceptably large UMSE's. As in the case of discriminant analysis, attention had to be focused on ways of smoothing the 0-1 loss function.

2. In discriminant analysis, a normally smoothed version of the 0-1 loss function performed well in terms of the UMSE of the corresponding estimator $H^{CMV}(\mathbf{X}; \tilde{p}(\mathbf{X}))$. This loss function is given in (4.4.1), and depends on the Anderson classification statistic, $W(\mathbf{x})$, and smoothing constants b_1 and b_2 . In logistic regression, the classification statistic $V(\mathbf{X})|(\mathbf{X} \in \Pi_i, t)$ is $N(\hat{\beta}_0 + \hat{\beta}'_1 \mu_i; \hat{\beta}'_1 \Sigma \hat{\beta}_1)$ distributed, $i = 0, 1$, provided that the data arise from normal populations. The conditional probability of misclassifying an entity from Π_0 , given the training data, is therefore

$$P[V(\mathbf{X}) > 0 | \mathbf{X} \in \Pi_0, t] = \Phi\left(\frac{\hat{\beta}_0 + \hat{\beta}'_1 \mu_0}{\hat{\beta}'_1 \Sigma \hat{\beta}_1}\right), \quad (4.6.4)$$

and that for an entity from Π_1 ,

$$P[V(\mathbf{X}) \leq 0 | \mathbf{X} \in \Pi_1, t] = \Phi\left(-\frac{\hat{\beta}_0 + \hat{\beta}'_1 \mu_1}{\hat{\beta}'_1 \Sigma \hat{\beta}_1}\right). \quad (4.6.5)$$

These probabilities depend on the unknown quantities μ_0 , μ_1 and Σ , and cannot be calculated. A possibility that suggests itself is to replace the unknown parameters in (4.6.4) and (4.6.5) by unbiased estimates, thereby obtaining estimates of the conditional probabilities of misclassification. If \mathbf{X} is used as an unbiased estimator of its own expectation, and the pooled sample covariance matrix \mathbf{S} is used to estimate Σ , the cross model validation criterion defined in (4.3.2.1) becomes

$$H^{CMV}(\mathbf{X}; p) = \frac{1}{n} \sum_{i=1}^n \sum_{j=0}^1 \left[\Phi\left((-1)^j \frac{\hat{\beta}_0 + \hat{\beta}'_1 \mathbf{x}_i}{\hat{\beta}'_1 \mathbf{S} \hat{\beta}_1}\right) I(y_i = j) \right]. \quad (4.6.6)$$

In the simulation study it was found that this approach generally did too much smoothing, causing the estimator $H^{CMV}(\mathbf{X}; \tilde{p}(\mathbf{X}))$ to over-estimate the post-selection actual error rate of the logistic classification rule based on the selected variables. It seems that smoothing constants similar to b_1 and b_2 in (4.4.1) are required in (4.6.4) and (4.6.5).

3. Another intuitively appealing option for the loss function in cross model validation is the posterior probability of wrong classification of the omitted case. For entities from Π_0 , these probabilities are given by

$$\hat{\tau}_1(\mathbf{x}_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}'_1 \mathbf{x}_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}'_1 \mathbf{x}_i}} \quad (4.6.7)$$

and for entities from Π_1 , by

$$\hat{\tau}_0(\mathbf{x}_i) = \frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{x}_i}} \quad (4.6.8)$$

If this approach is used,

$$H(\mathbf{X}; \mathbf{p}) = \frac{1}{n} \sum_{i=1}^n \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{x}_i) I(y_i=0)}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{x}_i}} \quad (4.6.9)$$

and the optimal model dimension $\tilde{p}(\mathbf{X})$ is chosen to minimise this quantity. Simulation experiments once more indicated that this approach did too much smoothing, and that the corresponding estimator $H^{\text{CMV}}(\mathbf{X}; \tilde{p}(\mathbf{X}))$ is conservatively biased in many parameter configurations.

From the empirical results for the cases discussed above, it seems that the loss function should be a combination of the 0-1 loss function and a smoothed version of this loss function. In addition, the transition between the 0-1 part of the loss function and its smoothed version should ideally depend on the separation between the two populations. This can be motivated as follows.

Consider the posterior probability, (4.6.7), of wrong classification of an entity from Π_0 . It would seem to be acceptable if a loss of zero is declared if this posterior probability becomes small enough, i.e. if its complement, the probability of correct classification, becomes larger than some cut-off point. This cut-off point should increase with the separation between the populations. If there is little or no separation between the populations, the mere fact that the posterior probability of correct classification exceeds 0.5 should be reason enough to declare a loss of zero. However, in cases where the populations are well separated, the posterior probability of correct classification should approach unity before a loss of zero is declared. The sample Mahalanobis distance, D , is a measure of the separation between the two populations, and the above considerations suggest the following method of loss calculation. For an entity from group Π_0 , calculate the posterior probability of misclassification, viz. $\hat{\tau}_1(\mathbf{x}_i) = e^{\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{x}_i} / (1 + e^{\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{x}_i})$, and take the loss for this omitted case equal to

$$\begin{cases} 0, & \text{if } \hat{\tau}_1(\mathbf{x}_i) < \min(\frac{1}{2}, 1/(1+D)) \\ 1, & \text{if } \hat{\tau}_1(\mathbf{x}_i) > \max(\frac{1}{2}, D/(1+D)) \\ \hat{\tau}_1(\mathbf{x}_i), & \text{if } \min(\frac{1}{2}, 1/(1+D)) \leq \hat{\tau}_1(\mathbf{x}_i) \leq \max(\frac{1}{2}, D/(1+D)). \end{cases}$$

(4.6.10)

Similar expressions hold for entities from Π_1 , with the posterior probability $\hat{\tau}_1(\mathbf{x}_i)$ replaced by $\hat{\tau}_0(\mathbf{x}_i) = 1 - \hat{\tau}_1(\mathbf{x}_i)$. Let A_1 be the subset of indices of $\{1, \dots, n_0\}$ for which the loss according to (4.6.10) is 1, and A_2 the subset for which the loss equals $\hat{\tau}_1(\mathbf{x}_i)$. Similarly, let B_1 and B_2 be these respective subsets for the cases in the training data set that come from Π_1 . Then,

$$H(\mathbf{X}; \mathbf{p}) = \frac{1}{n} \left\{ \#(A_1) + \#(B_1) + \sum_{i \in A_2} \hat{\tau}_1(\mathbf{x}_i) + \sum_{i \in B_2} \hat{\tau}_0(\mathbf{x}_i) \right\}. \quad (4.6.11)$$

Extensive simulation investigations indicated that this choice of loss function for the inner criterion leads to an estimator $H^{\text{CMV}}(\mathbf{X}; \tilde{\mathbf{p}}(\mathbf{X}))$ of the post-selection actual error rate that has good UMSE behaviour. This is the loss function for which results will be reported in Section 4.7.

Comparatively little has appeared in the literature on estimation of the error rate of a logistic classification rule based on a selected subset of variables. Notable exceptions are the papers by Gong (1986) and Efron and Gong (1983). These papers were briefly mentioned in Section 4.2. Since the bootstrap estimator discussed in these papers is compared in the simulation study of Section 4.7 with the CMV-estimator, this resampling procedure is now explained in greater detail.

Consider a given training data set, \mathbf{t} , and suppose a variable selection technique is applied to \mathbf{t} and a logistic classification rule is constructed based on the selected variables. How can the bootstrap be used to estimate the actual error rate of this classification rule? According to Efron and Gong (1983) and Gong (1986) it is essential to repeat the selection step on each bootstrap sample drawn from the given training data set. The following steps are recommended.

1. Calculate the post-selection apparent error rate when the classification rule based on the selected variables is applied to all cases in the training data set, \mathbf{t} . Call this apparent error rate ae_1 . It is well known that ae_1 is an optimistic estimate of the error rate of the rule being considered.
2. Generate a bootstrap sample, \mathbf{t}_1^* , from the training data set. Suppose the original training data set consists of n_0 cases from Π_0 and n_1 cases from Π_1 . Then \mathbf{t}_1^* must also have n_0 and n_1 cases from Π_0 and Π_1 respectively. Hence, \mathbf{t}_1^* is obtained by selecting n_0 cases randomly and with replacement from the n_0 cases from Π_0 in \mathbf{t} , and similarly for n_1 cases from the n_1 cases from Π_1 in \mathbf{t} .
3. Perform the variable selection step on \mathbf{t}_1^* , obtaining a bootstrap classification rule.

4. Apply the bootstrap classification rule to t and to t_1^* , obtaining the apparent error rates ae_2 and ae_3 , respectively.

5. Repeat steps 2-4 a large number of times, say B times. Calculate $\frac{1}{B} \sum_{i=1}^B (ae_{2i} - ae_{3i})$. This is an estimate of the optimism inherent in the apparent error rate when it is used to estimate the actual error rate of the classification rule.

6. The bootstrap estimate of the post-selection error rate is given by $ae_1 + \frac{1}{B} \sum_{i=1}^B (ae_{2i} - ae_{3i})$. Clearly, the bootstrap is used to estimate a bias correction factor that is used to improve the ordinary apparent error rate.

An essential part of the above process is that the variable selection step must be carried out anew for each bootstrap sample, as indicated in step 3. According to Gong (1986) the bootstrap bias correction method has little merit if only the variables that are originally selected from t are repeatedly applied to each bootstrap sample. This is in line with the principle that procedures in the "bootstrap world" should mimic as closely as possible those in the "real world" (cf. Efron and Tibshirani, 1993).

In Section 4.7 the bootstrap estimate of post-selection actual error rate will be compared to the CMV-estimator (4.6.11).

4.7 MONTE CARLO SIMULATION STUDY FOR LOGISTIC REGRESSION

A Monte Carlo simulation study was undertaken to compare the performance of cross model validation to that of the bootstrap procedure described in Section 4.6. The methods were evaluated for populations with different underlying distributions: the normal distribution, the double exponential distribution and the lognormal distribution. The covariance structure, $\Sigma = I$ was used for all the distributions. For the total number of available feature variables, the value $k = 10$ was used throughout. It is assumed that the feature vector X has mean vector $\mu_0 = 0$ in Π_0 , and that the first r elements of μ_1 , the mean vector of X in Π_1 , differ from zero. The values $r = 1, 5$ and 10 were used. For $r = 1$ and 5 , the elements of μ_1 were chosen as in (4.5.1), and for $r = 10$, as in (4.5.2).

In each case, only sample sizes $n_0 = n_1 = 50$ were considered. The codes N1, N2 and N3 are used to denote the normal cases with $r = 1$, $r = 5$ and $r = 10$, in that order. For the double exponential cases D1, D2 and D3 are used similarly, with L1, L2 and L3 being used for the lognormal cases. In the normal and double exponential cases, the performance of the techniques were evaluated at the following values of Δ^2 :

0,1,2,3 and 4. In the lognormal case, $\Delta^2 = 0, 0.5, 1, 1.5$ and 2 were used, because of the problem of non-existence of maximum likelihood estimates of the logistic regression coefficients when the populations are well separated.

The procedures are evaluated with respect to the accuracy with which they estimate the post-selection actual error rate. For this purpose, the bias and unconditional mean squared error (UMSE) of the error rate estimators are compared. Program 3 in the Appendix is an example of the Fortran program used in this part of the simulation study.

4.7.1 THE NORMAL CASE

In the normal case, a simulation study was performed to compare the estimation performance of the cross model validation error rate estimator to that of the bootstrap error rate estimator. To estimate the quantities used in the comparison, 200 Monte Carlo repetitions were used. For each repetition, a training data set was generated from the relevant normal distributions. The cross model validation procedure for logistic regression which was described in Section 4.6, was used to identify an optimal subset of the available feature variables, and to estimate the post-selection actual error rate associated with the logistic discriminant rule based on these variables. An all possible subsets selection procedure using C_p as criterion, was also used to identify an optimal subset, and the bootstrap method described in Section 4.6 was used to estimate the post-selection actual error rate associated with the logistic discriminant function based on this subset. In both cases, the actual error rates associated with the selected subsets were obtained by means of Monte Carlo simulation. To do this, a large number (500) of data cases were generated independently from the training data, and classified using the logistic discriminant rule based on each of the selected subsets of feature variables. With a view to estimating the bias and unconditional mean squared error of each of the error rate estimators, the difference and squared difference between the value of each error rate estimator and the corresponding post-selection actual error rate, were also calculated. To obtain the expected post-selection actual error rates, the 200 actual error rates obtained for each technique, were averaged. The bias associated with each technique was estimated by averaging the differences between the value of each error rate estimator and post-selection actual error rate over the 200 repetitions, i.e. $\hat{B}_j = \frac{1}{200} \sum_{i=1}^{200} (\hat{\alpha}_{ij} - \alpha_{ij}^{act})$, where $\hat{\alpha}_{ij}$ denotes a value of an error rate estimator obtained by means of technique j , $j = 1, 2$ for the i -th Monte Carlo repetition and α_{ij}^{act} denotes the actual error rate calculated for technique j for the i -th Monte Carlo repetition. The estimated unconditional mean squared error of each error rate estimator was obtained by averaging the squared differences between the relevant error rate estimator and the corresponding post-selection actual error rate, i.e.

$$\hat{U}_j = \frac{1}{200} \sum_{i=1}^{200} (\hat{\alpha}_{ij} - \alpha_{ij}^{act})^2.$$

If a data set was generated for which the maximum likelihood estimates of the logistic regression coefficients did not exist, the case was excluded from further analyses, and a new data set was generated, to obtain a total of 200 valid repetitions. The results of the simulation study were summarised by means of graphs, given in Figs. 4.20 - 4.22.

4.7.1.1 Expected Actual Error Rate

The expected actual error rate associated with the logistic discriminant function based on the variables selected by means of the cross model validation technique, as well as that associated with the subset selected by an all possible subsets approach based on the C_p -criterion, are displayed in graphs given in Fig. 4.20. It is clear that the classification performance of the rules based on these subsets, is virtually identical. Only in case N3 is there a slight difference between the expected actual error rates, the logistic discriminant rule based on variables selected by means of the cross model validation technique, yielding a lower expected actual error rate than the other procedure in this case.

4.7.1.2 Bias

Graphs of the bias of the cross model validation based error rate estimator and that of the bootstrap estimator, are given in Fig. 4.21. In all cases, the bootstrap estimator is considerably less biased than the CMV-estimator at small to moderate values of Δ^2 ($\Delta^2 \leq 2$), but the opposite is true at larger values of Δ^2 ($\Delta^2 > 2$).

4.7.1.3 Unconditional Mean Squared Error (UMSE)

In Fig. 4.22, graphs of the unconditional mean squared errors of the CMV-estimator and the bootstrap estimator are given. In case N1, the UMSE of the CMV-estimator is considerably less than that of the bootstrap estimator, except at $\Delta^2 = 0$. In case N2, the bootstrap estimator has lower UMSE at small values of Δ^2 ($\Delta^2 < 2$), while the CMV-estimator performs better at large values of Δ^2 ($\Delta^2 \geq 2$). In case N3, the bootstrap estimator outperforms the CMV-estimator at all Δ^2 -values.

In general, for normal data, neither of the two methods outperforms the other consistently. The bootstrap method performs better for populations that are not well separated, but is outperformed by the CMV - method at larger separations.

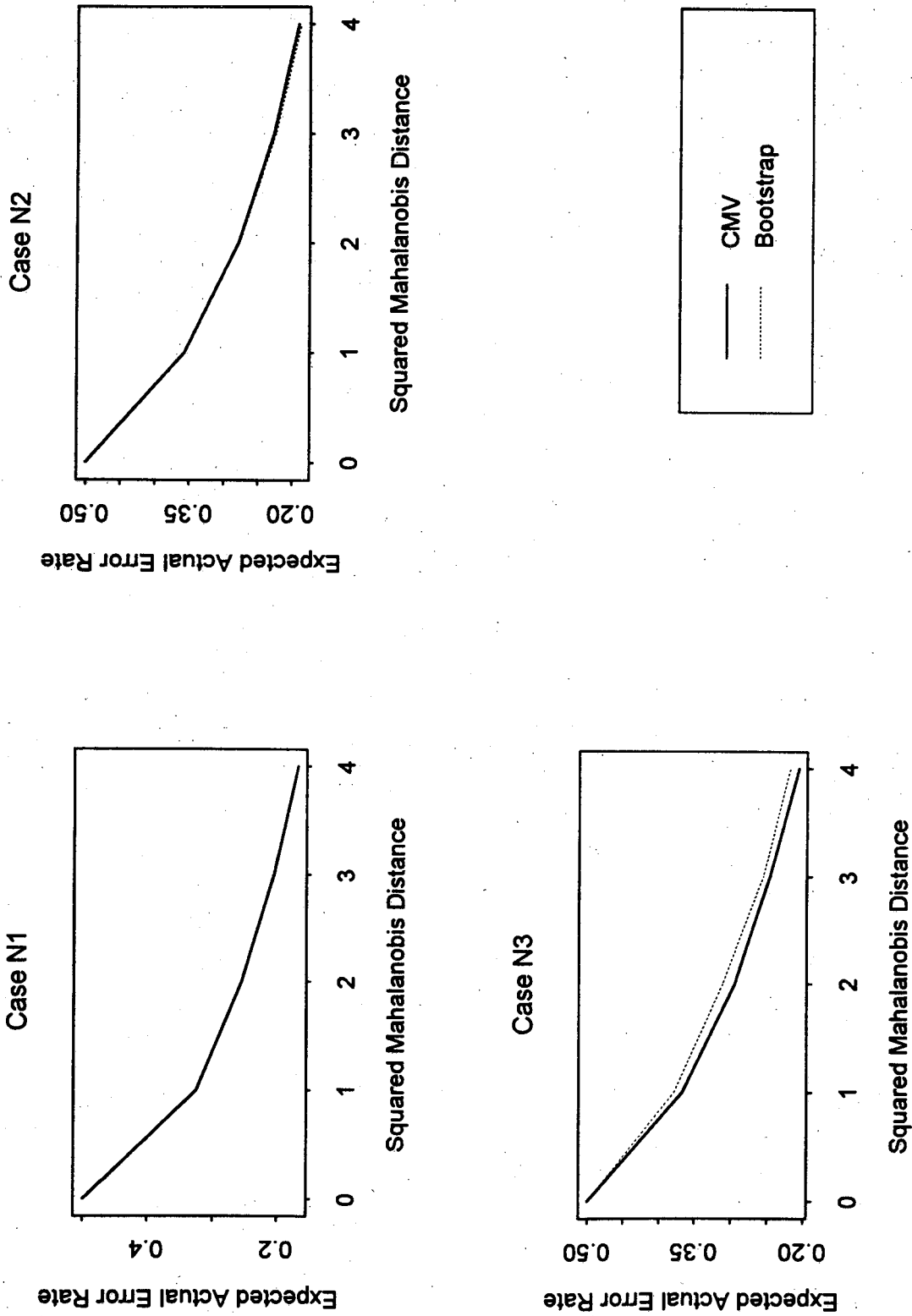


FIG. 4.20: EXPECTED ACTUAL ERROR RATE, NORMAL DATA

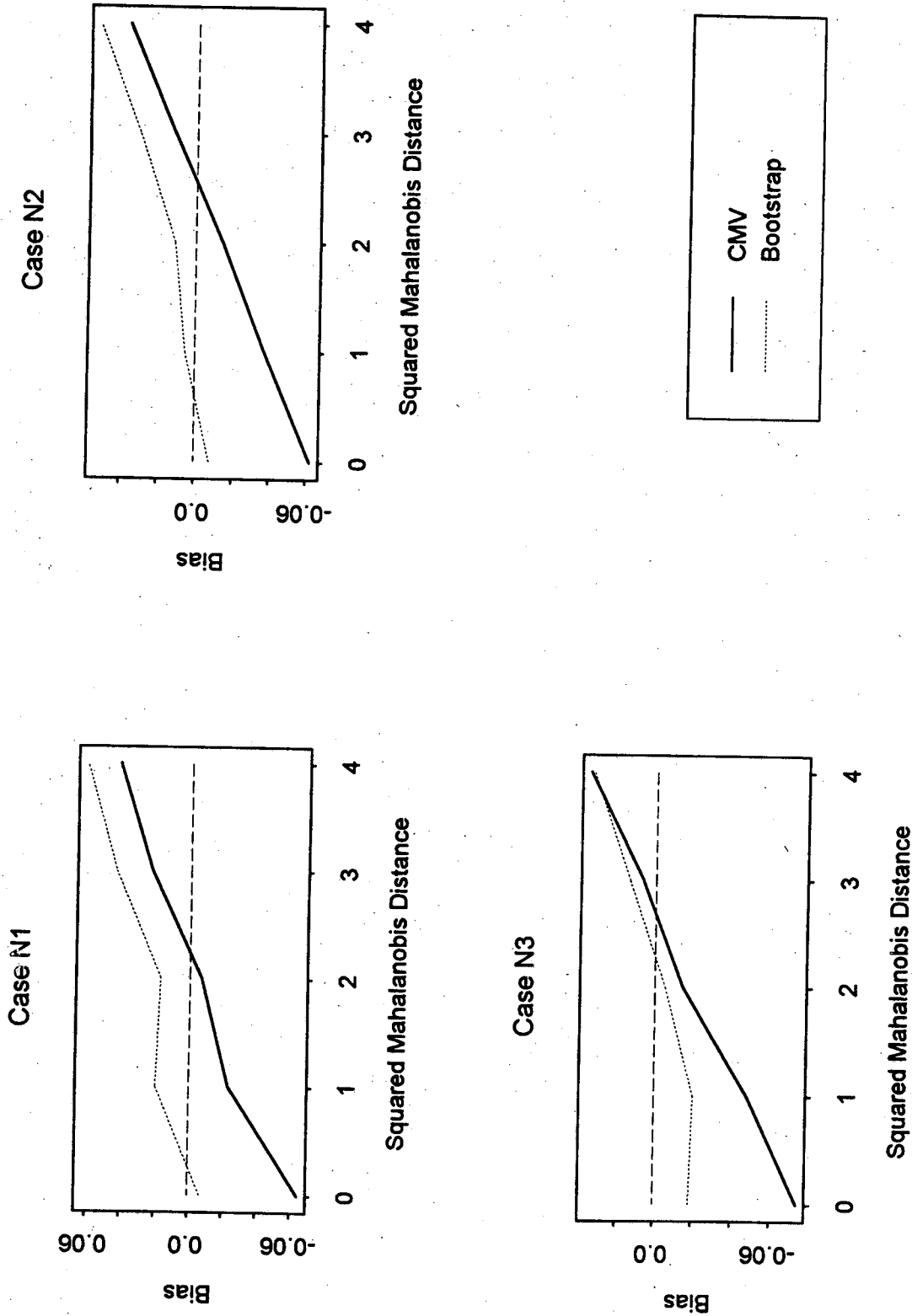


FIG. 4.21: BIAS OF ERROR RATE ESTIMATORS, NORMAL DATA

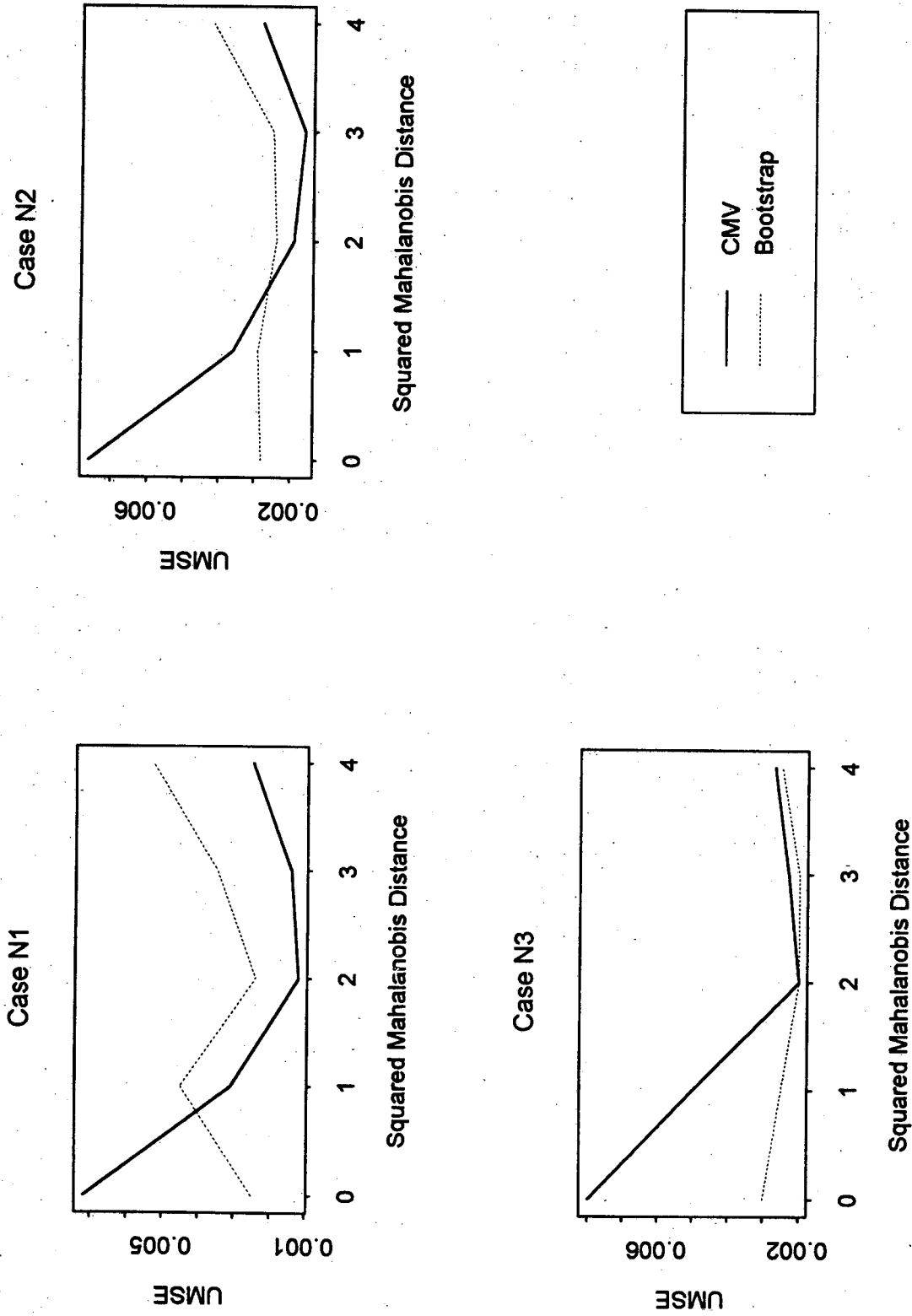


FIG. 4.22: UNCONDITIONAL MEAN SQUARED ERRORS OF ERROR RATE ESTIMATORS, NORMAL DATA

4.7.2 THE DOUBLE EXPONENTIAL CASE

The graphs displaying the results of the simulation study for the double exponential cases, are given in Figs. 4.23 - 4.25. These results are now discussed.

4.7.2.1 Expected Actual Error Rate

The differences in the expected actual error rates associated with the logistic discriminant function based on the variables selected by means of the two different selection procedures, are very small (see Fig. 4.23). The classification performance of the rules based on the different subsets, are therefore virtually identical.

4.7.2.2 Bias

Perusal of the graphs in Fig. 4.24 shows that the behaviour of the bias in the double exponential cases is largely the same as in the normal cases, discussed in Section 4.7.1. The bootstrap estimator is less biased at small values of Δ^2 ($\Delta^2 < 2$), while the CMV-estimator performs better with respect to bias at larger values of Δ^2 ($\Delta^2 \geq 2$).

4.7.2.3 Unconditional Mean Squared Error (UMSE)

In the double exponential cases, the UMSE of the CMV-estimator is less than that of the bootstrap estimator (except at $\Delta^2 = 0$) in case D1. In case D2, the bootstrap estimator has lower UMSE at small values of Δ^2 ($\Delta^2 < 2$), but the CMV-estimator performs better in terms of UMSE at larger Δ^2 - values ($\Delta^2 \geq 2$). In case D3, the bootstrap estimator outperforms the CMV-estimator at all values of Δ^2 . These conclusions follow from the graphs in Fig. 4.25.

As in the normal case, neither of the two estimators seems to be better than the other in all cases. The relative performance of the techniques is influenced by the data configuration and by the separation between the two populations.

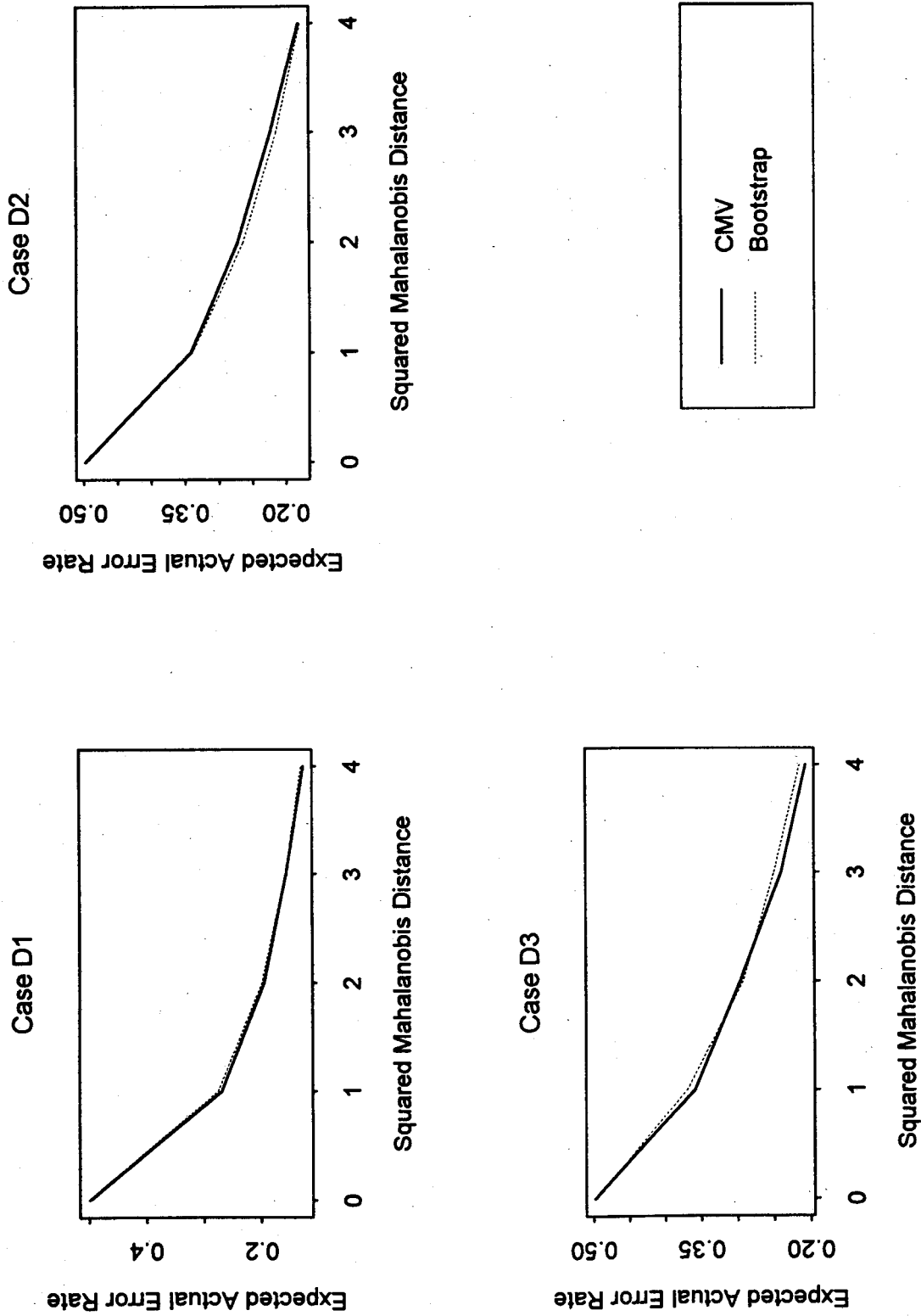


FIG. 4.23: EXPECTED ACTUAL ERROR RATE OF ERROR RATE ESTIMATORS, DOUBLE EXPONENTIAL DATA

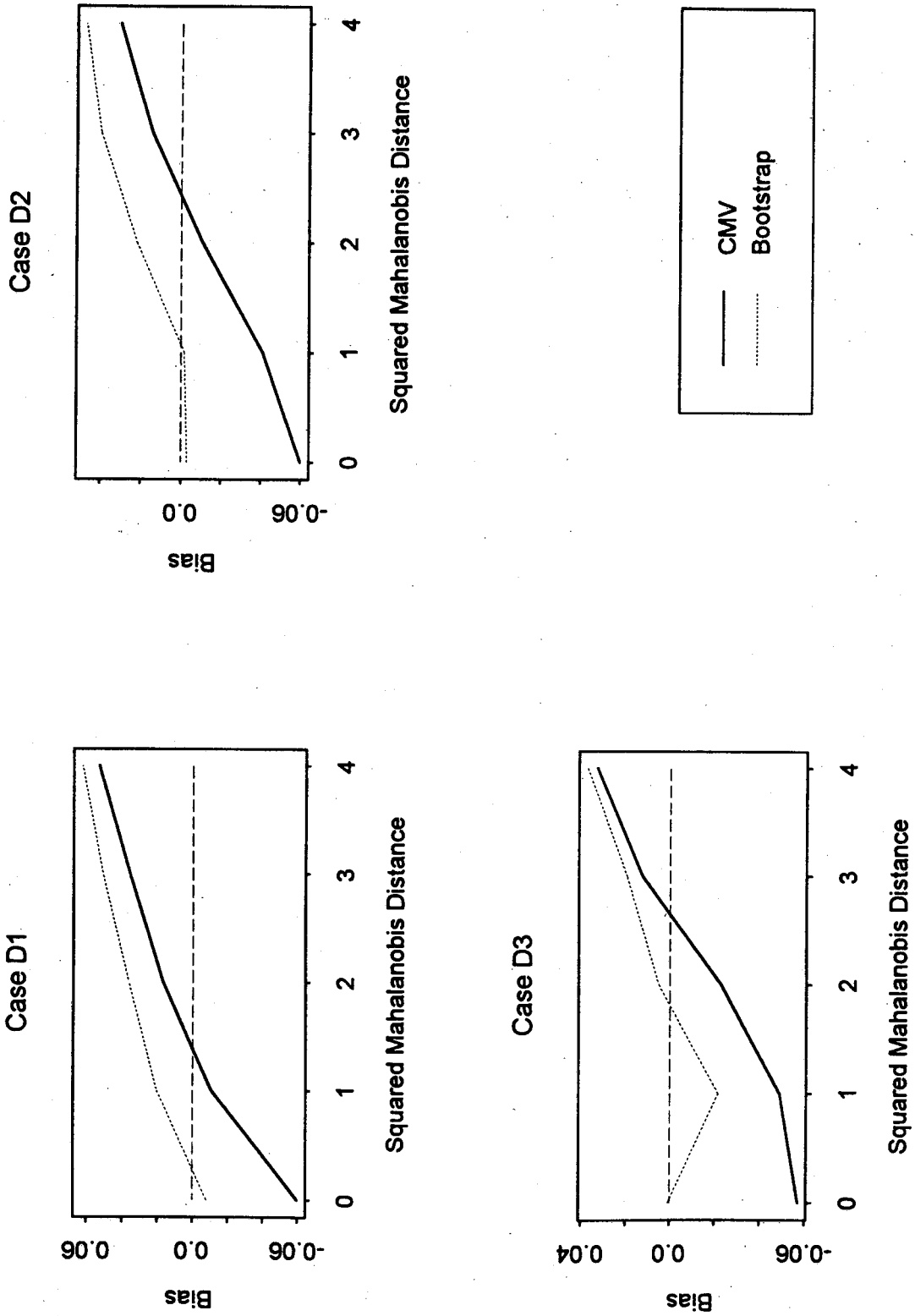


FIG. 4.24: BIAS OF ERROR RATE ESTIMATORS, DOUBLE EXPONENTIAL DATA

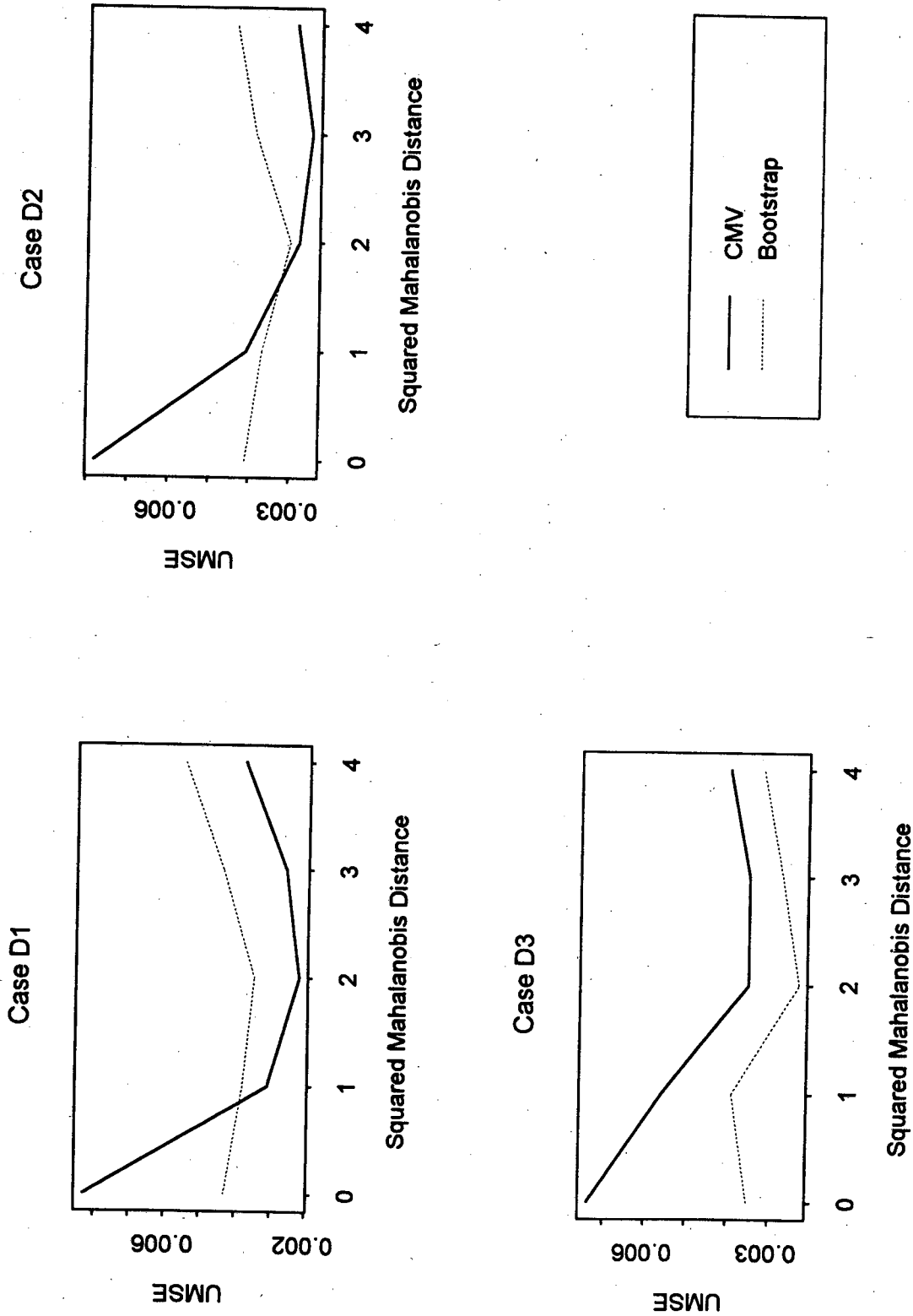


FIG. 4.25: UNCONDITIONAL MEAN SQUARED ERRORS OF ERROR RATE ESTIMATORS, DOUBLE EXPONENTIAL DATA

4.7.3 THE LOGNORMAL CASE

For the lognormal case, Figs. 4.26 - 4.28 contain the graphs of the simulation output.

4.7.3.1 Expected Actual Error Rate

The differences between the expected actual error rates (displayed in Fig. 4.26) associated with the logistic discriminant functions based on the subsets selected by means of the two methods considered, are larger in the lognormal case than in the normal and double exponential cases. In cases L2 and L3, the expected actual error rate of the logistic discriminant function based on variables selected by means of the CMV- technique, is lower than that of the other procedure. Using this function will therefore lead to slightly better classification.

4.7.3.2 Bias

From the graphs in Fig. 4.27 it is clear that the bias of the bootstrap estimator is lower than that of the CMV-estimator at small values of Δ^2 ($\Delta^2 < 1$), but at moderate to large Δ^2 -values ($\Delta^2 \geq 1$), the opposite is true.

4.7.3.3 Unconditional Mean Squared Error (UMSE)

The UMSE values attained by the error rate estimators (see Fig. 4.28), display similar behaviour in the lognormal cases than in the normal and double exponential cases. The most important difference is in case L1, where the UMSE of the CMV-estimator is larger than that of the bootstrap estimator at $\Delta^2 = 2$. In cases L2 and L3 the relative performance of the two techniques with respect to UMSE is similar to the corresponding normal and double exponential cases.

Once more, neither of the two techniques can be recommended in preference to the other, since the relative performance is again dependent on the separation between the two populations, as well as on the specific data configuration considered.

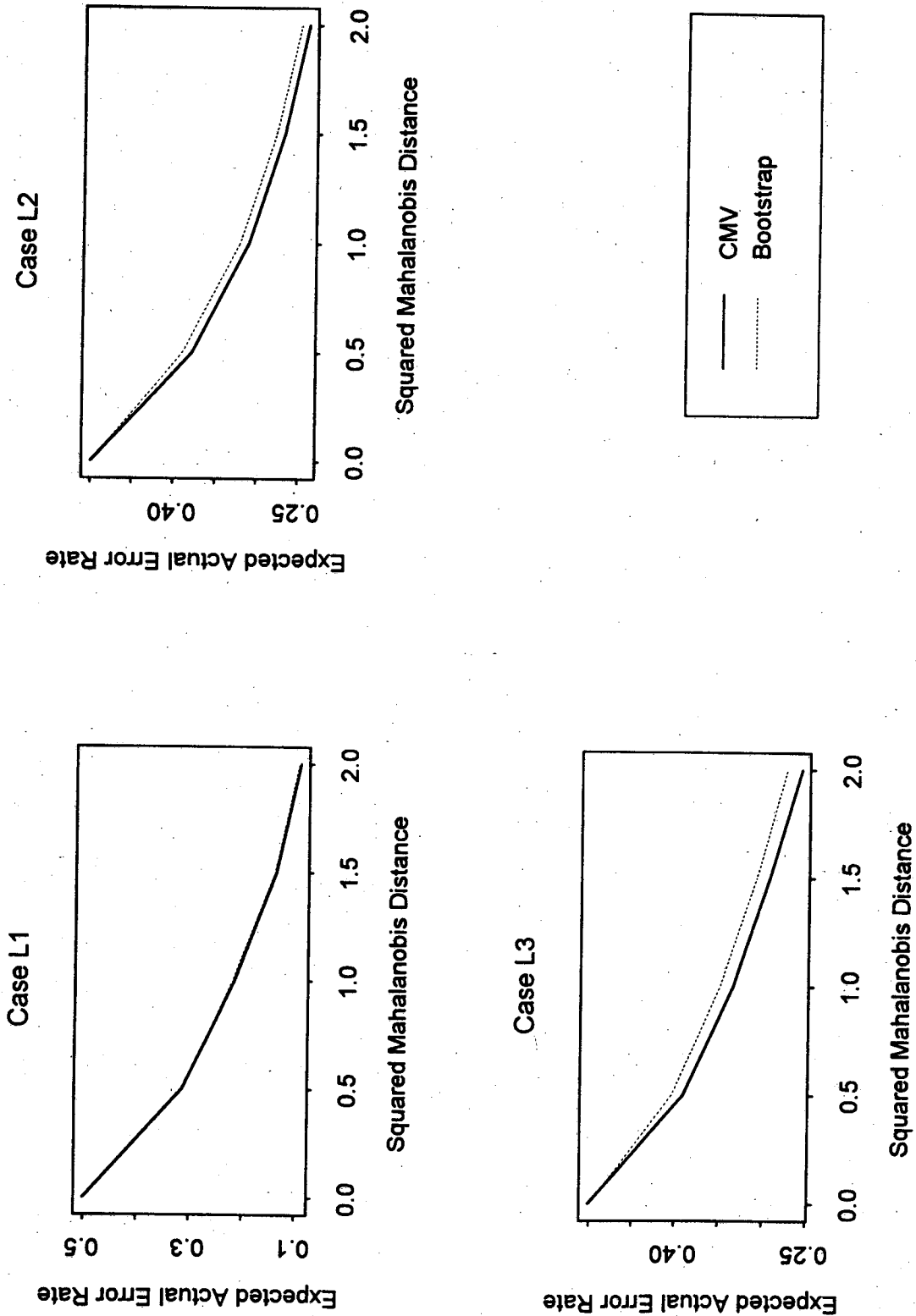


FIG. 4.26: EXPECTED ACTUAL ERROR RATE, LOGNORMAL DATA

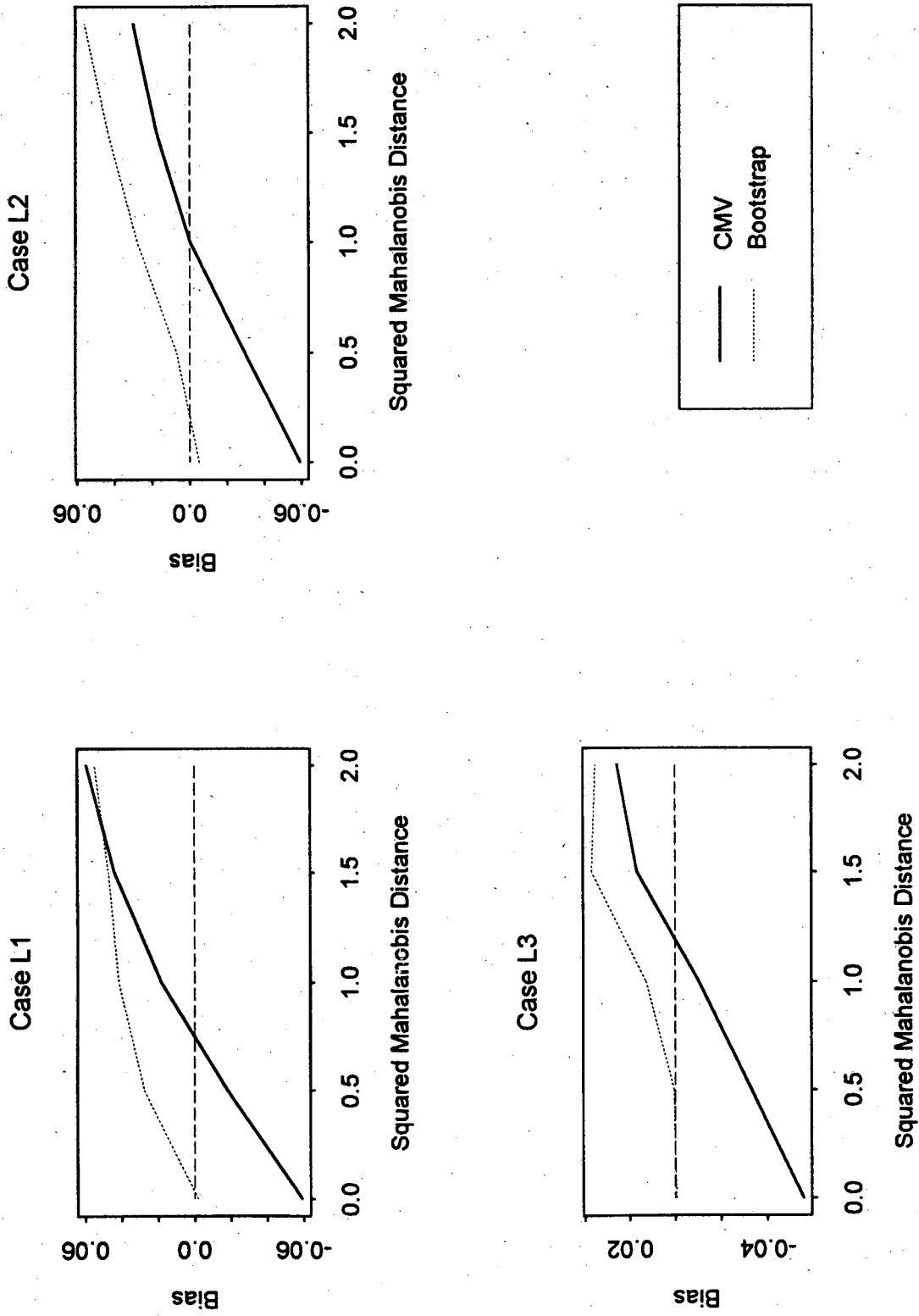


FIG. 4.27: BIAS OF ERROR RATE ESTIMATORS, LOGNORMAL DATA

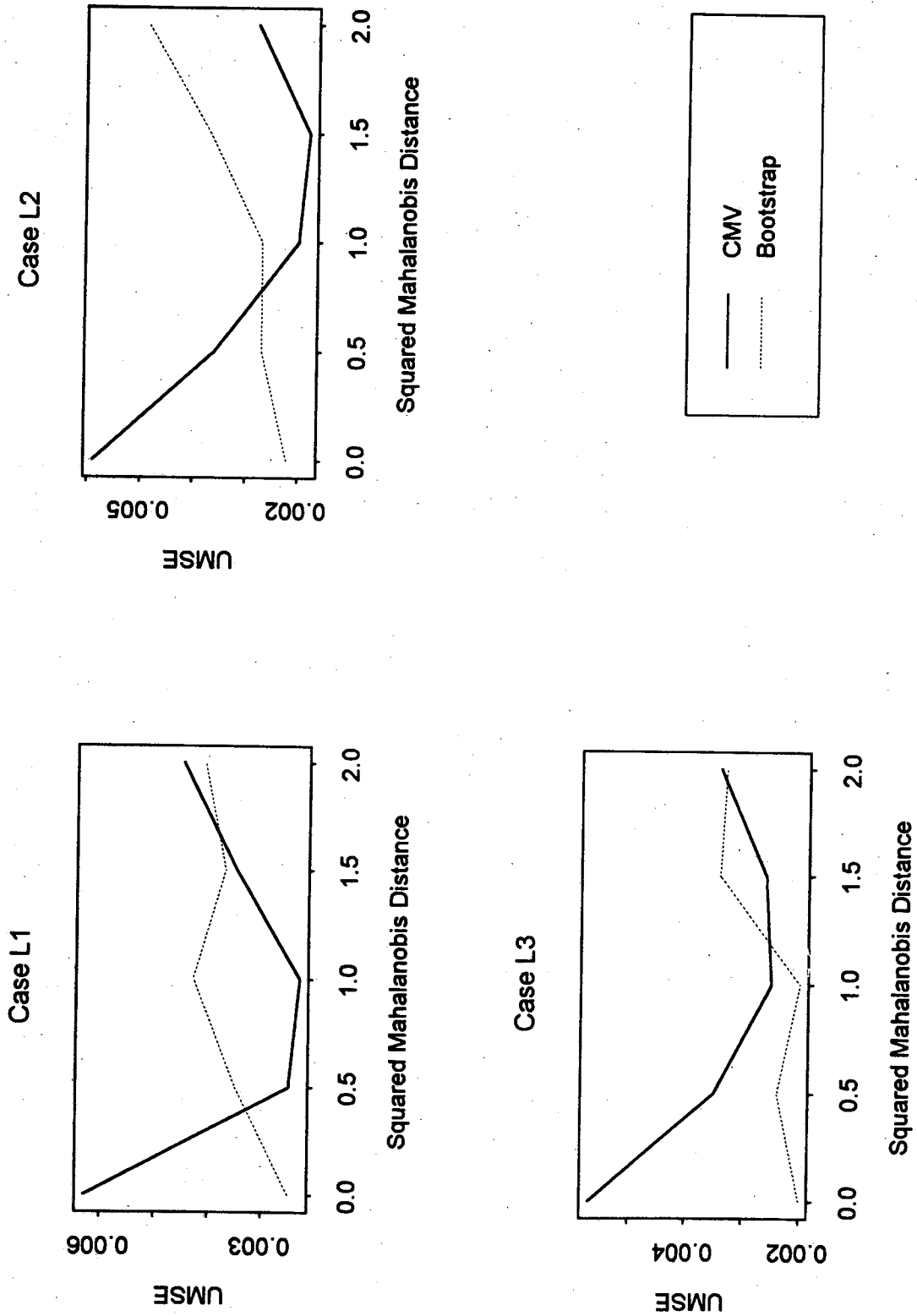


FIG. 4.28: UNCONDITIONAL MEAN SQUARED ERRORS OF ERROR RATE ESTIMATORS, LOGNORMAL DATA

4.8 COMPARISON OF THE PERFORMANCE OF CROSS MODEL VALIDATION IN DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

The cross model validation technique can be applied for variable selection and error rate estimation in both discriminant analysis and logistic regression. In Section 4.4 the cross model validation technique was applied in a discriminant analysis context, and subsequently its performance was evaluated by means of a simulation study in which it was compared to two other procedures for variable selection and error rate estimation in discriminant analysis (cf. Section 4.5). Application of the cross model validation technique in a logistic regression context, was discussed in Section 4.6, followed in Section 4.7 by a discussion of a simulation study in which the performance of the proposed cross model validation technique was compared to another procedure for variable selection and error rate estimation in logistic regression. In both these simulation studies, the cross model validation procedure was found to perform very well relative to the other methods considered, not only in selecting the seemingly relevant variables and forming a classification rule having a lower error expected actual error rate than that of the rules selected by the other methods, but also in estimating the resulting post-selection error rate accurately. An important issue that also needs to receive attention, is the relative performance of the cross model validation technique in discriminant analysis and in logistic regression. In this section, the selection performance of the cross model validation technique in discriminant analysis will be compared to its performance in logistic regression. In this comparison, the probabilities with which seemingly relevant and seemingly irrelevant variables are selected, will be considered. The following probabilities will be compared:

- the probability of correct selection (PCS), defined as the probability of selecting all the seemingly relevant variables and no seemingly irrelevant variables;
- the probability of over-selection (POS), defined as the probability of selecting all the seemingly relevant and some seemingly irrelevant variables;
- the probability of under-selection (PUS), defined as the probability of selecting only a subset of the seemingly relevant variables and no seemingly irrelevant variables;
- the probability of mixed selection (PMS), defined as the probability of selecting a subset (but not all) of the seemingly relevant variables, plus some seemingly irrelevant variables.

The classification performance of the linear discriminant rule and the logistic discriminant rule based on the variables selected by means of the cross model validation technique, will also be compared. This comparison will take place in terms of the post-selection expected actual error rates associated with the two discriminant rules.

O’Gorman and Woolson (1991) compared the selection performance of stepwise discriminant analysis to that of stepwise logistic regression by means of an extensive

Monte Carlo simulation study. They considered normal, lognormal and Bernoulli feature variables, and also mixtures of these variables. They used eight feature variables, of which four were seemingly relevant (i.e. the means differed between the two groups) and four seemingly irrelevant, and considered sample sizes $n_0 = n_1 = 50, 100, 200, 400$ as well as $n_0 = 50; n_1 = 200$. They calculated the four probabilities mentioned above (the PCS, POS, PUS and PMS) and compared these probabilities for variable selection by means of a fully stepwise selection procedure for discriminant analysis and logistic regression. They concluded that the differences in these probabilities for the two techniques are very small for sample sizes of 100 and larger, but that the probability of correct selection was higher for stepwise discriminant analysis than for stepwise logistic regression in cases where the sample sizes were small (i.e. $n_0 = n_1 = 50$). O'Gorman and Woolson (1991) did not compare the classification performance of the linear discriminant rule and the logistic discriminant rule based on the selected subsets, but concentrated on the selection performance in terms of the probabilities defined above.

In this section, the selection performance of the cross model validation technique for discriminant analysis and that for logistic regression are compared by considering the probabilities defined above. The post-selection actual error rates are also compared to evaluate the classification performance of the resulting linear and logistic discriminant functions.

The comparison was done for populations with different underlying distributions: the normal distribution, the double exponential distribution and the lognormal distribution. The covariance structure $\Sigma = I$ was used for all the distributions. For the total number of available feature variables, the value $k = 10$ was used throughout. It was assumed that the feature vector X has mean vector $\mu_0 = 0$ in Π_0 , and that the first r elements of μ_1 , the mean vector of X in Π_1 , differ from zero. The values $r = 1, 5$ and 10 were used. For $r = 1$ and 5 , the elements of μ_1 were chosen as in (4.5.1), and for $r = 10$, as in (4.5.2). In each case, sample sizes $n_0 = n_1 = 50$ were considered. The codes N1, N2 and N3 are used for the normal cases with $r = 1$, $r = 5$ and $r = 10$, in that order. For the double exponential cases D1, D2 and D3 are used similarly, while L1, L2 and L3 are used for the lognormal cases.

4.8.1 SELECTION PERFORMANCE

In Fig. 4.29 graphs of the probability of correct selection (PCS), the probability of over-selection (POS), the probability of under-selection (PUS) and the probability of mixed selection (PMS) for one of the normal cases, case N2, are given. Fig. 4.30 contain similar graphs for case D2, while graphs for case L2 appear in Fig. 4.31. These are the cases for which $r = 5$, i.e. there are 5 feature variables with respect to which the means of the two populations differ and 5 feature variables for which the two populations have identical means.

Perusal of the graphs displayed in Figs. 4.29 - 4.31, lead to the following conclusions.

1. For normal data, the PCS of the two procedures is nearly identical, the logistic regression procedure having a slightly lower PCS at some values of Δ^2 . For double exponential and lognormal variables, the difference in PCS is larger and increases with Δ^2 . This is similar to the findings of O'Gorman and Woolson (1991) for stepwise selection in discriminant analysis and logistic regression for sample sizes of 50.
2. For data from all three distributions considered, the discriminant analysis procedure had lower PUS and higher POS than the logistic regression procedure. This is an indication that the logistic regression procedure tended to select less variables than the discriminant analysis procedure. Using the logistic regression cross model validation selection procedure will therefore generally lead to a more parsimonious model.
3. In all cases considered, differences between the PMS of the two procedures are small.

The conclusion made in the second point above, is further illustrated by considering the cumulative distribution of the number of variables selected by each of the two techniques. Examples of such graphs, for case N2 at different values of Δ^2 , are given in Fig. 4.32. From these graphs (and similar graphs for other cases that are not shown here), it is clear that the logistic regression procedure tends to select less variables than the discriminant analysis procedure.

4.8.2 CLASSIFICATION PERFORMANCE

Graphs displaying the expected actual error rates associated with the linear discriminant function and the logistic discriminant function based on the subsets selected by means of the cross model validation procedure for each technique, are given in Figs. 4.33 - 4.35. In the normal cases (see Fig. 4.33) the differences in the post-selection error rates are very small. Only in cases N2 and N3, the linear discriminant rule yields a slightly lower post-selection error rate than the logistic discriminant rule at large values of Δ^2 ($\Delta^2 \geq 2$). For the double exponential cases, the differences in the post-selection error rates are slightly larger. Once more, the post-selection error rate associated with the linear discriminant rule is lower at large values of Δ^2 than that attained by the logistic discriminant rule for cases D2 and cases D3. For lognormal data, the same is true for cases L2 and L3, but for case L1, the logistic discriminant rule yields a slightly lower error rate at $\Delta^2 \geq 1$. However, in this case the differences are very small. In summary, if correct classification of new cases is the main concern, using the linear discriminant function based on variables selected by means of cross model validation, may be preferable. If it is of importance to select a parsimonious rule, the logistic discriminant function may be a better option, and the price paid in terms of correct classification of new cases, will be very small.

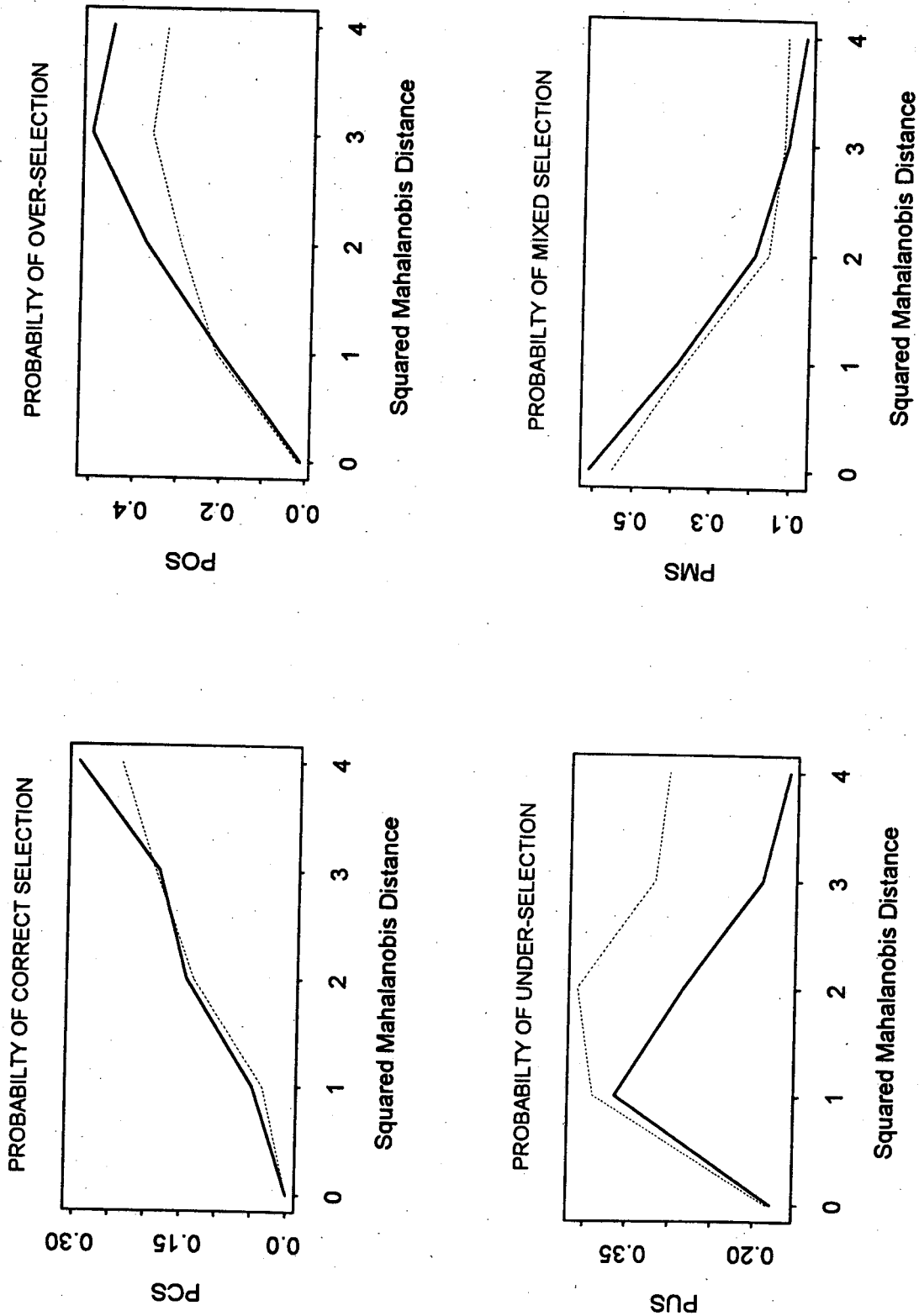


FIG. 4.29: COMPARISON OF SELECTION PERFORMANCE OF DA AND LR, NORMAL DATA, CASE N2

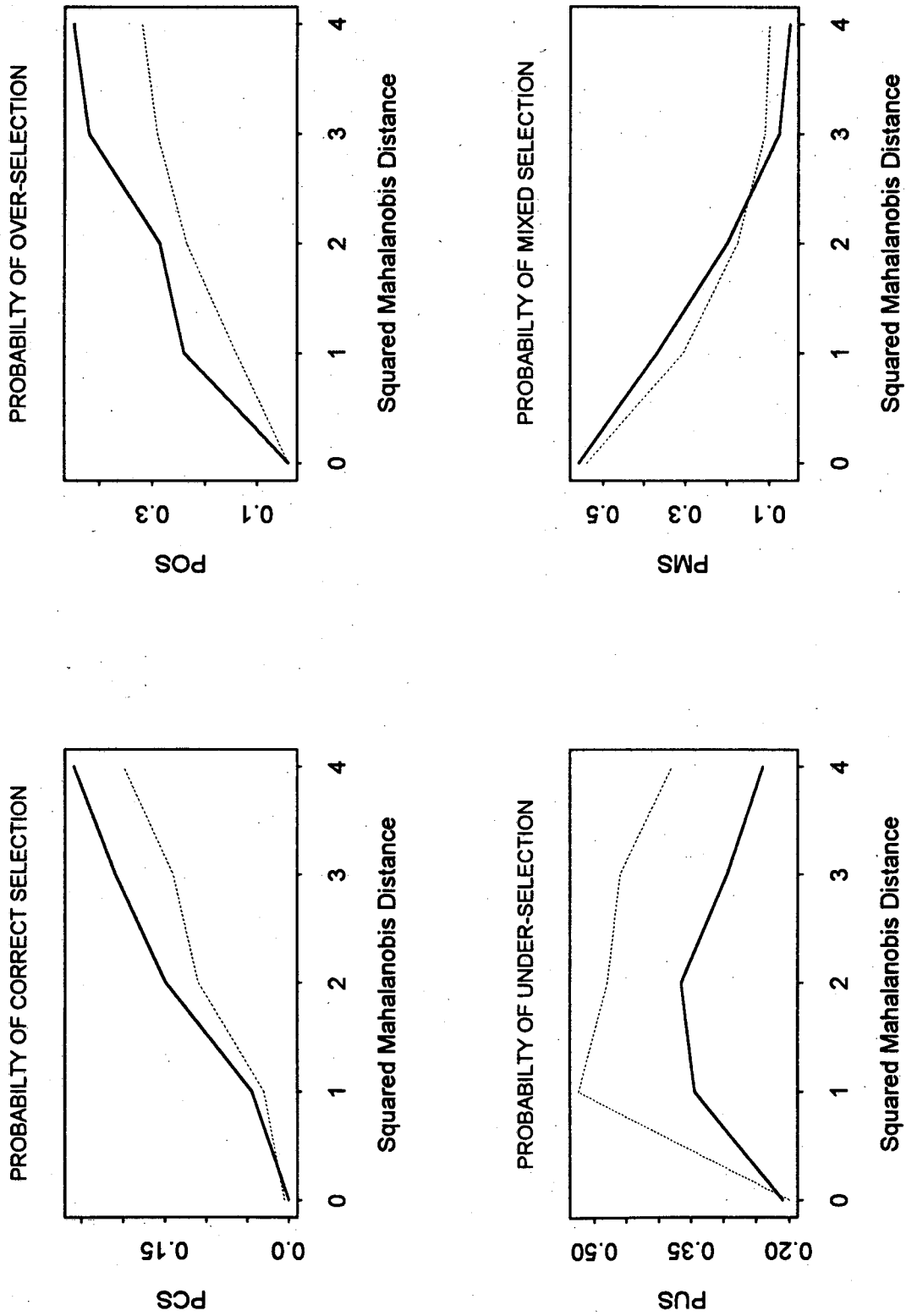


FIG. 4.30: COMPARISON OF SELECTION PERFORMANCE OF DA AND LR, DOUBLE EXPONENTIAL DATA, CASE D2

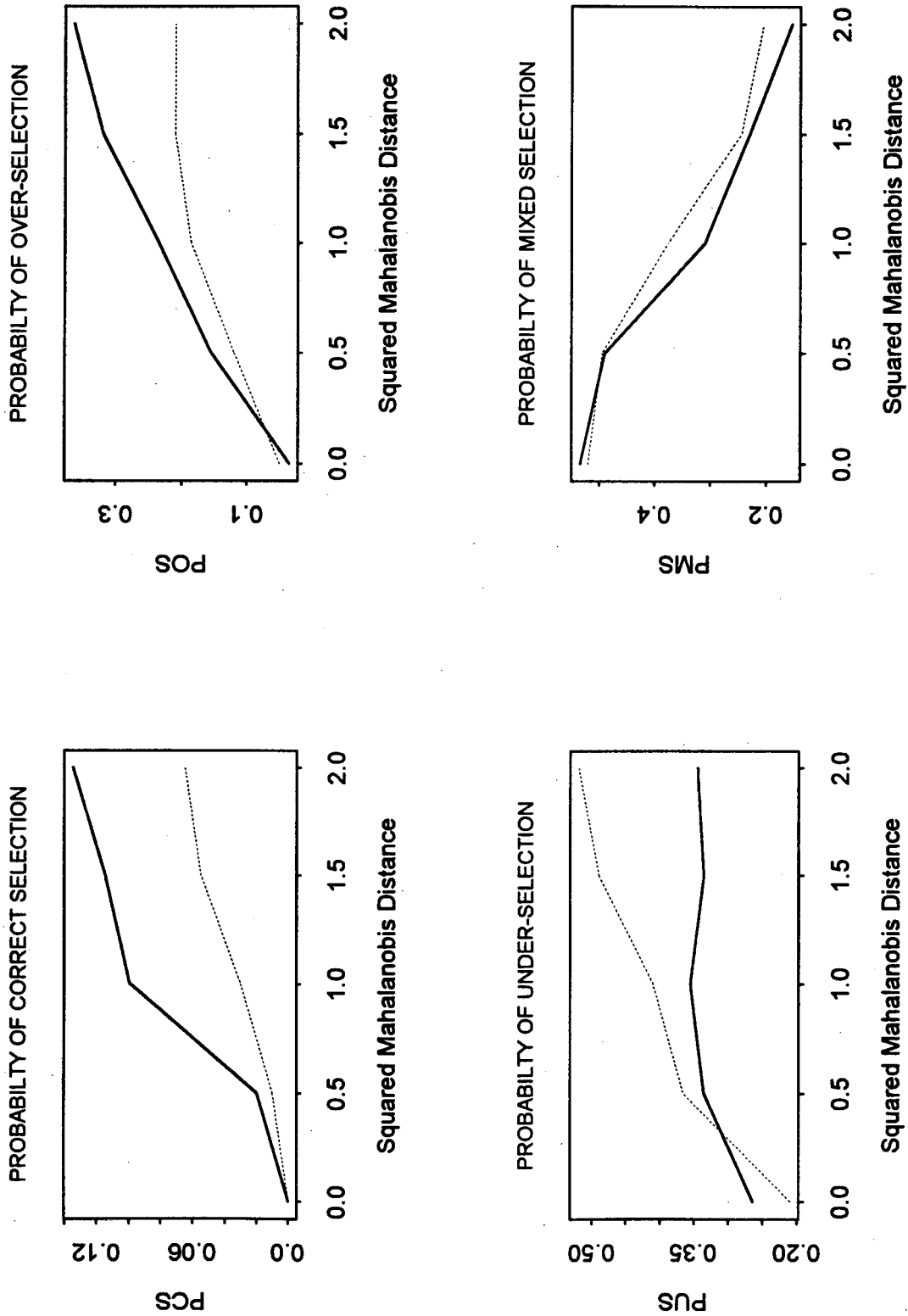


FIG. 4.31: COMPARISON OF SELECTION PERFORMANCE OF DA AND LR, LOGNORMAL DATA, CASE L2

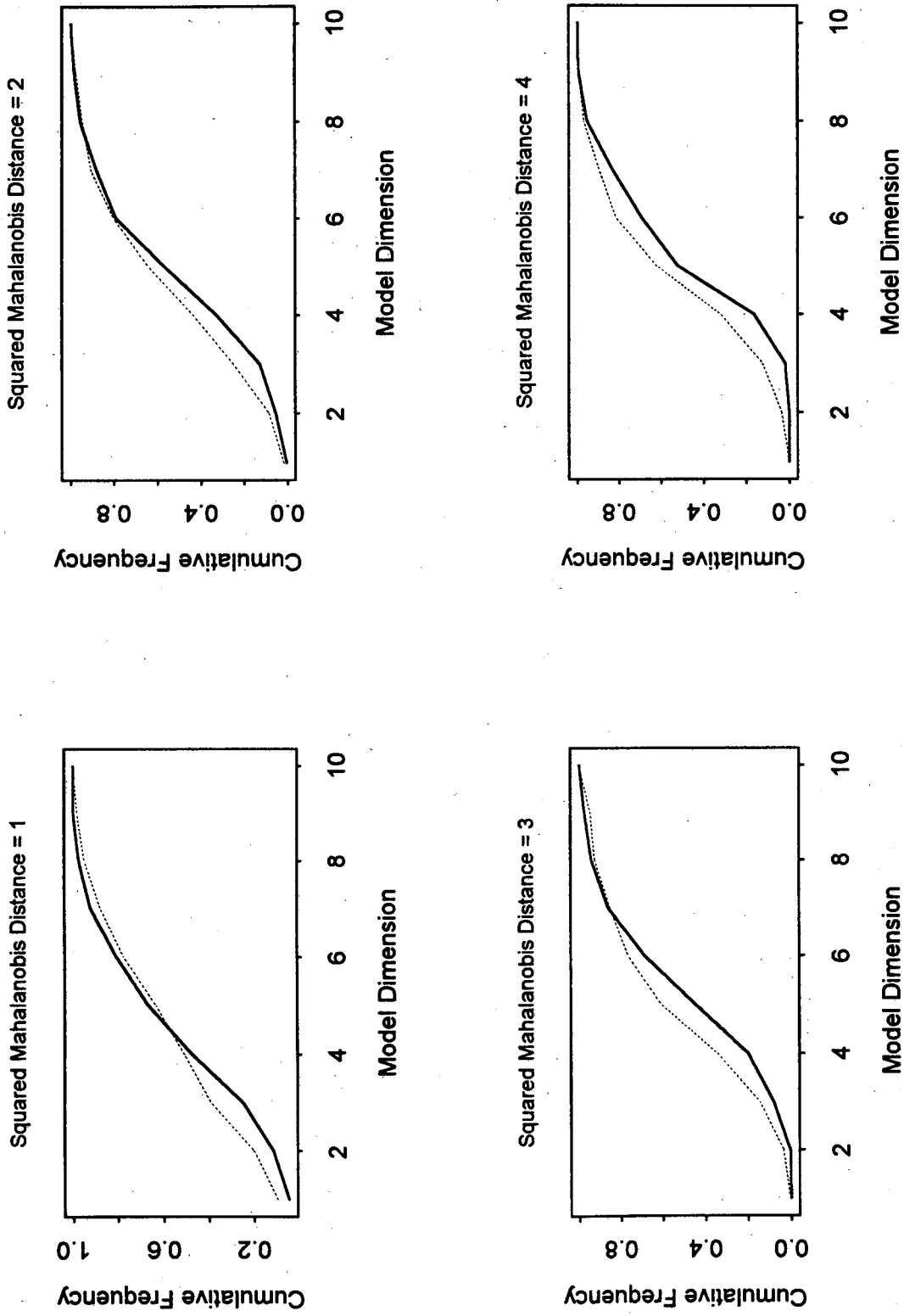


FIG. 4.32: CUMULATIVE FREQUENCY PLOT, CASE N2

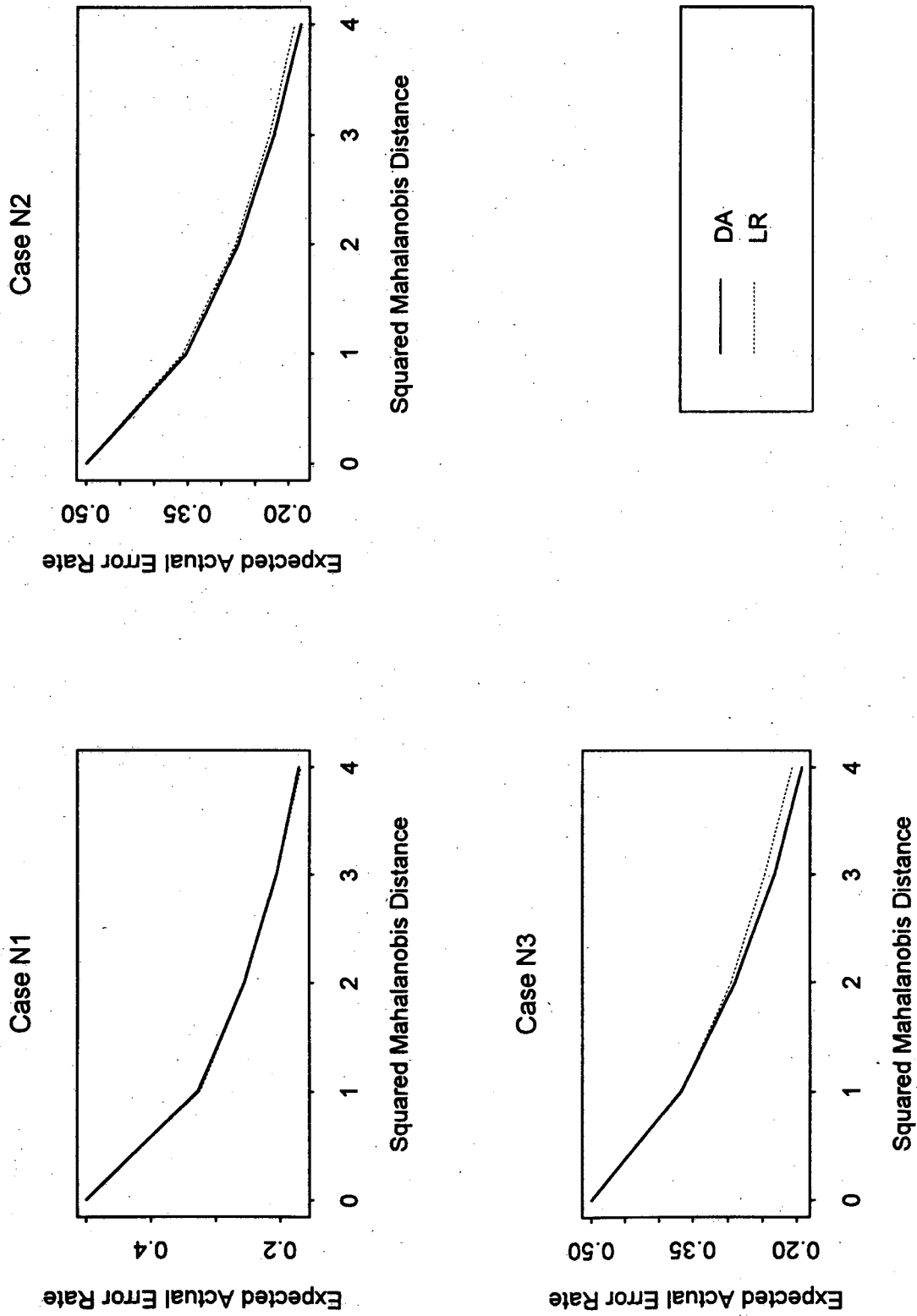


FIG. 4.33: EXPECTED ACTUAL ERROR RATE, NORMAL DATA

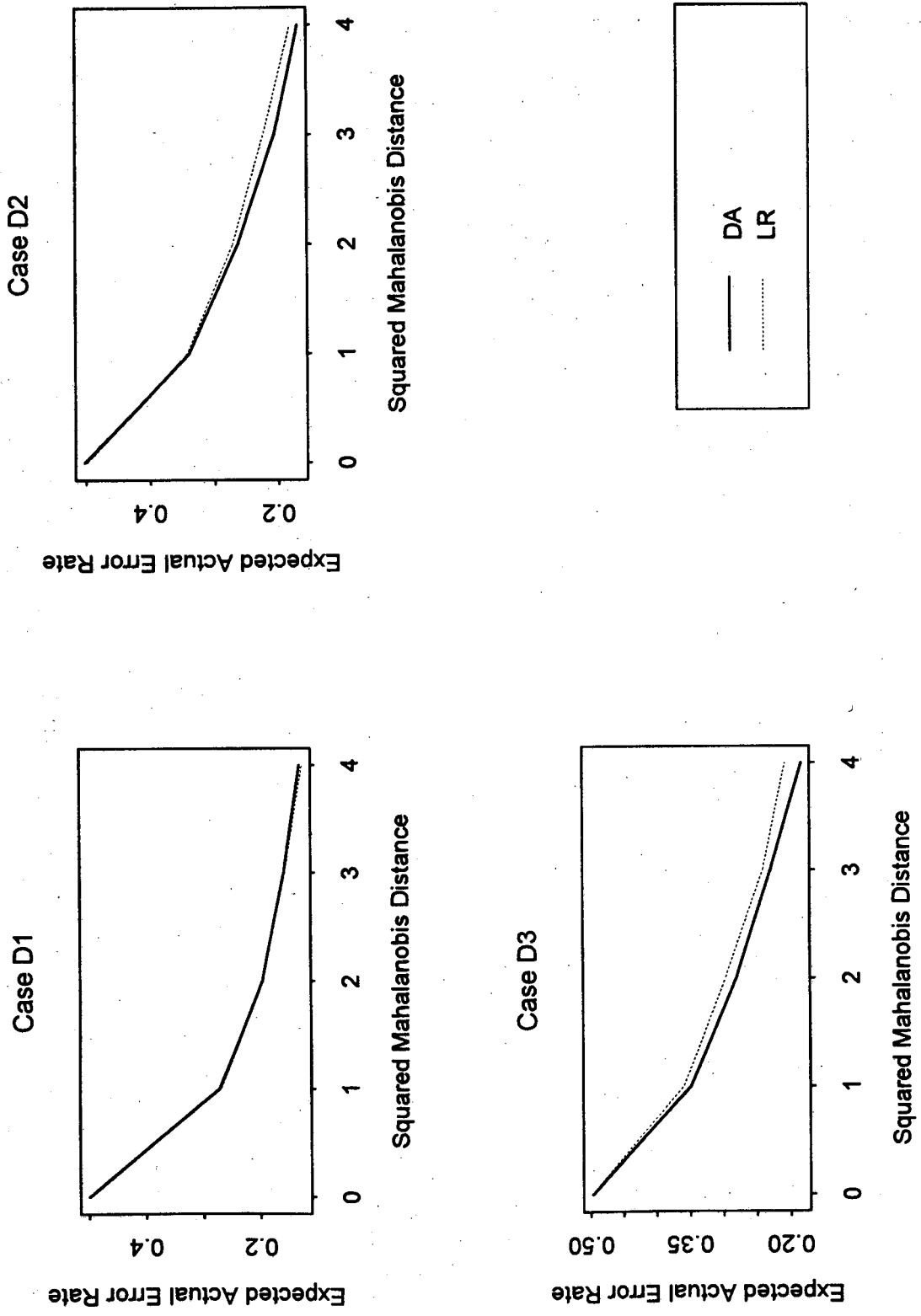


FIG. 4.34: EXPECTED ACTUAL ERROR RATE, DOUBLE EXPONENTIAL DATA

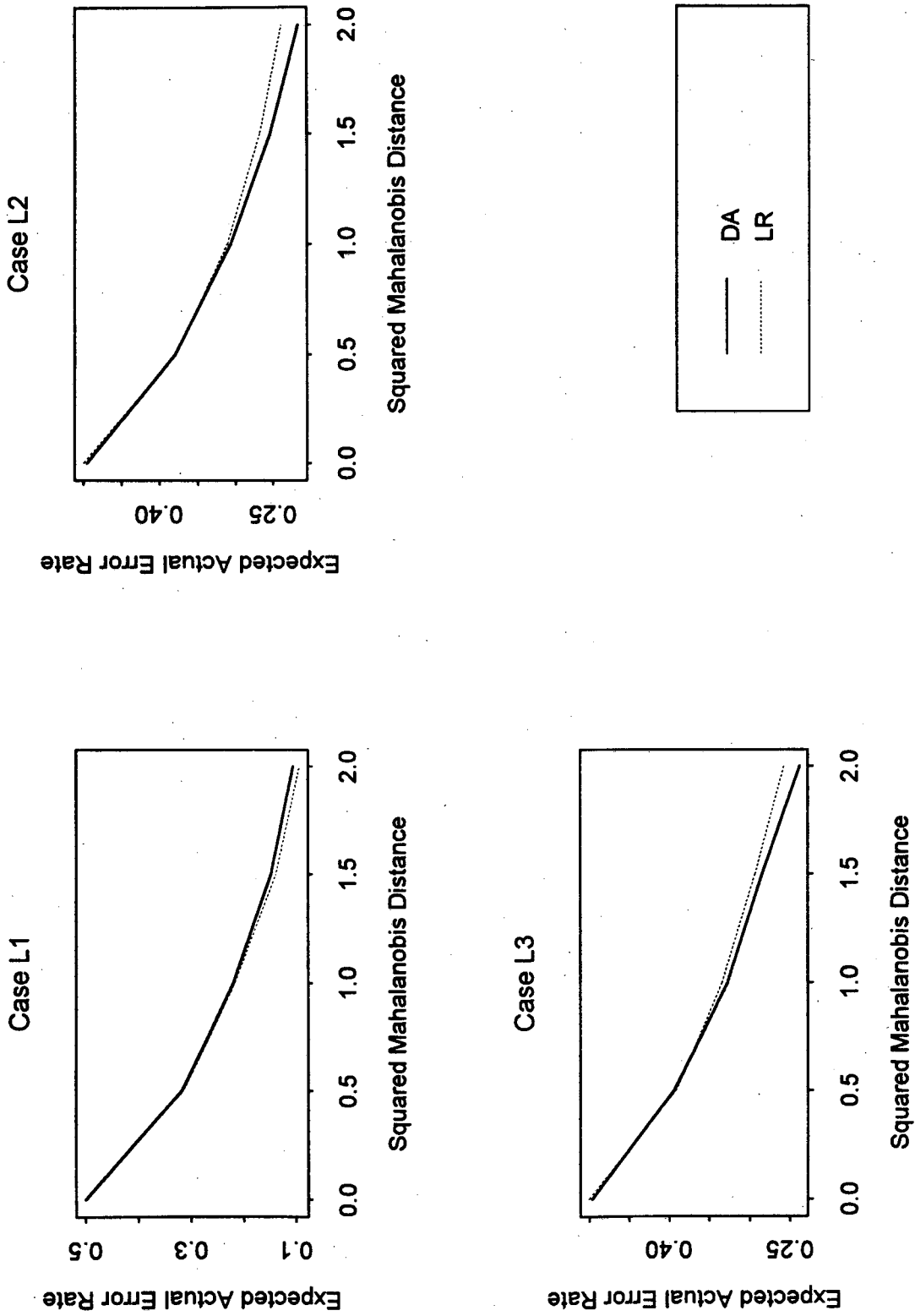


FIG. 4.35: EXPECTED ACTUAL ERROR RATE, LOGNORMAL DATA

4.9 APPLICATION OF CROSS MODEL VALIDATION AND OTHER TECHNIQUES TO REAL LIFE DATA SETS

The cross model validation techniques proposed in this chapter for variable selection in discriminant analysis and logistic regression, were applied to two real life data sets. In each case, the other techniques discussed earlier in this chapter, were also applied to the data, and the results obtained are compared to that obtained by means of cross model validation.

4.9.1 CORPORATE FAILURE DATA

The prediction of corporate failure is important to shareholders and creditors, in order to identify companies that are at risk of being declared insolvent. Discriminant analysis and logistic regression are often used to differentiate between solvent and insolvent companies, as well as for prediction of future failure. Olivier (1990) investigated prediction of corporate failure based on ratio variables for trade and manufacturing companies listed on the Johannesburg Stock Exchange (JSE). He used a data set consisting of 24 insolvent companies that were delisted from the JSE between 1970 and 1988 as a result of financial failure, as well as 55 solvent companies. For the insolvent companies, data were obtained from financial reports that were published one to five years prior to failure. Only the data pertaining to the one year lag are considered here. For the solvent companies, those that were listed on the JSE in 1982, and were still listed in 1988 were considered, to avoid the possibility of including a company that could fail in the immediate future. Financial reports of 1982 were used to obtain the data for these companies.

The data consisted of observations on 35 ratio variables (referred to as X1 to X35), such as nett income before taxes to total assets, increase in turnover to turnover in previous year and cash flow before taxes to total debt. One of the aims of Olivier (1990) was to identify a subset of the 35 ratio variables which discriminates well between the solvent and insolvent firms, and that could be used for the prediction of future failures. To achieve this, he used an F-based fully stepwise selection procedure with F-to-enter = 4 and F-to-delete = 2.996. This procedure selected the following variables: X7, X23, X35. To estimate the post-selection error rate, Olivier (1990) calculated the cross validation (leave-one-out) error rate, and obtained a value of 0.038. Since the same data set is used for the selection of variables and for the estimation of the post-selection error rate, it seems reasonable to suspect that this estimate gives an optimistic impression of the classification performance of the rule based on X7, X23, X35, when used to predict new cases.

The cross model validation technique with F-based forward selection as inner criterion (referred to as CMV-1 in Chapter 4), as well as the cross model validation technique with an all possible subsets approach based on R^2 as inner criterion (referred to as CMV-2 in Chapter 4), were applied to the data set to select a subset of variables for

inclusion in a linear discriminant function. As suggested in Section 4.4, a graph of the CMV - criterion for each possible model dimension p is plotted against p and used as an aid in determining the final model dimension. The graph for the CMV-1 technique appears in Fig 4.36, and that for the CMV-2 technique, in Fig. 4.37. From these graphs, it is clear that the CMV-criterion is not a monotone function of model dimension. Perusal of Fig. 4.36 shows that the minimum CMV-value (0.1616) is attained at model dimension $p = 3$. Addition of further variables, leads to a sharp increase in the value of the criterion, and only at model dimensions 29 and 31 does the criterion approach the minimum value quoted above. In this case, the choice of a model dimension of $p = 3$ is clear. It should be noted that even if the CMV-criterion attained a smaller value than 0.1616 at say model dimension 29, one would be hesitant to select 29 variables for inclusion in the linear discriminant function. Applying the inner criterion (F-based forward selection) to the full data set to select a model of the optimal dimension (three), led to selection of the following variables: X7, X23, X35. The selected subset therefore contains exactly the same variables that were selected by Olivier (1990). The error rate estimate yielded by the CMV-1 procedure, is 0.1616 (the value of the CMV-criterion at the optimal model dimension), which is much larger than the leave-one-out estimate used by Olivier (1990). Since the cross model validation technique is specifically aimed at reducing selection bias, it seems reasonable that this estimate gives a better indication of the performance of the linear discriminant rule based on X7, X23, X35, than the leave-one-out estimate. F-based forward selection with α - to - enter = 0.15, was also applied to the data, and selected the same subset (X7, X23, X35). The NS_k^* - estimate (cf. Snapinn and Knoke, 1989) was also calculated for the rule based on this subset, and had a value of 0.2140.

Fig. 4.37 contains a graph of $CMV(p)$ against p for the CMV-2 procedure (using all possible subsets selection based on R^2 as inner criterion). A comparison of this graph to the graph in Fig. 4.36 reveals that, except at model dimensions 1, 2 and 35, the CMV-criterion of the CMV-2 technique is lower than that of the CMV-1 technique at the same dimension. This is intuitively clear from the following explanation. When applying the CMV-technique, the best model of each dimension is found by applying the inner criterion to the data set with one case omitted. The linear discriminant function based on the selected variables is then used to classify the omitted case, and a measure of loss is calculated. The CMV-criterion associated with each dimension, is the average loss for that dimension, averaged over all omitted cases. When a forward selection procedure is used as inner criterion, as is the case in the CMV-1 procedure, the optimal model of dimension p ($p \geq 2$) is found by comparing only models containing all variables included in the optimal model of dimension $p - 1$, plus an additional variable not yet included in the model. When an all possible subsets selection procedure is used as inner criterion, as in the CMV-2 procedure, all possible subsets of dimension p are considered, and the optimal model may be one that was never considered in a forward selection procedure. It is therefore reasonable to expect the CMV-criterion of the CMV-2 procedure to be lower (or at least not higher) than that of the CMV-1 procedure at the same dimension.

The minimum value of the CMV-criterion for the CMV-2 procedure, 0.1416, is attained at model dimension 7. When the inner criterion (all possible subsets selection based on R^2) is applied to the full data set to select a model of this dimension, the following variables are selected: X4, X5, X6, X9, X16, X25, X31. The value of CMV(7), 0.1416, is used as an estimate of the error rate of the linear discriminant rule based on these variables.

From this example, it is evident that the optimal model dimension and the variables selected for inclusion into the discriminant function, may be quite different when using the two different inner criteria. For this data set, the computing time on a Hewlett Packard 712/60 for the CMV-1 procedure was approximately 7 minutes, while the time for the CMV-2 procedure was approximately 42 minutes. With the increase in computer power, use of an all possible subsets approach as inner criterion is entirely feasible and is recommended in preference to use of a forward selection procedure as inner criterion.

The results of the analyses described above, are summarised in Table 4.1.

The logistic regression cross model validation procedure was also applied to the data set, but the maximum likelihood estimates of the logistic regression coefficients did not exist, because the two populations are very well separated.

TABLE 4.1 RESULTS OF VARIABLE SELECTION AND ERROR RATE ESTIMATION, CORPORATE FAILURE DATA SET

SELECTION METHOD (error rate estimator)	RATIO VARIABLES SELECTED	VALUE OF ERROR RATE ESTIMATOR
Stepwise selection (leave-one-out estimator)	X7, X23, X35	0.0380
CMV-1 (CMV-estimator)	X7, X23, X35	0.1616
CMV-2 (CMV-estimator)	X4, X5, X6, X9, X16, X25, X31	0.1416
Forward selection (NS_k^* - estimator)	X7, X23, X35	0.2140

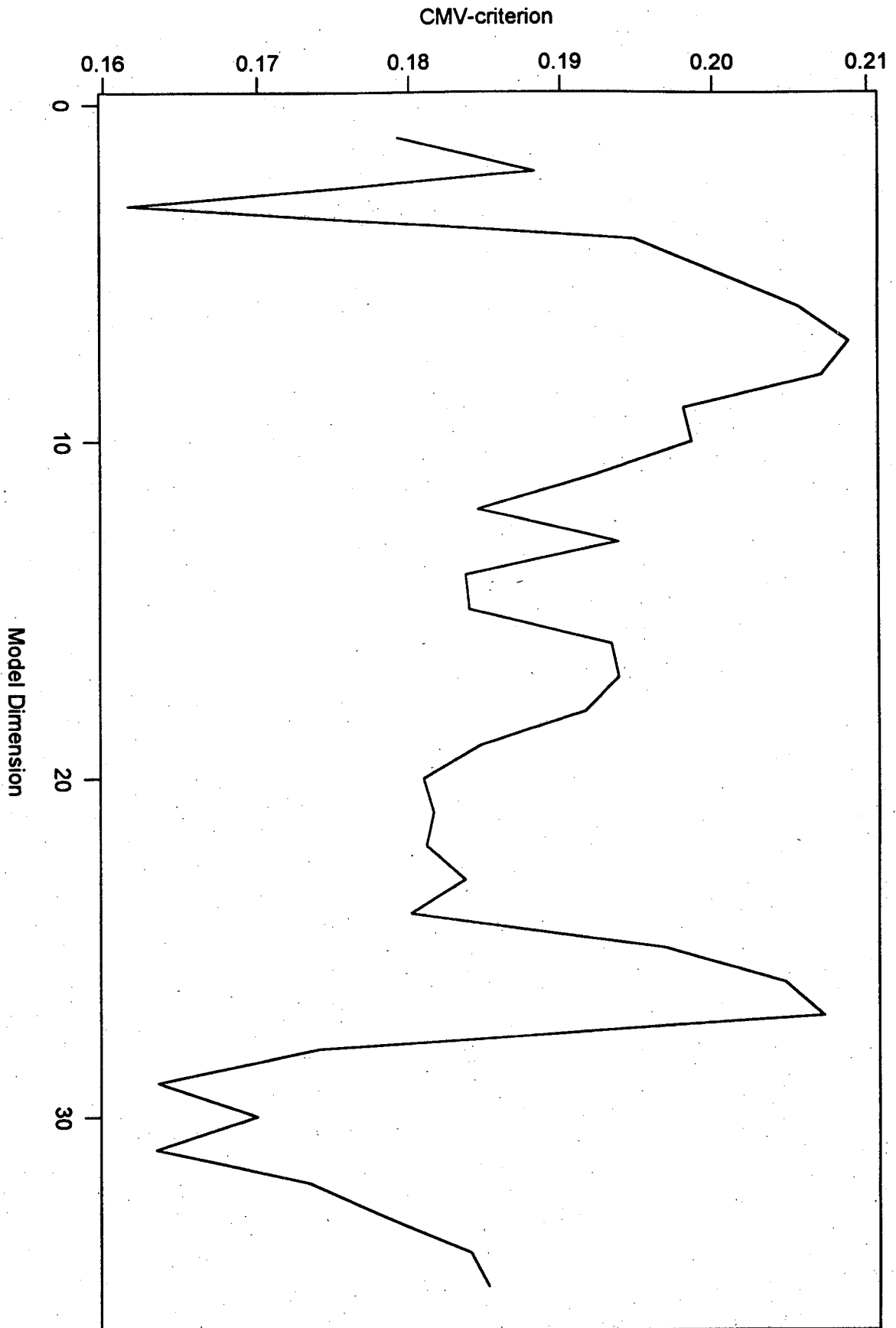


FIG. 4.36: PLOT OF CMV(p) AGAINST p, COMPANY DATA, CMV-1 PROCEDURE

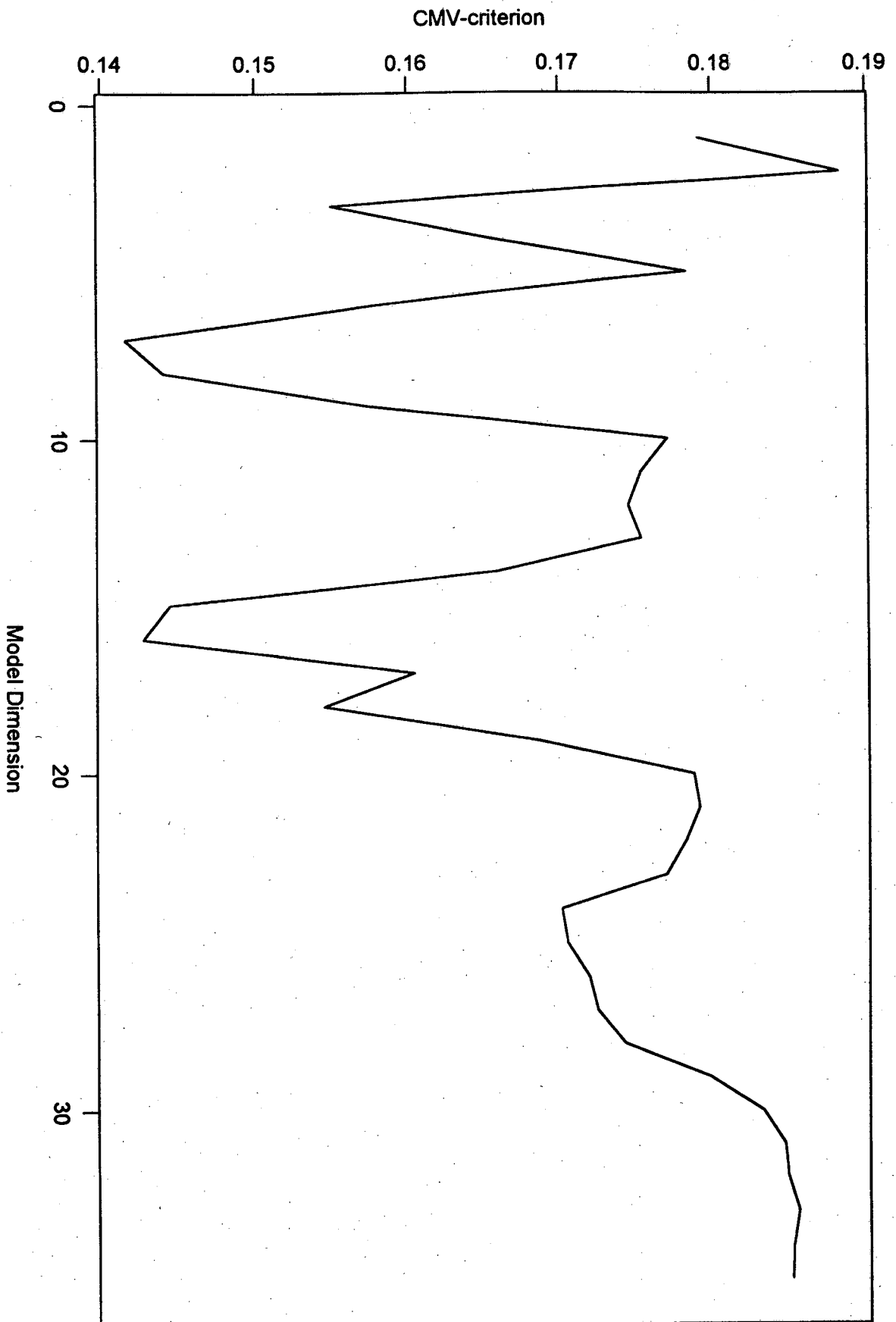


FIG. 4.37: PLOT OF CMV(p) AGAINST p, COMPANY DATA, CMV-2 PROCEDURE

4.9.2 SWISS BANK NOTE DATA

The techniques used to analyse the corporate failure data set, were also applied to a data set on genuine and forged Swiss bank notes, given by Flury and Riedwyl (1988). The data set consists of observations on 100 genuine and 100 forged thousand franc bills. The following six variables were observed:

- X1, the length of the bill,
- X2, the width of the bill measured on the left,
- X3, the width of the bill measured on the right,
- X4, the width of the margin at the bottom,
- X5, the width of the margin at the top,
- X6, the length of the image diagonal.

The aim is to select a subset of the variables that best differentiates between the genuine and forged bills.

The same techniques used in Section 4.9.1 to analyse the corporate failure data, were applied to the Swiss bank note data. For this data set, the value of $CMV(p)$ at each value of p ($p = 1, \dots, 6$) are exactly the same for the CMV-1 and CMV-2 techniques. A graph displaying the values of $CMV(p)$ against p , appears in Fig. 4.38. From this graph, it is evident that the optimal model dimension for the bank note data, is $p = 3$. This example illustrates why use of a graph or implementation of the strategy involving ϕ , as described in Section 4.4, is recommended, rather than Hjorth's suggestion of choosing the model dimension yielding the absolute minimum. For this data set, the absolute minimum of the CMV-criterion (0.0050002) occurs at model dimension 4, while the value of the CMV-criterion at dimension 3 is 0.0050016. Implementation of the procedure involving ϕ , described in Section 4.4, or use of the graph in Fig. 4.38, would lead to a choice of model dimension 3. It is indeed questionable whether an additional variable should be included if the resulting improvement in the classification performance (based on the CMV-estimates of the error rate) is as small as 0.0000014.

Applying either F-based forward selection or all possible subsets selection based on R^2 to the full data set to select a subset of the optimal dimension (3), leads to selection of the following variables: X4, X5, X6. The cross model validation estimate of the post-selection error rate, is 0.005.

Forward selection with α -to-enter = 0.15, selects a subset containing variables X2, X3, X4, X5 and X6. The NS_k^* -estimate (cf. Snapinn and Knoke, 1989) was calculated for the rule based on this subset, and had a value of 0.0049. A fully stepwise selection procedure similar to that applied by Olivier (1990) to the corporate failure data, selected the same subset. The leave-one-out error rate based on this subset is equal to 0.005.

The results of the analyses are summarised in Table 4.2. The computing times on a Hewlet Packard 712/60 computer was 27 seconds for the CMV-1 procedure and 16 seconds for the CMV-2 procedure. It is interesting that for a relatively small number of variables, the all possible subsets procedure takes less time than the stepwise procedure.

The logistic regression cross model validation procedure was also applied to the data set, but because of the large separation between the two groups, the maximum likelihood estimates of the coefficients do not exist.

TABLE 4.2 RESULTS OF VARIABLE SELECTION AND ERROR RATE ESTIMATION, SWISS BANK NOTE DATA SET

SELECTION METHOD (error rate estimator)	VARIABLES SELECTED	VALUE OF ERROR RATE ESTIMATOR
Stepwise selection (leave-one-out estimator)	X2, X3, X4, X5, X6	0.0050
CMV-1 (CMV-estimator)	X4, X5, X6	0.0050
CMV-2 (CMV-estimator)	X4, X5, X6	0.0050
Forward selection (NS_k^* - estimator)	X2, X3, X4, X5, X6	0.0049

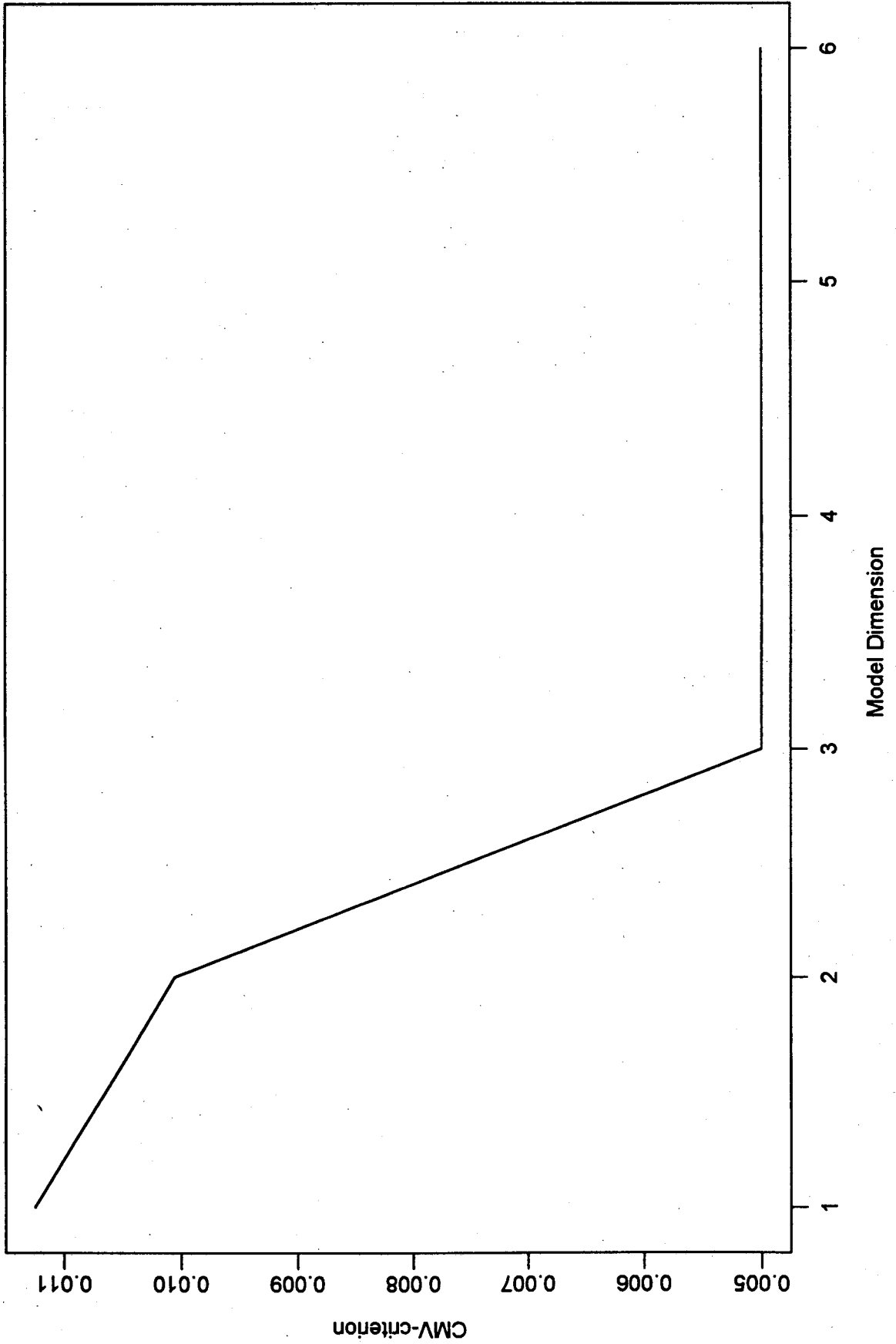


FIG. 4.38: PLOT OF CMV(p) AGAINST p, SWISS BANK NOTE DATA

4.10 CONCLUSIONS AND RECOMMENDATIONS

It is worthwhile to summarise the main conclusions emanating from the extensive simulation study reported in this chapter. Firstly, a few general conclusions.

1. The usefulness of logistic regression as a classification technique is limited by the non-existence of the maximum likelihood estimates of the parameters in the classification function when the populations are fairly well separated. This problem was encountered during the simulation study, and also in both examples discussed in Section 4.9.
2. An allocatory approach to variable selection should be used if the classification performance of the rule being constructed is of prime importance. In such cases the selection criterion is typically an error rate estimator. Using an error rate estimator based on a 0-1 loss function has a disadvantage in this context, viz. that it can easily lead to more than one model being identified as optimal. This problem can be overcome by using a smoothed version of the 0-1 loss function.
3. It is well known that naive error rate estimators such as the apparent error rate, is optimistically biased in a non-selection context. Somewhat less well known is the fact that estimators that perform acceptably in a non-selection context do not take selection induced bias into account, and consequently do not fare well when post-selection error rate has to be estimated. A need therefore exists for error rate estimators developed specifically for application in a selection context.

In this chapter the problems of variable selection and subsequent error rate estimation were addressed by introducing the cross model validation technique for discriminant analysis and logistic regression. This technique has a number of advantages.

1. The cross model validation technique is based on separatory as well as allocatory considerations. A separatory approach is simpler to implement and is sufficient when only models of the same dimension are compared, as is the case during the first stage of cross model validation. However, the decision regarding a final model dimension should be based on allocatory considerations, as is done in cross model validation.
2. The CMV-technique combines variable selection and subsequent error rate estimation in a sensible way, rather than considering these closely related problems separately.
3. Both in terms of variable selection and error rate estimation, the CMV-technique performs excellently. Application of the technique yields a rule with good classification properties, and at the same time provides an accurate estimate of the error rate of this rule.

4. Validity of the CMV-technique does not depend on assumptions regarding the distribution of the feature variables. The technique was found to perform well for data from the normal distribution as well as a number of heavy-tailed and skewed alternatives.

5. In the practical application of the technique a plot of the CMV-criterion against model dimension provides very useful information. It enables a user of the technique to weigh the complexity of the model against its expected classification accuracy, thereby making it easier to reach a decision on the model that should be selected. This is clearly illustrated in the two examples in Section 4.9.

Although the cross model validation technique is numerically intensive, this is not a serious disadvantage. In practical applications of the technique, it is only applied once to a given data set, and this is easily done if a suitable computer program is available.

CHAPTER 5

PRE-TEST VARIABLE SELECTION

5.1 INTRODUCTION

The topics of *preliminary test estimation* and *preliminary test variable selection* (for the sake of brevity, pre-test estimation and variable selection) have received considerable attention in the literature (cf. Venter and Steel, 1994, and the references therein). The following simple example illustrates what is meant by these terms. Consider a $N(\theta,1)$ distributed random variable X and suppose θ has to be estimated. An example of a *pre-test estimator* is

$$\hat{\theta} = XI(|X| > c) \quad (5.1.1)$$

where c is a pre-specified constant. In (5.1.1), θ is estimated by 0 if $|X| \leq c$, and by X otherwise. Using $\hat{\theta}$ to estimate θ is equivalent to first testing the hypothesis $H: \theta = 0$. If this hypothesis is rejected, i.e. if $|X| > c$, θ is estimated by X , and if accepted, i.e. if $|X| \leq c$, θ is estimated by 0. The constant c can be chosen to fix the significance level at which the hypothesis H is tested, and its choice naturally also influences the properties of $\hat{\theta}$. For example, the mean squared error of $\hat{\theta}$ is

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E[X - \theta - XI(|X| \leq c)]^2 \\ &= 1 + E\{[X^2 - 2X(X - \theta)]I(|X| \leq c)\} \\ &= 1 + E\{[\theta^2 - (X - \theta)^2]I(|X| \leq c)\} \end{aligned} \quad (5.1.2)$$

where the expectation is taken with respect to the $N(\theta,1)$ distribution of X .

The above example can also be used to explain what is meant by *pre-test variable selection*. Consider the case of linear regression, and suppose X is the least squares estimator of a regression coefficient θ . Then accepting $H: \theta = 0$ would imply that the variable corresponding to θ should be excluded from the model being fitted, and rejecting H would lead to inclusion of the variable. Although this is an oversimplification of what occurs in practice, the basic idea underlying pre-test selection is well illustrated.

In Section 5.2 general aspects of pre-test variable selection are discussed in more detail. Two important cases are distinguished, viz. the case where c in (5.1.1) does not depend on the data, and the case where c is data-dependent. Pre-test selection procedures introduced by Venter and Steel (1993, 1994), that use a data-dependent specification of c , are also discussed. In Section 5.3 it is shown how one of the pre-test variable selection procedures of Venter and Steel (1994) can be applied in discriminant analysis. The limitations of this procedure are also emphasised. Section 5.4 is devoted to a discussion of error rate estimation following pre-test variable selection. A cross validation based approach is proposed, and attempts to reduce the variance of the resulting error rate estimator, are described. The chapter closes in Section 5.5 with a description of a simulation study that was undertaken to investigate the operating characteristics of the proposed selection and estimation procedure. The main conclusions are: provided that the underlying assumptions of normally distributed and independent feature variables are satisfied, pre-test selection performs very well from a separatory point of view, while the proposed post-selection error rate estimator has very low bias but a fairly large variance.

5.2 GENERAL ASPECTS OF PRE-TEST SELECTION

A number of important points on pre-test variable selection can best be illustrated by considering the following simplified model selection situation (cf. Venter and Steel, 1993 and 1994). Let X_1, \dots, X_k be independent random variables, with $X_i \sim N(\theta_i, \sigma^2)$ distributed, $i = 1, \dots, k$. Let \mathbf{X} and $\boldsymbol{\theta}$ be the k -vectors with elements X_1, \dots, X_k and $\theta_1, \dots, \theta_k$ respectively. Assume initially that the value of σ^2 is known. This assumption will later be relaxed, and the required modifications will be discussed. The model selection problem is to use the data to select a member from the *some zeros family of models*. This family has 2^k members, of which a typical one states that $\theta_i \neq 0$ if and only if $i \in \mathcal{J}$, where \mathcal{J} is a subset of the set of indices $\mathcal{K} = \{1, \dots, k\}$. For the sake of brevity, the model corresponding to a given subset \mathcal{J} will be referred to as model \mathcal{J} .

How should \mathcal{J} be selected from \mathcal{K} ? An answer to this question can be formulated in terms of a pre-test estimator of $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_1$ be the k -vector with i -th component

$$\hat{\theta}_{1i} = X_i I(i \in \mathcal{J}) \quad (5.2.1)$$

for $i = 1, \dots, k$, with \mathcal{J} a given subset of \mathcal{K} . The worth of $\hat{\boldsymbol{\theta}}_1$ as an estimator of $\boldsymbol{\theta}$ can be judged in terms of its mean squared error, given by

$$\begin{aligned} R_1(\theta, \sigma^2) &= E \left[\sum_{i=1}^k (\hat{\theta}_{1i} - \theta_i)^2 \right] \\ &= k\sigma^2 + \sum_{\{i: i \in J\}} (\theta_i^2 - \sigma^2). \end{aligned} \quad (5.2.2)$$

Since $R_1(\theta, \sigma^2)$ depends on θ (and σ^2), its value is unknown. However, with σ^2 known an unbiased estimate of (5.2.2) is

$$k\sigma^2 + \sum_{\{i: i \in J\}} (X_i^2 - 2\sigma^2). \quad (5.2.3)$$

A strategy that can now be used to select J is to choose $J = J(\mathbf{X})$ to minimise the estimate (5.2.3). This implies

$$J(\mathbf{X}) = \{i: |X_i| > \sqrt{2}\sigma\}. \quad (5.2.4)$$

In (5.2.4), the notation $J(\mathbf{X})$ emphasises that the subset $J(\mathbf{X})$ of \mathcal{K} is selected data-dependently. Selecting model $J(\mathbf{X})$ from the some zeros family implies that the θ_i 's corresponding to the X_i 's for which $|X_i| > \sqrt{2}\sigma$ are considered non-zero. Following selection of $J(\mathbf{X})$, the corresponding pre-test estimator has i -th component

$$\tilde{\theta}_{1i} = X_i I(|X_i| > \sqrt{2}\sigma), \quad (5.2.5)$$

for $i = 1, \dots, k$. The mean squared error of this estimator is given by

$$\begin{aligned} \tilde{R}_1(\theta, \sigma^2) &= E \left[\sum_{i=1}^k (\tilde{\theta}_{1i} - \theta_i)^2 \right] \\ &= k\sigma^2 + E \sum_{i=1}^k \left\{ [\theta_i^2 - (X_i - \theta_i)^2] I(|X_i| \leq \sqrt{2}\sigma) \right\}. \end{aligned} \quad (5.2.6)$$

This approach for selecting $J(\mathbf{X})$ is a component-wise approach, since the pre- and post-selection estimators (5.2.1) and (5.2.5) respectively, both consider the cases corresponding to $i = 1, \dots, k$ separately.

A second possibility for selecting $J(\mathbf{X})$ that treats the components in a combined manner was introduced by Venter and Steel (1993, 1994). Let $Z_1 < Z_2 < \dots < Z_k$ be the order statistics of $|X_1|, \dots, |X_k|$, and put $Z_0 = 0$. Suppose the some zeros model being considered specifies $q < k$ of the θ_i 's to be zero. Then it makes sense that these

should be the θ_i 's corresponding to the q smallest absolute observations. This point of view implies that the pre-test estimator with i -th component (5.2.1) should be replaced by $\hat{\theta}_2$, with i -th component

$$\hat{\theta}_{2i} = X_i I(|X_i| > Z_q), \quad (5.2.7)$$

for $i = 1, \dots, k$. The only problem that remains is how to specify the integer q from the data, $0 \leq q \leq k$. Venter and Steel (1993, 1994) proposed an approach similar to the one summarised in (5.2.2) - (5.2.4), viz. to estimate the mean squared error of $\hat{\theta}_2$ and to choose q from $\{0, 1, \dots, k\}$ to minimise this estimate. Details in this regard are now provided.

From (5.2.7) it follows that

$$\hat{\theta}_{2i} = X_i - g_i(\mathbf{X}) \quad (5.2.8)$$

where

$$g_i(\mathbf{X}) = X_i I(|X_i| \leq Z_q). \quad (5.2.9)$$

Let $Z_1(i) < Z_2(i) < \dots < Z_{k-1}(i)$ be the order statistics of the $k-1$ $|X_j|$'s with $j \neq i$, and put $Z_0(i) = 0$, $Z_k(i) = \infty$. Suppose $|X_i| = Z_{l_i}$, $i = 1, \dots, k$. Then $Z_j(i) = Z_j$ for $j = 1, \dots, l_i - 1$ and $Z_j(i) = Z_{j+1}$ for $j = l_i, \dots, k-1$. Hence,

$$I(|X_i| \leq Z_q) = \begin{cases} 0, & \text{if } q < l_i \\ 1, & \text{if } q \geq l_i, \end{cases}$$

and

$$I(|X_i| \leq Z_q(i)) = \begin{cases} 0, & \text{if } q < l_i \\ 1, & \text{if } q \geq l_i, \end{cases}$$

and therefore

$$I(|X_i| \leq Z_q) = I(|X_i| \leq Z_q(i)). \quad (5.2.10)$$

Similarly,

$$Z_q I(|X_i| > Z_q) = \begin{cases} 0, & \text{if } q \geq l_i \\ Z_q, & \text{if } q < l_i, \end{cases}$$

and

$$Z_q(i)I(|X_i| > Z_q(i)) = \begin{cases} 0 & , \text{ if } q \geq l_i \\ Z_q(i) & , \text{ if } q < l_i, \end{cases}$$

and since $Z_q(i) = Z_q$ for $q < l_i$,

$$Z_q I(|X_i| > Z_q) = Z_q(i) I(|X_i| > Z_q(i)). \quad (5.2.11)$$

From (5.2.10) it is clear that (5.2.9) can also be written in the form

$$g_i(\mathbf{X}) = X_i I(|X_i| \leq Z_q(i)). \quad (5.2.12)$$

The mean squared error of $\hat{\theta}_{2i}$ is thus given by

$$\begin{aligned} E(\hat{\theta}_{2i} - \theta_i)^2 &= E[X_i - \theta_i - X_i I(|X_i| \leq Z_q(i))]^2 \\ &= \sigma^2 + E[X_i^2 I(|X_i| \leq Z_q(i))] - 2E[(X_i - \theta_i)X_i I(|X_i| \leq Z_q(i))]. \end{aligned} \quad (5.2.13)$$

Now consider $E[(X - \theta)XI(|X| \leq a)]$, where X is $N(\theta, \sigma^2)$ distributed and a is a constant. By using partial integration and the identity $\int x\phi(x)dx = -\phi(x)$, with $\phi(x)$ the density function of the standard normal distribution, it is found that

$$\begin{aligned} E[(X - \theta)XI(|X| \leq a)] &= \sigma^2 \left\{ b_1\phi(b_1) - b_2\phi(b_2) + \int_{b_1}^{b_2} \phi(y)dy \right\} \\ &\quad - \sigma\theta\{\phi(b_2) - \phi(b_1)\} \\ &= \sigma \left\{ (\sigma b_1 + \theta)\phi(b_1) - (\sigma b_2 + \theta)\phi(b_2) + \sigma \int_{b_1}^{b_2} \phi(y)dy \right\}, \end{aligned} \quad (5.2.14)$$

where $b_1 = -(a + \theta)/\sigma$ and $b_2 = (a - \theta)/\sigma$. By conditioning on $Z_q(i)$ in (5.2.13), and using (5.2.14), it follows that

$$\begin{aligned} E(\hat{\theta}_{2i} - \theta_i)^2 &= \sigma^2 + E(X_i^2 - 2\sigma^2)I(|X_i| \leq Z_q(i)) \\ &\quad + 2\sigma E Z_q(i) \left[\phi\left(\frac{Z_q(i) - \theta_i}{\sigma}\right) + \phi\left(\frac{Z_q(i) + \theta_i}{\sigma}\right) \right], \end{aligned}$$

for $i = 1, \dots, k$. If (5.2.10) is once more applied, the total mean squared error of $\hat{\theta}_2$ is found to be

$$\begin{aligned} R_2(\theta, \sigma^2) &= E \left[\sum_{i=1}^k (\hat{\theta}_{2i} - \theta_i)^2 \right] \\ &= k\sigma^2 + E \sum_{i=1}^k (X_i^2 - 2\sigma^2)I(|X_i| \leq Z_q) \\ &\quad + 2\sigma E \sum_{i=1}^k Z_q(i) \left[\phi\left(\frac{Z_q(i) - \theta_i}{\sigma}\right) + \phi\left(\frac{Z_q(i) + \theta_i}{\sigma}\right) \right]. \end{aligned} \tag{5.2.15}$$

A plug-in estimator can be used for the last term on the right hand side of (5.2.15), and this entails replacing θ_i by X_i . Still assuming σ^2 to be known, an estimator of (5.2.15) is therefore

$$\begin{aligned} &= k\sigma^2 + \sum_{i=1}^q (Z_i^2 - 2\sigma^2) + 2\sigma \sum_{i=1}^k Z_q(i) \left[\phi\left(\frac{Z_q(i) - X_i}{\sigma}\right) + \phi\left(\frac{Z_q(i) + X_i}{\sigma}\right) \right]. \end{aligned} \tag{5.2.16}$$

This expression can be simplified by splitting the sum in the final term according to $|X_i| \leq Z_q(i)$ and $|X_i| > Z_q(i)$. In the former case, $Z_q(i) = Z_{q+1}$, and in the latter case, $Z_q(i) = Z_q$. An estimator of (5.2.15) is then found to be

$$\begin{aligned} \tilde{r}(\mathbf{X}, q) &= k\sigma^2 + \sum_{i=1}^q (Z_i^2 - 2\sigma^2) + 2\sigma Z_{q+1} \sum_{i=1}^q \left[\phi\left(\frac{Z_{q+1} - Z_i}{\sigma}\right) + \phi\left(\frac{Z_{q+1} + Z_i}{\sigma}\right) \right] \\ &\quad + 2\sigma Z_q \sum_{i=q+1}^k \left[\phi\left(\frac{Z_q - Z_i}{\sigma}\right) + \phi\left(\frac{Z_q + Z_i}{\sigma}\right) \right]. \end{aligned} \tag{5.2.17}$$

Since (5.2.15) is a non-negative quantity, it makes sense to replace (5.2.17) by

$$r(\mathbf{X}, q) = \max\{0, \tilde{r}(\mathbf{X}, q)\}. \quad (5.2.18)$$

Two special cases are $q = 0$ and $q = k$.

For $q = 0$, $\hat{\theta}_{2i} = X_i$ for all i , and hence

$$r(\mathbf{X}, 0) = k\sigma^2. \quad (5.2.19)$$

For $q = k$, $\hat{\theta}_{2i} = 0$ for all i , and hence

$$E\left[\sum_{i=1}^k (\hat{\theta}_{2i} - \theta_i)^2\right] = \sum_{i=1}^k \theta_i^2.$$

This is estimated unbiasedly by $\sum_{i=1}^k X_i^2 - k\sigma^2$, and hence

$$r(\mathbf{X}, k) = \max\left\{0, \sum_{i=1}^k X_i^2 - k\sigma^2\right\}. \quad (5.2.20)$$

Venter and Steel (1994) propose that q should be selected from $\{0, 1, \dots, k\}$ to minimise (5.2.18). Denote this value of q by \hat{q} . Then the subset $J = J(\mathbf{X})$ corresponding to \hat{q} is given by

$$J(\mathbf{X}) = \{i: |X_i| > Z_{\hat{q}}\}. \quad (5.2.21)$$

Selecting model $J(\mathbf{X})$ from the some zeros family implies that the θ_i 's corresponding to observations X_i for which $|X_i| \leq$ the \hat{q} -th absolute order statistic, are viewed as zero, while the θ_i 's corresponding to the remaining X_i 's are considered to be non-zero.

The post-selection pre-test estimator has i -th component

$$\tilde{\theta}_{2i} = X_i I(|X_i| > Z_{\hat{q}}), \quad (5.2.22)$$

for $i = 1, \dots, k$, and mean squared error

$$\tilde{R}_2(\theta, \sigma^2) = k\sigma^2 + E\sum_{i=1}^k X_i^2 I(|X_i| \leq Z_{\hat{q}}) - 2E\sum_{i=1}^k X_i (X_i - \theta_i) I(|X_i| \leq Z_{\hat{q}}). \quad (5.2.23)$$

A comparison of (5.2.5) and (5.2.22) reveals the similarity and the difference between the two post-selection pre-test estimators. Whereas $|X_i|$ is compared with a fixed constant in (5.2.5), it is compared with a data dependent quantity in (5.2.22). Venter and Steel (1994) used simulation to investigate the mean squared errors of these estimators, and they found that (5.2.22) performs well. They refer to the criterion (5.2.18) as the PT_q -criterion (pre-test q criterion). This term will also be used in the remainder of this chapter.

In all of the above it was assumed that the value of σ^2 is known. Suppose this is not the case, but an estimator S^2 of σ^2 is available, where S^2 is distributed independently of X . Then (5.2.17) - (5.2.20) can still be used to select a model $J(X)$ by replacing σ^2 with S^2 in these expressions.

5.3 THE PT_q -CRITERION IN DISCRIMINANT ANALYSIS

In this section it is shown how the PT_q -criterion can be applied for variable selection in two-group discriminant analysis. Unfortunately, this requires rather restrictive assumptions to be made, viz. the feature variables

- (i) are independent,
- (ii) are normally distributed, and
- (iii) have the same variance.

These are the assumptions underlying derivation of the PT_q -criterion in Section 5.2.

The following notation is required. Let x_{ijl} be the l -th observation on the j -th feature variable in group i , where $i = 0,1$; $j = 1, \dots, k$ and $l = 1, \dots, n_i$, and let

$$\mathbf{x}_{ij} = [x_{ij1} \ x_{ij2} \ \dots \ x_{ijk}]'$$

If the training sample cases are selected randomly, the corresponding random vectors \mathbf{X}_{ij} are independently distributed with \mathbf{X}_{ij} having the k -variate $N(\mu_i, \Sigma)$ distribution. Assumptions (i) and (iii) above imply that $\Sigma = \sigma^2 \mathbf{I}_k$, with σ^2 the common variance of the feature variables. Let $\mathbf{Z}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} \mathbf{X}_{ij}$, $i = 0,1$. Then \mathbf{Z}_0 and \mathbf{Z}_1 are independent random vectors, with

$$\mathbf{Z}_i \sim N\left(\mu_i, \frac{\sigma^2}{n_i} \mathbf{I}_k\right), \quad i = 0,1. \quad (5.3.1)$$

Put $\mathbf{T} = \left(\frac{n_0 n_1}{n_0 + n_1} \right)^{\frac{1}{2}} (\mathbf{Z}_0 - \mathbf{Z}_1)$. Then,

$$\mathbf{T} \sim N(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_k), \quad (5.3.2)$$

where

$$\boldsymbol{\theta} = \left(\frac{n_0 n_1}{n_0 + n_1} \right)^{\frac{1}{2}} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1). \quad (5.3.3)$$

The distribution of the random vector \mathbf{T} is now the same as that of the random vector \mathbf{X} of Section 5.2, and therefore the PT_q -criterion can be applied to identify the elements of $\boldsymbol{\theta}$ that may be regarded as zero. From (5.3.3) it is clear that $\theta_j = 0$ implies $\mu_{0j} = \mu_{1j}$, i.e. these are the feature variables with respect to which the two groups do not differ. The subset $J(\mathbf{X})$ in (5.2.21), identified by applying the PT_q -criterion to the components of \mathbf{T} , therefore contains the indices of the variables that are selected for inclusion into the discriminant function.

A problem that remains before the PT_q -criterion can be applied, is that the value of σ^2 is unknown and has to be estimated from the available data. Since the random variables X_{ij} are independently $N(\mu_{ij}, \sigma^2)$ distributed for $i = 1, \dots, n_i$, it follows from standard theory that

$$\sum_{i=1}^{n_i} (X_{ij} - \bar{X}_{ij})^2 \sim \sigma^2 \chi_{n_i-1}^2, \quad (5.3.4)$$

independently for $i = 0, 1$, where $\bar{X}_{ij} = \frac{1}{n_i} \sum_{t=1}^{n_i} X_{ijt}$. Hence, defining

$$S^2 = \frac{1}{d} \sum_{i=0}^1 \sum_{j=1}^k \sum_{t=1}^{n_i} (X_{ijt} - \bar{X}_{ij})^2, \quad (5.3.5)$$

with $d = k(n_0 + n_1 - 2)$, it follows that S^2 is an unbiased estimator of σ^2 , and that $dS^2 \sim \sigma^2 \chi_d^2$, independent of \mathbf{T} .

Application of the PT_q -criterion for variable selection in a discriminant analysis context can therefore be summarised as follows:

1. Calculate the values

$$t_j = \left(\frac{n_0 n_1}{n_0 + n_1} \right)^{\frac{1}{2}} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} x_{0ij} - \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1ij} \right), \quad (5.3.6)$$

for $j = 1, \dots, k$, and

$$s^2 = \frac{1}{d} \sum_{i=0}^1 \sum_{j=1}^k \sum_{l=1}^{n_i} (x_{ijl} - \bar{x}_{ij})^2. \quad (5.3.7)$$

2. Let $z_1 < z_2 < \dots < z_k$ be the observed order statistics corresponding to $|t_1|, \dots, |t_k|$. Calculate the PT_q -criterion $r(t, q)$ defined in (5.2.18) for $q = 0, 1, \dots, k$, replacing σ in (5.2.17) by s .

3. Suppose \hat{q} minimises $r(t, q)$ over $\{0, 1, \dots, k\}$. Then the variables that are selected for inclusion into the discriminant function correspond to the indices for which $|t_i| > z_{\hat{q}}$. If $\hat{q} = k$, no variables are selected and the discriminant function contains only an intercept.

Derivation of the PT_q -criterion in Section 5.2 depends strongly on the assumptions stated at the start of this section. This somewhat limits the applicability of the criterion, since it can be expected that the performance of a selection technique employing the PT_q -criterion will be strongly affected by departures from these assumptions. Application of such a technique should therefore only be considered if the required assumptions are satisfied.

Variable selection in discriminant analysis using the PT_q -criterion aims at identifying the variables that best separate the two populations, i.e. it concentrates directly on the separatory rather than the allocatory aspect. However, since the feature variables are assumed independent, it is to be expected that insofar as PT_q -selection correctly identifies those feature variables that separate the populations well, it will also yield a classification rule with good allocatory properties, i.e. a low expected actual error rate. This is clearly illustrated in the discussion of the simulation study results in Section 5.5.

5.4 ERROR RATE ESTIMATION

In Section 5.3, pre-test variable selection in discriminant analysis using the PT_q -criterion was discussed. As mentioned in Chapter 4, an important and difficult

problem when a discriminant function is formed using a selected subset of the available feature variables, is accurate estimation of the post-selection actual error rate. This error rate gives an indication of the accuracy with which the linear discriminant function based on the selected subset will predict new cases. The following cross validation strategy for estimation of the post-selection error rate when using the PT_q -method to select variables for inclusion in a linear discriminant function, is proposed. The notation introduced in Section 4.3.2 is used.

Let X be the $(n \times k)$ matrix of observations on the feature variables and denote the data with the j -th observation (row) deleted by $X_{(j)}$. Let y be the n -dimensional vector of observations on the response variable and let $y_{(j)}$ denote the response vector with observation j deleted. The following procedure is applied to obtain an error rate estimator.

1. Apply the PT_q -selection procedure as described in Section 5.3, to the data in $X_{(j)}$ to select a subset of the k available feature variables.
2. Use the Anderson classification statistic (2.1.7) based only on the selected variables to classify the omitted case X_j , and record the 0-1 loss associated with this classification.
3. Average the loss over all n cases, and use the average loss as an error rate estimator.

It is important to note that the selection process is repeated for each deleted case, implying that a different set of variables may be selected with each different case being omitted. This is in line with the recommendations of Snapinn and Knoke (1989) and Ganeshanandam and Krzanowski (1990) that the leave-one-out step should precede the selection step to effectively reduce selection induced bias.

In preliminary simulation studies, it was found that the estimator proposed above is virtually unbiased, but has a fairly large variance, resulting in UMSE-values comparable to that of the CMV-estimator (cf. Chapter 4) which had much larger bias. In an attempt to reduce the variance of this error rate estimator, several ways of smoothing the 0-1 loss function were investigated. These will now briefly be discussed.

1. The normally smoothed version of the 0-1 loss function suggested by Snapinn and Knoke (1985) and used in the cross model validation technique described in Chapter 4, was used to obtain an error rate estimator for the PT_q -technique. Although this did reduce the variance of the error rate estimator, it was accompanied by an increase in the bias. This resulted in UMSE-values that were largely similar to those obtained when using the 0-1 loss function, and was therefore not considered to be an improvement.

2. Another option that was investigated, was using the posterior probability of misclassification of the omitted case as loss function. For a case, \mathbf{x}_j , from Π_0 , this probability is given by

$$\hat{\tau}_1(\mathbf{x}_j) = \frac{e^{-0.5D_{1j}^2}}{e^{-0.5D_{0j}^2} + e^{-0.5D_{1j}^2}}, \quad (5.4.1)$$

and for a case from Π_1 , by

$$\hat{\tau}_0(\mathbf{x}_j) = \frac{e^{-0.5D_{0j}^2}}{e^{-0.5D_{0j}^2} + e^{-0.5D_{1j}^2}}, \quad (5.4.2)$$

where $D_{ij}^2 = (\mathbf{x} - \bar{\mathbf{x}}_{ij})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_{ij})$, $i = 0, 1$; $j = 1, \dots, n_0 + n_1$.

Simulation studies that were carried out using this loss function, indicated that the resulting reduction in variance is once more not effective in reducing the UMSE of the error rate estimator, since its bias is again increased.

3. The loss function that was used in the cross model validation technique for logistic regression (cf. Section 4.6), was also applied in the PT_q -procedure. For a case \mathbf{x}_j , from Π_0 , the posterior probability of misclassification (5.4.1) is calculated, and the loss is obtained as follows:

$$L(\mathbf{x}_j) = \begin{cases} 0, & \text{if } \hat{\tau}_1(\mathbf{x}_j) < \min(\frac{1}{2}, 1/(1+D)) \\ 1, & \text{if } \hat{\tau}_1(\mathbf{x}_j) > \max(\frac{1}{2}, D/(1+D)) \\ \hat{\tau}_1(\mathbf{x}_j), & \text{if } \min(\frac{1}{2}, 1/(1+D)) \leq \hat{\tau}_1(\mathbf{x}_j) \leq \max(\frac{1}{2}, D/(1+D)), \end{cases} \quad (5.4.3)$$

where D is the sample Mahalanobis distance between the two populations based on the selected variables. Similar expressions are used for cases from group Π_1 , with $\hat{\tau}_0(\mathbf{x}_j)$ replacing $\hat{\tau}_1(\mathbf{x}_j)$.

The results obtained when implementing this loss function, are similar to that obtained by the previous two smoothing methods. The reduction in variance is again accompanied by an increase in bias, resulting in UMSE-values that are similar to that obtained when using a 0-1 loss function.

Based on the results of these initial simulation studies, it was decided to use the 0-1 loss function, since it yielded an estimator that has the lowest bias of all strategies

considered and UMSE-values similar to those obtained by the other strategies. This loss function was employed in the detailed Monte Carlo simulation study in which the selection and estimation performance of the PT_q -method is compared to that of the cross model validation technique. The results of this simulation study is reported in Section 5.5.

5.5 MONTE CARLO SIMULATION STUDY

To evaluate the performance of the PT_q -technique for the selection of variables for inclusion in a linear discriminant function, and for estimation of the resulting post-selection error rates, a Monte Carlo simulation study was undertaken. Since this technique is only applicable in the case of independent normal feature variables with equal variances, only such cases were included in the study. The cases NS11, NS21, NS31 and NS41, as well as the corresponding mixed and large sample cases (coded by replacing S in the codes for the small sample cases by M and L respectively), defined in Section 4.5.1.1, were included in the study. The selection and estimation performance of the PT_q -technique in these cases are compared to that of the cross model validation technique with F-based forward selection as inner criterion. The comparison of selection performance is done in terms of the expected post-selection actual error rate as well as the probability of correct selection (PCS). To judge estimation performance, the bias and unconditional mean squared errors of the error rate estimators were compared. These quantities were estimated for the PT_q -method by means of simulation, using 5000 repetitions. For each repetition a training data set was generated from the relevant normal distributions and a subset of variables was selected by applying the PT_q -criterion, as described in Section 5.3. The post-selection actual error rate associated with the selected subset, was calculated using (2.2.9). The error rate estimate proposed in Section 5.4 was also calculated. In order to estimate the bias and unconditional mean squared error of the error rate estimator, the difference and squared difference between the value of the error rate estimator and the post-selection actual error rate, were also calculated. The 5000 actual error rates were averaged to obtain the expected post-selection actual error rates, while the probability of correct selection was estimated by calculating the fraction of repetitions in which all the seemingly relevant variables and no seemingly irrelevant variables were selected. The bias of the PT_q -estimator was estimated by averaging the differences between the value of the error rate estimator and the post-selection actual error rate over the 5000 repetitions, i.e. $\hat{B}_{PT_q} = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\alpha}_i - \alpha_i^{act})$, where $\hat{\alpha}_i$ denotes the value of the error rate estimator of the PT_q -technique obtained for the i -th Monte Carlo repetition and α_i^{act} denotes the actual error rate (2.2.9) calculated for the i -th Monte Carlo repetition. The squared differences between the PT_q -estimator and the post-selection actual error rate were averaged to obtain an estimate of the unconditional mean squared error of

the PT_q -estimator, i.e. $\hat{U}_{PT_q} = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\alpha}_i - \alpha_i^{act})^2$. In the Appendix, Program 4 is given as an example of the Fortran program used in this simulation study.

The results of the simulation study were summarised by means of graphs. A representative selection of these graphs, displaying typical cases, is given in Figs. 5.1 - 5.4. In Fig. 5.1, graphs of the post-selection expected actual error rates are given, while Fig. 5.2 displays the PCS associated with the procedures. Fig. 5.3 contains graphs of the bias of the two error rate estimators, and graphs of the unconditional mean squared errors of the error rate estimators are given in Fig. 5.4.

5.5.1 SELECTION PERFORMANCE

The selection performance of the techniques is firstly evaluated. Two aspects are considered, viz. the post-selection expected actual error rate and the probability of correct selection associated with the techniques.

5.5.1.1 Expected Actual Error Rate

In all the cases considered, the post-selection expected actual error rate of the PT_q -technique is consistently slightly lower than that of the cross model validation procedure, except at very small values of Δ^2 ($\Delta^2 = 0,1$), where the error rates are approximately equal (see Fig. 5.1 for cases NS11, NS31, NM21 and NL41). The differences are generally larger in the small and mixed sample cases than in the large sample cases. This is an indication that a classification function based on variables selected by applying the PT_q -criterion, will in general perform better in terms of accurate classification of new cases.

5.5.1.2 Probability of Correct Selection (PCS)

The PT_q -technique consistently outperforms the CMV-technique with respect to the probability of selection all the seemingly relevant variables and no seemingly irrelevant variables. In all the cases considered, the PCS associated with PT_q -selection, is higher than that associated with CMV-selection. In cases NS11, NM11 and NL11 (see Fig. 5.2 for case NS11), the PT_q -procedure yields PCS-values in excess of 0.8, even at moderate values of Δ^2 ($\Delta^2 \geq 2$), while the PCS associated with the CMV-technique is in the region of 0.5 at the same values of Δ^2 . In cases NS21, NM21 and NL21 (see Fig. 5.2 for case NM21), the PCS-values were generally lower than in the previous cases, but increased quite sharply with Δ^2 . The PT_q -procedure once more outperformed the CMV-procedure. The difference in the performance of the two

techniques was largest for small sample cases. In cases NS31, NM31 and NL31 (see Fig. 5.2 for case NM31), the PCS associated with PT_q -selection, is again larger than 0.8 for large values of Δ^2 ($\Delta^2 \geq 6$). In these cases, the CMV-procedure yielded a maximum PCS of approximately 0.4. In cases NS41, NM41 and NL41 (see Fig. 5.2 for case NL41), the PCS associated with PT_q -selection, is again larger than that of the CMV-procedure, reaching a maximum value of 0.5 at $\Delta^2 = 9$, while the PCS associated with the CMV-procedure is close to 0 even at such a large separation. The PT_q -procedure is clearly superior with respect to selecting variables that best separate the two populations.

5.5.2 ESTIMATION PERFORMANCE

To evaluate the estimation accuracy of the three procedures, the bias and unconditional mean squared errors (UMSE) of the error rate estimators are compared.

5.5.2.1 Bias

When considering the bias of the error rate estimators, displayed in Fig. 5.3, it is clear that the PT_q -estimator is virtually unbiased, especially in small sample cases (see Fig. 5.3 for cases NS31 and NS41) and large sample cases (see Fig. 5.3 for NL21). In the mixed sample cases (see Fig. 5.3 for NM11), the PT_q -estimator is slightly biased at small values of Δ^2 , but much less so than the CMV-estimator.

5.5.2.2 Unconditional Mean Squared Error

A representative selection of graphs displaying the unconditional mean squared errors of the PT_q -estimator and the CMV-estimator, appears in Fig. 5.4. In the small sample cases (see Fig. 5.4 for cases NS31 and NS41), the UMSE of the CMV-estimator is lower (except at $\Delta^2 = 0$) than that of the PT_q -estimator. In the mixed sample cases (see Fig. 5.4 for case NM21), the performance varies: the PT_q -estimator performs better at small values of Δ^2 , but the CMV-estimator has lower UMSE at moderate to large values of Δ^2 ($\Delta^2 > 2$). In the large sample cases (see Fig. 5.4 for case NL11), the differences in the unconditional mean squared errors are small. The CMV-estimator yields slightly lower values than the PT_q -estimator.

In conclusion, if the necessary assumptions underlying the PT_q -method are satisfied, and especially if the main aim is to select variables which separate the populations well, the PT_q -selection technique is recommended. The classification performance of a rule selected by means of the PT_q -criterion, will also be slightly better than that of its

competitors (the PT_q - method performs better than the CMV- method, which outperformed the other methods considered in Section 4.5). In terms of estimation accuracy, the proposed cross validation based error rate estimator performs the best of all the estimators considered in this section as well as in Section 4.5 with respect to bias, and yields slightly larger UMSE - values than the CMV- estimator (which outperformed the other two estimators considered in Section 4.5) only in some of the cases considered.

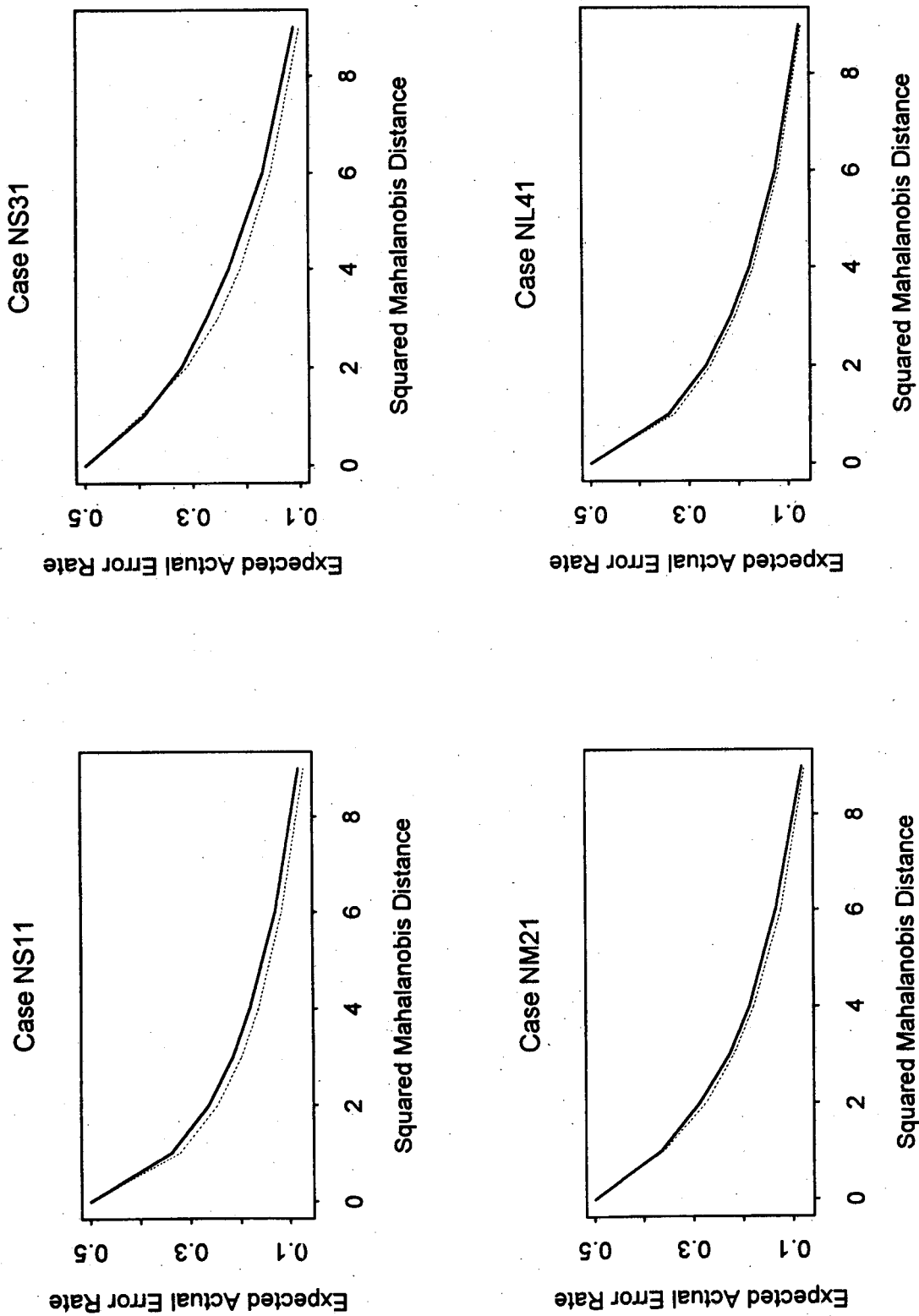


FIG. 5.1: EXPECTED ACTUAL ERROR RATE, UNCORRELATED NORMAL DATA

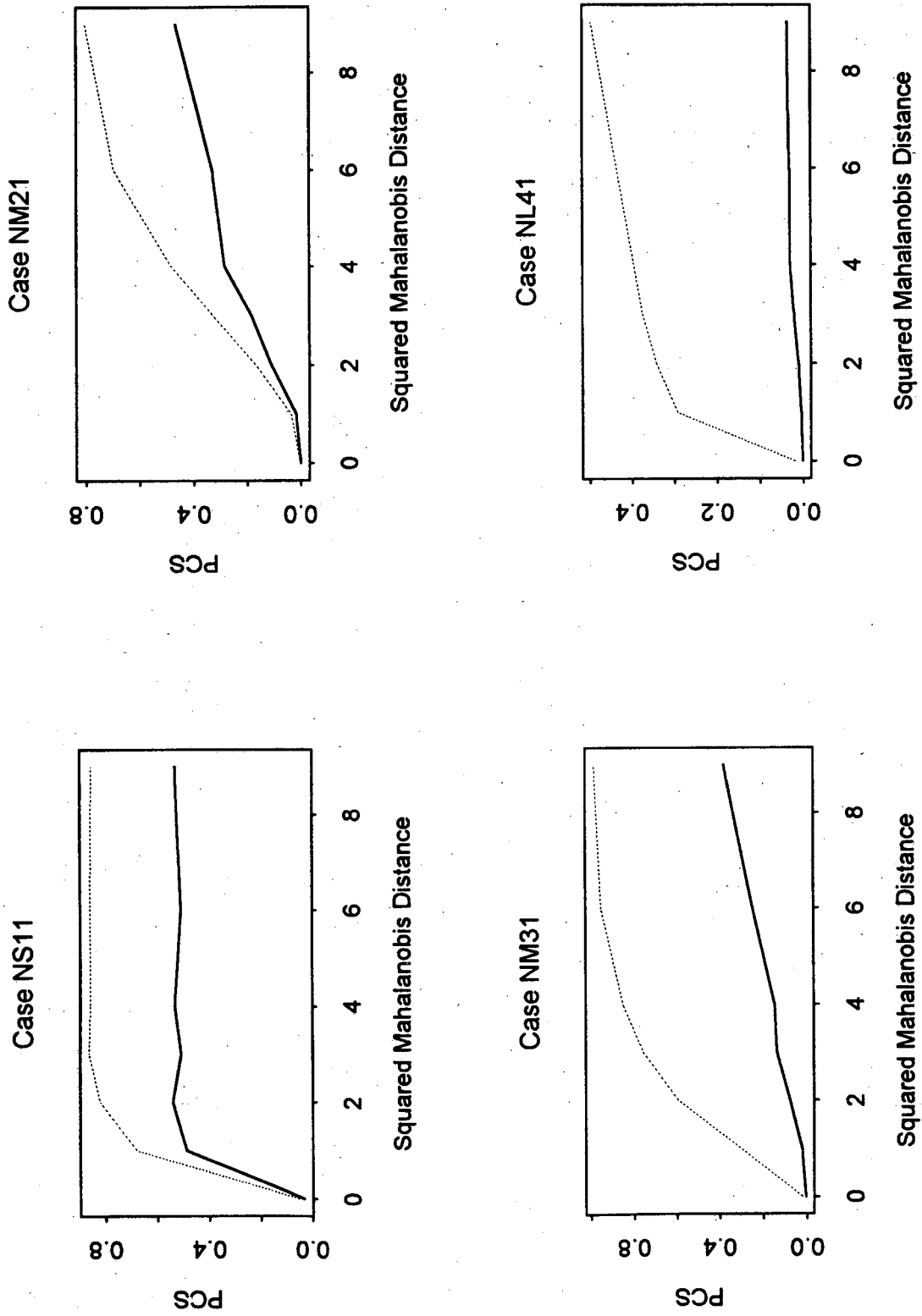


FIG. 5.2: PROBABILITY OF CORRECT SELECTION, UNCORRELATED NORMAL DATA

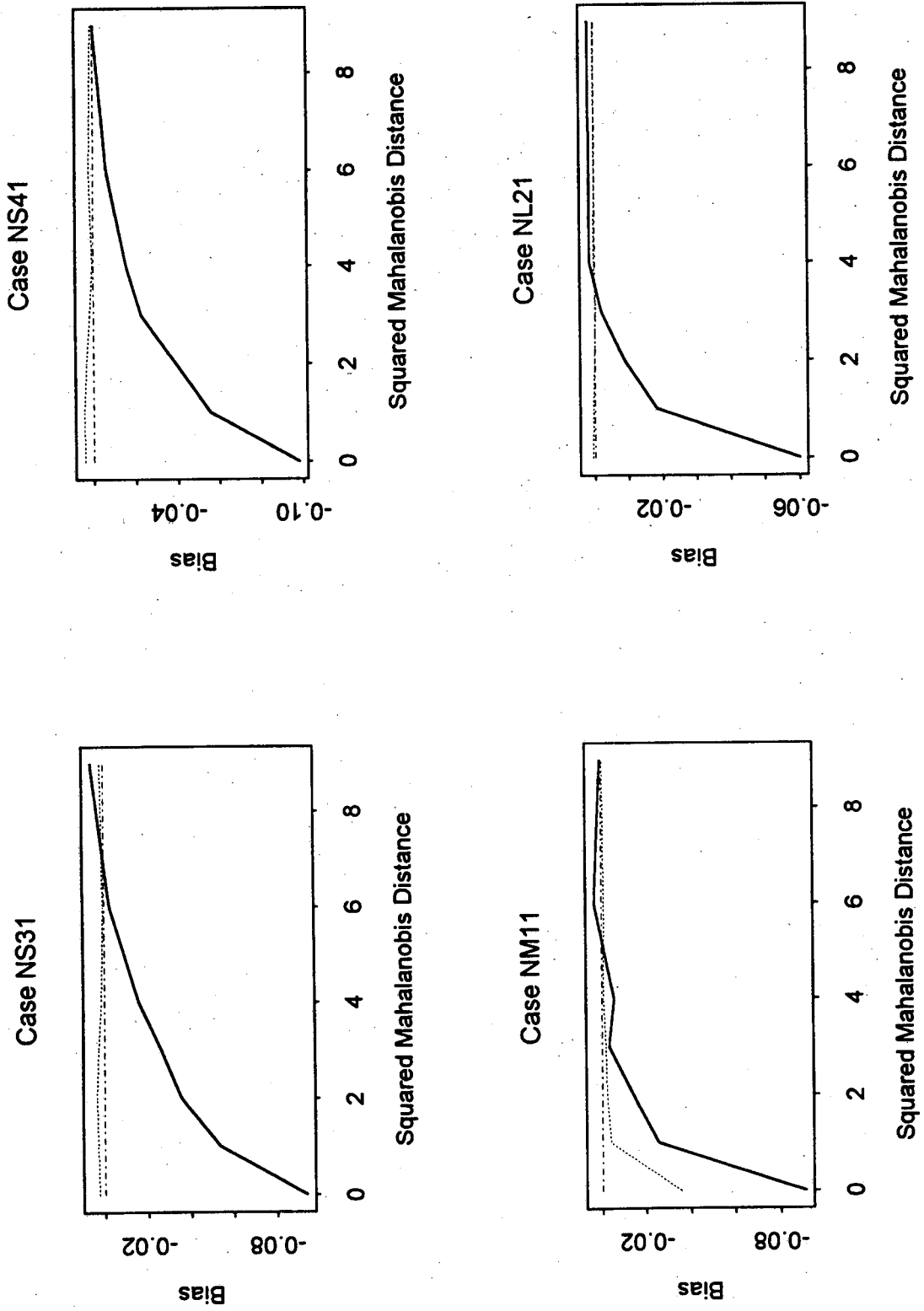
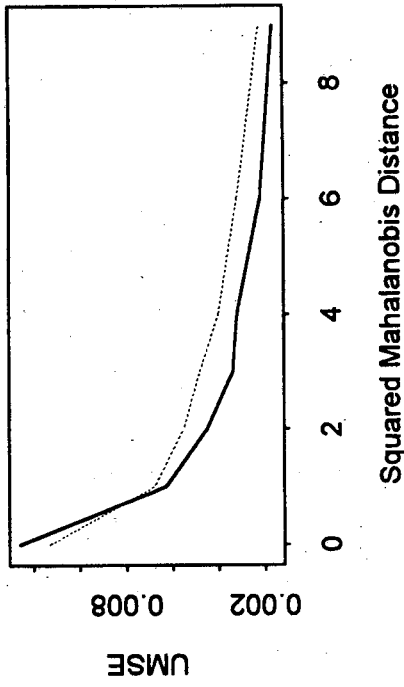
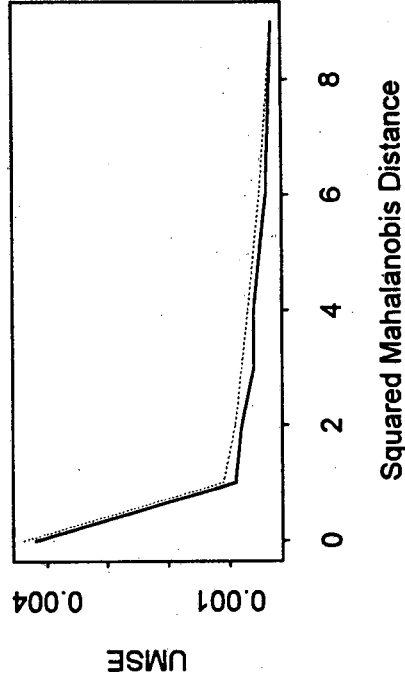


FIG. 5.3: BIAS OF ERROR RATE ESTIMATORS, UNCORRELATED NORMAL DATA

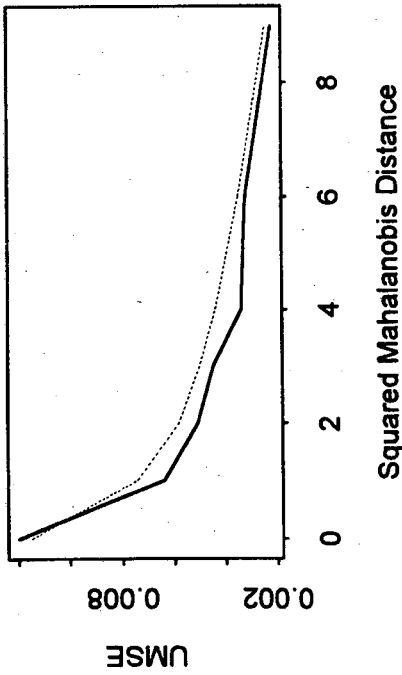
Case NS41



Case NL11



Case NS31



Case NM21

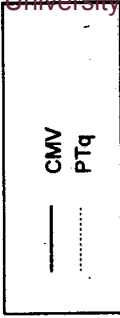
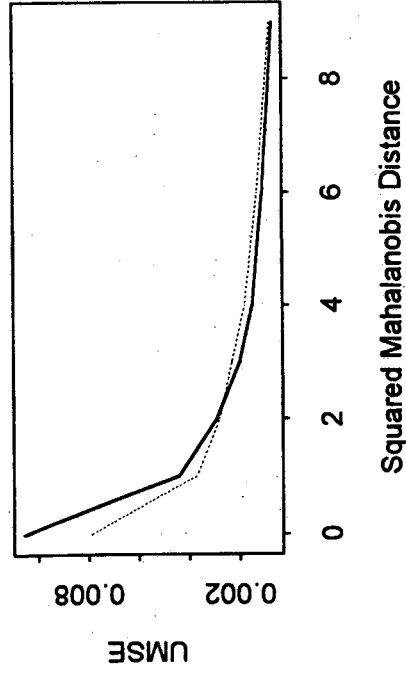


FIG. 5.4: UNCONDITIONAL MEAN SQUARED ERROR OF ERROR RATE ESTIMATORS, UNCORRELATED NORMAL DATA

CHAPTER 6

SUMMARY AND DIRECTIONS FOR FURTHER RESEARCH

Various aspects of the pre- and post-selection classification performance of the linear discriminant function and the logistic discriminant function were studied in this thesis. The main conclusions emanating from this study are summarised in this chapter, and a number of directions for future research are indicated.

The results of the simulation study reported in Chapter 2 show that the pre-selection classification performance of the linear discriminant function is better than that of the logistic discriminant function if the feature variables have a normal or a double exponential distribution, while the reverse is true for lognormal feature variables. It was also found that increasing the ratio of the number of variables to the training sample size, favoured the linear discriminant function. From these results it would seem that the linear discriminant function is preferable for data from symmetric distributions, while logistic discrimination should generally be the method of choice for data from skew distributions. Further examples of skew and symmetric distributions could be investigated to add weight to this conclusion.

A part of Chapter 2 was devoted to a comparison of the fully polychotomous and individualised binary approaches to logistic regression when more than two groups are available. The fully polychotomous approach generally performed better, except for a few lognormal cases. It was also found in Chapter 2 that logistic discrimination suffers from a serious disadvantage which limits applicability of the technique, viz. the non-existence of the maximum likelihood estimates of the logistic regression coefficients when the populations are well separated.

As an introduction to the investigation of post-selection classification performance, the effect of the number of variables in a classification function on its actual error rate, was studied from various points of view in Chapter 3. The correlation structure amongst the feature variables was found to have a profound influence on this effect. Consequently, variables should not be considered singly when a decision has to be made regarding their inclusion into or exclusion from a classification function. A distinction was also made between separatory and allocatory selection criteria, and the actual error rates resulting when such criteria are used to select a pre-specified number of feature variables, were investigated. It became clear that in such cases there is little to choose between these two types of selection criteria. Since applying a separatory criterion is typically much simpler and these criteria are more readily available than allocatory criteria, use of a separatory criterion to choose between models of the same dimension can be recommended. However, if the classification performance of the rule being constructed is of prime importance, the choice of a final model dimension should ideally be based on an allocatory criterion, i.e. an error rate estimate.

The findings in Chapter 3 were used in Chapter 4 to develop a new selection technique for discriminant analysis and logistic regression, viz. cross model validation. One of the main advantages of this technique is that it combines variable selection and estimation of the accuracy of the resulting classification function, rather than considering these two closely related problems separately. An extensive simulation study was undertaken to investigate the properties of cross model validation, and it was found to perform well with respect to selection and estimation. In addition, the two examples discussed in Chapter 4 showed that application of the technique is fairly straightforward and that it provides the user with useful information regarding the estimated classification accuracy associated with each possible model dimension.

There are a number of aspects of cross model validation that require further research. These include its application in the case of more than two groups and in cases where the assumption of homoscedasticity is not valid.

Chapter 5 was devoted to an investigation into a pre-test type selection criterion, originally proposed in a non-classification context. It was shown how this criterion can be adapted for application in discriminant analysis. Simulation was used to study the properties of the criterion, and it was found to perform well in the rather restricted setting of uncorrelated normally distributed feature variables. Further research can be directed at adapting the procedure for application in other settings.

APPENDIX

PROGRAM 1

C IN THIS PROGRAM MONTE CARLO SIMULATION IS USED TO COMPARE
 C THE PERFORMANCE OF THE LINEAR DISCRIMINANT FUNCTION AND THE
 C LOGISTIC DISCRIMINANT FUNCTION IN THE CASE OF THREE GROUPS.
 C THE EXPECTED ACTUAL ERROR RATES OF THE PROCEDURES ARE COMPARED FOR
 C TRAINING DATA GENERATED FROM DOUBLE EXPONENTIAL POPULATIONS.
 C PROVISION IS MADE FOR EQUI-CORRELATED FEATURE VARIABLES.

C

C PARAMETERS :

C NATTRS=THE NUMBER OF FEATURE VARIABLES

C N1/2/3=THE TRAINING SAMPLE SIZE FROM GROUP 1/2/3

C NDATA=N1+N2+N3 : THE TOTAL SAMPLE SIZE

C NMC=THE NUMBER OF MONTE CARLO REPETITIONS

C NB=THE NUMBER OF CASES FROM EACH GROUP GENERATED TO ESTIMATE
 C THE ACTUAL ERROR RATES

C KLASS=THE NUMBER OF GROUPS

C RHO=THE CORRELATION BETWEEN THE NORMAL FEATURE VARIABLES THAT IS
 C REQUIRED TO ENSURE A GIVEN CORRELATION BETWEEN THE DOUBLE
 C EXPONENTIAL FEATURE VARIABLES

C

C THE FOLLOWING IMSL-SUBROUTINES ARE USED IN THE MAIN PROGRAM:

C 1. DLINDS: FINDS THE INVERSE OF A GIVEN COVARIANCE MATRIX

C 2. DCHFAC: FINDS THE CHOLESKY DECOMPOSITION OF A GIVEN MATRIX

C 3. DRNMVN: GENERATES VALUES FROM A MULTIVARIATE NORMAL DISTRIBUTION

C 4. DNORDF: CALCULATES THE CUMULATIVE DISTRIBUTION FUNCTION OF THE
 C STANDARD NORMAL DISTRIBUTION

C

IMPLICIT DOUBLE PRECISION (A-H,O-Z)

PARAMETER (NATTRS=10,N1=25,N2=25,N3=25,NDATA=N1+N2+N3,

&N12=N1+N2,NATP1=NATTRS+1,NMC=1000,NB=5000,KLASS=3,RHO=0.905D0)

DIMENSION AMU(3,NATTRS),SIGMAM(NATTRS,NATTRS)

DIMENSION U1(N1,NATTRS),U2(N2,NATTRS),U3(N3,NATTRS)

DIMENSION RNX1(N1,NATTRS),RNX2(N2,NATTRS),RNX3(N3,NATTRS)

DIMENSION X1(N1,NATTRS),X2(N2,NATTRS),X3(N3,NATTRS)

DIMENSION RSIG(NATTRS,NATTRS),RESP(NDATA)

DIMENSION SIGINV(NATTRS,NATTRS),BETA(NATP1,KLASS-1)

DIMENSION XX(NDATA,NATP1),XPOLY(NDATA,NATTRS)

DIMENSION ACTDA(1000,10),ACTLR(1000,10),ADA(10),ALR(10)

DIMENSION COEF12(NATP1,1),COEF13(NATP1,1)

DIMENSION ICLASS(NDATA),NOCONV(10)

EXTERNAL DLINDS,DCHFAC,DRNMVN,DNORDF

CHARACTER*70 FOUT1,FOUT2,FOUT3

FOUT1='/da.d'

FOUT2='/lr.d'

FOUT3='/dalr.d'

NITER=100

DSMALL=0.1D0

```
C
C PROVIDE APPROPRIATE VALUES FOR THE COMMON COVARIANCE MATRIX AND
C THE MEAN VECTOR OF GROUP 1
C
DO 2 I=1,NATTRS
SIGMAM(I,I)=1.0D0
DO 1 J=1,NATTRS
IF (J.NE.I) SIGMAM(I,J)=RHO
1 CONTINUE
2 CONTINUE

DO 5 I=1,NATTRS
AMU(I,I)=0.0D0
5 CONTINUE
CALL DLINDS(NATTRS,SIGMAM,NATTRS,SIGINV,NATTRS)
C
C COMPUTE THE CHOLESKY DECOMPOSITION OF THE COVARIANCE MATRIX.
C THIS IS LATER REQUIRED TO GENERATE NORMAL VALUES
C
TOL=1.0D2*DMACH(4)
CALL DCHFAC(NATTRS,SIGMAM,NATTRS,TOL,IRANK,RSIG,NATTRS)

S11=SIGINV(1,1)
S12=SIGINV(1,2)
S122=S12*S12
T1=(NATTRS-1.0D0)*S122/(S11*S11)
T2=(NATTRS-2.0D0)*S12/S11
C
C THE VECTOR ICLASS CONTAINS THE RESPONSE VARIABLE INDICATING GROUP
C MEMBERSHIP. IT IS REQUIRED AS INPUT FOR SUBROUTINE POLY.
C
DO 8 I=1,N1
ICLASS(I)=0
8 CONTINUE
DO 9 I=N1+1,N12
ICLASS(I)=1
9 CONTINUE
DO 10 I=N12+1,NDATA
ICLASS(I)=2
10 CONTINUE
C
C WE COME TO THE LOOP THAT ENABLES US TO LOOK AT DIFFERENT
C SEPARATIONS BETWEEN THE GROUPS
C
DO 500 IS=0,8
IF (IS.LE.4) D2=0.5D0*IS
IF (IS.EQ.5) D2=3.0D0
IF (IS.EQ.6) D2=4.0D0
IF (IS.EQ.7) D2=6.0D0
IF (IS.EQ.8) D2=9.0D0
C
C SET UP THE MEAN VECTORS OF GROUPS 2 AND 3 TO ENSURE THAT THE
C MAHALANOBIS DISTANCE BETWEEN ANY TWO OF THE GROUPS IS EQUAL TO D2
C
D1=DSQRT(D2/S11)
```



```

T3=3.0D0*D1*D1*S11/4.0D0
B=DSQRT(T3/((NATTRS-1.0D0)*S11*(1.0D0-T1+T2)))
A=D1/2.0D0-(NATTRS-1.0D0)*S12*B/S11
DO 12 J=1,NATTRS
AMU(2,J)=0.0D0
AMU(3,J)=B
12 CONTINUE
AMU(2,1)=D1
AMU(3,1)=A
NOCONV(IS+1)=0
C
C THE MONTE CARLO LOOP STARTS AT STATEMENT 14, WITH MC AS COUNTER
C
C FIRST, GENERATE THE TRAINING DATA SETS FROM THE MULTIVARIATE NORMAL
C DISTRIBUTION AND TRANSFORM TO THE REQUIRED DOUBLE EXPONENTIAL
C DISTRIBUTION
C
MC=1
14 CALL DRNMVN(N1,NATTRS,RSIG,NATTRS,X1,N1)
CALL DRNMVN(N2,NATTRS,RSIG,NATTRS,X2,N2)
CALL DRNMVN(N3,NATTRS,RSIG,NATTRS,X3,N3)
DO 18 J=1,NATTRS
DO 15 I=1,N1
U1(I,J)=DNORDF(X1(I,J)/(DSQRT(SIGMAM(J,J))))
RNX1(I,J)=GINV(U1(I,J))+AMU(1,J)
15 CONTINUE
DO 16 I=1,N2
U2(I,J)=DNORDF(X2(I,J)/(DSQRT(SIGMAM(J,J))))
RNX2(I,J)=GINV(U2(I,J))+AMU(2,J)
16 CONTINUE
DO 17 I=1,N3
U3(I,J)=DNORDF(X3(I,J)/(DSQRT(SIGMAM(J,J))))
RNX3(I,J)=GINV(U3(I,J))+AMU(3,J)
17 CONTINUE
18 CONTINUE
C
C RESP IS THE RESPONSE VARIABLE INDICATING GROUP MEMBERSHIP
C
DO 25 I=1,N1
RESP(I)=0.0D0
25 CONTINUE
DO 28 I=N1+1,N12
RESP(I)=1.0D0
28 CONTINUE
DO 30 I=N12+1,NDATA
RESP(I)=2.0D0
30 CONTINUE
C
C A MATRIX XX(NDATA,NATTRS+1) IS FORMED. THE FIRST NATTRS COLUMNS CONTAIN
C THE FEATURE VARIABLES AND COLUMN NATP1=NATTRS+1 CONTAINS THE RESPONSE
C VARIABLE INDICATING GROUP MEMBERSHIP.
C XPOLY(NDATA,NATTRS) IS THE XX-MATRIX WITHOUT THE LAST COLUMN CONTAINING
C THE RESPONSE VARIABLE.
C
DO 45 J=1,NATTRS

```

```

DO 35 I=1,N1
XX(I,J)=RNX1(I,J)
XPOLY(I,J)=XX(I,J)
35 CONTINUE
DO 38 I=1,N2
XX(N1+I,J)=RNX2(I,J)
XPOLY(N1+I,J)=XX(N1+I,J)
38 CONTINUE
DO 40 I=1,N3
XX(N12+I,J)=RNX3(I,J)
XPOLY(N12+I,J)=XX(N12+I,J)
40 CONTINUE
45 CONTINUE

DO 50 I=1,N1
XX(I,NATP1)=RESP(I)
50 CONTINUE
DO 53 I=1,N2
XX(N1+I,NATP1)=RESP(N1+I)
53 CONTINUE
DO 55 I=1,N3
XX(N12+I,NATP1)=RESP(N12+I)
55 CONTINUE
C
C SUBROUTINE POLY IS CALLED TO OBTAIN THE MAXIMUM LIKELIHOOD ESTIMATES OF
C THE LOGISTIC REGRESSION COEFFICIENTS (BETA). IF THE ITERATIVE PROCESS
C FOR CALCULATION OF THE COEFFICIENTS DOES NOT CONVERGE, THE WHOLE CASE IS
C EXCLUDED FROM THE ANALYSIS AND A NEW DATA SET IS GENERATED. IW IS USED
C AS AN INDICATOR FOR THIS PURPOSE. THE VECTOR NOCONV IS USED TO KEEP
C RECORD OF THE NUMBER OF TIMES THAT THIS HAPPENS AT EACH VALUE OF IS
C (CORRESPONDING TO DIFFERENT VALUES OF THE SQUARED MAHALANOBIS DISTANCE
C BETWEEN THE POPULATIONS).
C
IW=0
CALL POLY(IW,ICLASS,NITER,NDATA,KLASS,NATTRS,DSMALL,XPOLY,BETA)
IF (IW.EQ.1) THEN
  NOCONV(IS+1)=NOCONV(IS+1)+1
  GOTO 14
ENDIF

DO 90 J=1,NATP1
COEF12(J,1)=BETA(J,1)
COEF13(J,1)=BETA(J,2)
90 CONTINUE
C
C SUBROUTINE ERROR CALCULATES THE ACTUAL ERROR RATE ASSOCIATED WITH
C BOTH THE LINEAR DISCRIMINANT FUNCTION (ACTD) AND THE LOGISTIC
C DISCRIMINANT FUNCTION (ACTL).
C
CALL ERROR(AMU,SIGMAM,RSIG,XX,COEF12,COEF13,ACTD,ACTL)

ACTDA(MC,IS+1)=ACTD
ACTLR(MC,IS+1)=ACTL

MC=MC+1

```

IF (MC.LE.NMC) GOTO 14

500 CONTINUE

C

C THIS IS THE END OF THE MONTE CARLO SIMULATION LOOP

C

C THE ACTUAL ERROR RATES ARE ACCUMULATED IN ADA (FOR DISCRIMINANT ANALYSIS)

C AND ALR (FOR LOGISTIC REGRESSION) RESPECTIVELY, AND AVERAGES OVER ALL THE

C MONTE CARLO REPETITIONS ARE TAKEN TO OBTAIN ESTIMATES OF THE EXPECTED

C ACTUAL ERROR RATES

C

DO 502 J=1,IS

ADA(J)=0.0D0

ALR(J)=0.0D0

DO 501 I=1,NMC

ADA(J)=ADA(J)+ACTDA(I,J)

ALR(J)=ALR(J)+ACTLR(I,J)

501 CONTINUE

ADA(J)=ADA(J)/NMC

ALR(J)=ALR(J)/NMC

502 CONTINUE

OPEN(1,FILE=FOUT1,ACCESS='APPEND')

OPEN(2,FILE=FOUT2,ACCESS='APPEND')

OPEN(3,FILE=FOUT3,ACCESS='APPEND')

DO 510 I=1,NMC

WRITE(1,620) (ACTDA(I,J),J=1,IS)

WRITE(2,620) (ACTLR(I,J),J=1,IS)

510 CONTINUE

WRITE(3,*)

WRITE(3,630) (ADA(J),J=1,IS)

WRITE(3,630) (ALR(J),J=1,IS)

WRITE(3,*)

WRITE(3,640) (NOCONV(J),J=1,IS)

CLOSE(1)

CLOSE(2)

CLOSE(3)

620 FORMAT(10(F10.5,2X))

630 FORMAT(7(F10.5,2X))

640 FORMAT(10I5)

1000 STOP

END

SUBROUTINE ERROR(AMU,SIGMAM,RSIG,XX,COEF12,COEF13,ACTD,ACTL)

C

C SUBROUTINE ERROR USES SIMULATION TO CALCULATE THE ACTUAL ERROR RATES OF

C BOTH THE LINEAR DISCRIMINANT FUNCTION (ACTD) AND THE LOGISTIC

C DISCRIMINANT FUNCTION (ACTL).

C

C A LARGE NUMBER (NB) OF CASES FROM EACH GROUP ARE GENERATED.

C TO ESTIMATE THE ERROR RATE OF THE LINEAR DISCRIMINANT FUNCTION,

C THE SUBROUTINE WDIST IS USED TO CALCULATE THE SQUARED MAHALANOBIS

C DISTANCE BETWEEN EACH GENERATED CASE AND EACH OF THE THREE GROUP MEANS.

```

C THE CASE IS THEN CLASSIFIED INTO THE GROUP YIELDING THE MINIMUM DISTANCE.
C
C TO ESTIMATE THE ERROR RATE OF THE LOGISTIC DISCRIMINANT FUNCTION THE
C POSTERIOR PROBABILITY OF EACH CASE TO BELONG TO EACH OF THE THREE
C GROUPS ARE CALCULATED. THE CASE IS THEN CLASSIFIED INTO THE GROUP
C YIELDING THE MAXIMUM POSTERIOR PROBABILITY.
C
C INPUT : AMU=THE MATRIX CONTAINING THE MEANS OF THE THREE GROUPS
C SIGMAM=THE COMMON COVARIANCE MATRIX
C RSIG=THE MATRIX OBTAINED FROM THE CHOLESKY DECOMPOSITION
C OF THE COVARIANCE MATRIX
C XX=THE DATA MATRIX
C COEF12=LOGISTIC REGRESSION COEFFICIENTS FOR GROUPS 1 AND 2
C COEF13=LOGISTIC REGRESSION COEFFICIENTS FOR GROUPS 1 AND 3
C OUTPUT : ACTD/ACTL=THE ACTUAL ERROR RATES OF DA/LR
C
C IMPLICIT DOUBLE PRECISION (A-H,O-Z)
C PARAMETER (NATTRS=10,N1=25,N2=25,N3=25,NDATA=N1+N2+N3,
C &N12=N1+N2,NATP1=NATTRS+1,NMC=1000,NB=5000,KLASS=3,RHO=0.905D0)
C DIMENSION XX(NDATA,NATTRS+1),S(NATTRS,NATTRS),SINV(NATTRS,NATTRS)
C DIMENSION XM1(NATTRS),XM2(NATTRS),XM3(NATTRS),XV(NATTRS)
C DIMENSION AMU(3,NATTRS),XB(NB,NATTRS),RSIG(NATTRS,NATTRS)
C DIMENSION COEF12(NATP1,1),COEF13(NATP1,1),SIGMAM(NATTRS,NATTRS)
C DIMENSION U1(NATTRS)
C
C CALCULATE THE SAMPLE GROUP MEANS, THE POOLED COVARIANCE MATRIX AND
C ITS INVERSE
C
C CALL AVGVARV(XX,S,SINV,XM1,XM2,XM3)
C SUMADA1=0.0D0
C SUMADA2=0.0D0
C SUMADA3=0.0D0
C SUMALR1=0.0D0
C SUMALR2=0.0D0
C SUMALR3=0.0D0
C
C NB CASES ARE GENERATED FROM GROUP1 AND CLASSIFIED USING THE LINEAR
C DISCRIMINANT FUNCTION AND THE LOGISTIC DISCRIMINANT FUNCTION.
C THE NUMBER OF MISCLASSIFIED CASES FOR GROUP1 FOR BOTH DA (SUMADA1)
C AND LR(SUMALR1) ARE DETERMINED.
C
C CALL DRNMVN(NB,NATTRS,RSIG,NATTRS,XB,NB)
C DO 50 IB=1,NB
C DO 5 J=1,NATTRS
C U1(J)=DNORDF(XB(IB,J)/(DSQRT(SIGMAM(J,J))))
C XV(J)=GINV(U1(J))+AMU(1,J)
5 CONTINUE
C CALL WDIST(XM1,XM2,XM3,XV,SINV,D1,D2,D3)
C AMIN=D1
C IF (D2.LT.AMIN) AMIN=D2
C IF (D3.LT.AMIN) AMIN=D3
C IF (DABS(AMIN-D1).GT.0.000001D0) SUMADA1=SUMADA1+1.0D0
C
C SUM1=COEF12(1,1)
C DO 20 J=1,NATTRS

```

```

SUM1=SUM1+(XV(J)*COEF12(J+1,1))
20 CONTINUE
SUM2=COEF13(1,1)
DO 25 J=1,NATTRS
SUM2=SUM2+(XV(J)*COEF13(J+1,1))
25 CONTINUE
EPOWER1=DEXP(SUM1)
EPOWER2=DEXP(SUM2)
DENOM=1.0D0+EPOWER1+EPOWER2
POST1=1.0D0/DENOM
POST2=EPOWER1/DENOM
POST3=EPOWER2/DENOM
AMAX=POST1
IF (POST2.GT.AMAX) AMAX=POST2
IF (POST3.GT.AMAX) AMAX=POST3
IF (DABS(AMAX-POST1).GT.0.000001D0) SUMALR1=SUMALR1+1.0D0
50 CONTINUE
C
C NB CASES ARE GENERATED FROM GROUP2 AND CLASSIFIED USING THE LINEAR
C DISCRIMINANT FUNCTION AND THE LOGISTIC DISCRIMINANT FUNCTION.
C THE NUMBER OF MISCLASSIFIED CASES FOR GROUP2 FOR BOTH DA (SUMADA2)
C AND LR (SUMALR2) ARE DETERMINED.
C
CALL DRNMVN(NB,NATTRS,RSIG,NATTRS,XB,NB)
DO 90 IB=1,NB
DO 55 J=1,NATTRS
U1(J)=DNORDF(XB(IB,J)/(DSQRT(SIGMAM(J,J))))
XV(J)=GINV(U1(J))+AMU(2,J)
55 CONTINUE
CALL WDIST(XM1,XM2,XM3,XV,SINV,D1,D2,D3)
AMIN=D1
IF (D2.LT.AMIN) AMIN=D2
IF (D3.LT.AMIN) AMIN=D3
IF (DABS(AMIN-D2).GT.0.000001D0) SUMADA2=SUMADA2+1.0D0
SUM1=COEF12(1,1)
DO 60 J=1,NATTRS
SUM1=SUM1+(XV(J)*COEF12(J+1,1))
60 CONTINUE
SUM2=COEF13(1,1)
DO 65 J=1,NATTRS
SUM2=SUM2+(XV(J)*COEF13(J+1,1))
65 CONTINUE
EPOWER1=DEXP(SUM1)
EPOWER2=DEXP(SUM2)
DENOM=1.0D0+EPOWER1+EPOWER2
POST1=1.0D0/DENOM
POST2=EPOWER1/DENOM
POST3=EPOWER2/DENOM
AMAX=POST1
IF (POST2.GT.AMAX) AMAX=POST2
IF (POST3.GT.AMAX) AMAX=POST3
IF (DABS(AMAX-POST2).GT.0.000001D0) SUMALR2=SUMALR2+1.0D0
90 CONTINUE
C
C NB CASES ARE GENERATED FROM GROUP3 AND CLASSIFIED USING THE LINEAR

```

C DISCRIMINANT FUNCTION AND THE LOGISTIC DISCRIMINANT FUNCTION.
 C THE NUMBER OF MISCLASSIFIED CASES FOR GROUP3 FOR BOTH DA (SUMADA3)
 C AND LR (SUMALR3) ARE DETERMINED.

```

C
CALL DRNMVN(NB,NATTRS,RSIG,NATTRS,XB,NB)
DO 140 IB=1,NB
DO 95 J=1,NATTRS
U1(J)=DNORDF(XB(IB,J)/(DSQRT(SIGMAM(J,J))))
XV(J)=GINV(U1(J))+AMU(3,J)
95 CONTINUE
CALL WDIST(XM1,XM2,XM3,XV,SINV,D1,D2,D3)
AMIN=D1
IF (D2.LT.AMIN) AMIN=D2
IF (D3.LT.AMIN) AMIN=D3
IF (DABS(AMIN-D3).GT.0.000001D0) SUMADA3=SUMADA3+1.0D0
SUM1=COEF12(1,1)
DO 100 J=1,NATTRS
SUM1=SUM1+(XV(J)*COEF12(J+1,1))
100 CONTINUE
SUM2=COEF13(1,1)
DO 105 J=1,NATTRS
SUM2=SUM2+(XV(J)*COEF13(J+1,1))
105 CONTINUE
EPOWER1=DEXP(SUM1)
EPOWER2=DEXP(SUM2)
DENOM=1.0D0+EPOWER1+EPOWER2
POST1=1.0D0/DENOM
POST2=EPOWER1/DENOM
POST3=EPOWER2/DENOM
AMAX=POST1
IF (POST2.GT.AMAX) AMAX=POST2
IF (POST3.GT.AMAX) AMAX=POST3
IF (DABS(AMAX-POST3).GT.0.000001D0) SUMALR3=SUMALR3+1.0D0
140 CONTINUE

ACTD=(SUMADA1+ SUMADA2+ SUMADA3)/(3.0D0*NB)
ACTL=(SUMALR1+SUMALR2+ SUMALR3)/(3.0D0*NB)
RETURN
END

```

SUBROUTINE WDIST(XM1,XM2,XM3,XV,SINV,D1,D2,D3)

```

C
C THIS SUBROUTINE CALCULATES THE DISTANCE OF A SPECIFIC DATA CASE FROM THE
C SAMPLE MEAN OF EACH OF THE THREE GROUPS (D1, D2 AND D3 RESPECTIVELY).
C THESE DISTANCES ARE THEN USED TO CLASSIFY THE DATA CASE INTO ONE OF THE
C THREE GROUPS.
C INPUT : XM1=THE MEAN OF GROUP1
C         XM2=THE MEAN OF GROUP2
C         XM3=THE MEAN OF GROUP3
C         SINV=THE INVERSE OF THE POOLED COVARIANCE MATRIX
C         XV=THE CASE TO BE CLASSIFIED
C OUTPUT : D1=THE SQUARED MAHALANOBIS DISTANCE BETWEEN CASE XV AND THE
C          MEAN OF GROUP1
C          D2=THE SQUARED MAHALANOBIS DISTANCE BETWEEN CASE XV AND THE

```

```

C      MEAN OF GROUP2
C      D3=THE SQUARED MAHALANOBIS DISTANCE BETWEEN CASE XV AND THE
C      MEAN OF GROUP3
C

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (NATTRS=10,N1=25,N2=25,N3=25,NDATA=N1+N2+N3,
&N12=N1+N2,NATP1=NATTRS+1,NMC=1000,NB=5000,KLASS=3,RHO=0.905D0)
DIMENSION XV(NATTRS),SINV(NATTRS,NATTRS)
DIMENSION XM1(NATTRS),XM2(NATTRS),XM3(NATTRS)
SUM1=0.0D0
SUM2=0.0D0
SUM3=0.0D0
DO 95 I1=1,NATTRS
DO 90 I2=1,NATTRS
V1=XV(I1)-XM1(I1)
V2=XV(I2)-XM1(I2)
V3=XV(I1)-XM2(I1)
V4=XV(I2)-XM2(I2)
V5=XV(I1)-XM3(I1)
V6=XV(I2)-XM3(I2)
SUM1=SUM1+V1*SINV(I1,I2)*V2
SUM2=SUM2+V3*SINV(I1,I2)*V4
SUM3=SUM3+V5*SINV(I1,I2)*V6
90 CONTINUE
95 CONTINUE
D1=SUM1
D2=SUM2
D3=SUM3
RETURN
END

```

```

SUBROUTINE AVGVARV(XX,S,SINV,XM1,XM2,XM3)

```

```

C
C THIS SUBROUTINE CALCULATES THE GROUP MEANS (XM1, XM2 AND XM3) AND THE
C POOLED COVARIANCE MATRIX (S) AND ITS INVERSE (SINV). THE IMSL
C SUBROUTINE DCORVC IS USED FOR THIS PURPOSE.
C

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (NATTRS=10,N1=25,N2=25,N3=25,NDATA=N1+N2+N3,
&N12=N1+N2,NATP1=NATTRS+1,NMC=1000,NB=5000,KLASS=3,RHO=0.905D0)
DIMENSION XX(NDATA,NATP1)
DIMENSION XX1(N1,NATTRS),XX2(N2,NATTRS),XX3(N3,NATTRS)
DIMENSION XM1(NATTRS),XM2(NATTRS),XM3(NATTRS)
DIMENSION S(NATTRS,NATTRS),SINV(NATTRS,NATTRS)
DIMENSION S1(NATTRS,NATTRS),S2(NATTRS,NATTRS),S3(NATTRS,NATTRS)
EXTERNAL DCORVC,DLINDS
DO 10 I=1,N1
DO 5 J=1,NATTRS
XX1(I,J)=XX(I,J)
5 CONTINUE
10 CONTINUE
DO 20 I=1,N2
DO 15 J=1,NATTRS
XX2(I,J)=XX(N1+I,J)

```

```

15 CONTINUE
20 CONTINUE
  DO 30 I=1,N3
  DO 25 J=1,NATTRS
  XX3(I,J)=XX(N1+N2+I,J)
25 CONTINUE
30 CONTINUE
  IDO=0
  NVAR=NATTRS
  IFRQ=0
  IWT=0
  MOPT=0
  ICOPT=0
  LDCOV=NATTRS
  LDINCD=1
  NROW=N1
  LDX=N1
  CALL DCORVC(IDO,NROW,NVAR,XX1,LDX,IFRQ,IWT,MOPT,
&    ICOPT,XM1,S1,LDCOV,INCD,LDINCD,NOBS,
&    NMISS,SUMWT)
  NROW=N2
  LDX=N2
  CALL DCORVC(IDO,NROW,NVAR,XX2,LDX,IFRQ,IWT,MOPT,
&    ICOPT,XM2,S2,LDCOV,INCD,LDINCD,NOBS,
&    NMISS,SUMWT)
  NROW=N3
  LDX=N3
  CALL DCORVC(IDO,NROW,NVAR,XX3,LDX,IFRQ,IWT,MOPT,
&    ICOPT,XM3,S3,LDCOV,INCD,LDINCD,NOBS,
&    NMISS,SUMWT)
  NDM3=NDATA-3
  DO 40 I=1,NATTRS
  DO 35 J=1,NATTRS
  S(I,J)=(N1-1)*S1(I,J)+(N2-1)*S2(I,J)+(N3-1)*S3(I,J)/NDM3
35 CONTINUE
40 CONTINUE
  CALL DLINDS(NATTRS,S,NATTRS,SINV,NATTRS)
  RETURN
  END

```

FUNCTION GINV(U)

```

C
C TRANSFORMS A RANDOM NUMBER, U, TO AN OBSERVATION FROM THE STANDARD
C DOUBLE EXPONENTIAL DISTRIBUTION.
C

```

```

  IMPLICIT DOUBLE PRECISION (A-H,O-Z)
  IF (U.LT.0.5D0) T=DLOG(2.0D0*U)/DSQRT(2.0D0)
  IF (U.GE.0.5D0) T=-DLOG(2.0D0*(1.0D0-U))/DSQRT(2.0D0)
  GINV=T
  RETURN
  END

```

```

C
C SUBROUTINE POLY ESTIMATES THE LOGISTIC REGRESSION COEFFICIENTS IN

```


C A POLYCHOTOMOUS LOGISTIC REGRESSION.
 C IT WAS OBTAINED FROM THE EVALUATION ASSISTANT PACKAGE OF HENERY AND
 C GAMA, AND IS GIVEN IN ITS ORIGINAL FORM.
 C

```
subroutine poly(iw,iclass,niter,ndata,klass,nattrs,dsmall,
&          x,beta)
```

```
implicit real*8 (a-h,o-z)
real*8 xwx(nattrs+1,klass-1,nattrs+1,klass-1)
real*8 beta(nattrs+1,klass-1), delbeta(nattrs+1,klass-1)
real*8 betaold(nattrs+1,klass-1), x(ndata,nattrs)
real*8 allpro(ndata,klass), sumpro(klass)
real*8 resid(nattrs+1,klass-1), prob(klass)
real*8 invxwx(nattrs+1,klass-1,nattrs+1,klass-1)
real*8 oldinv(nattrs+1,klass-1,nattrs+1,klass-1)
real*8 mean(nattrs,klass), xwork(nattrs+1), xbar(nattrs)
real*8 betax(klass-1), prod(klass-1)
real*8 betinv(klass-1,klass-1), betvar(klass-1,klass-1)
integer*4 nfreq(klass),iclass(ndata)
data one, two, three, four, five/1.d0, 2.d0,3.d0,4.d0,5.d0/
```

```
nparam = (nattrs+1) * (klass-1)
```

```
c
c olddev is the previous deviance (an arbitrarily large no. initially)
c iterations stop if delta(deviance) < big
```

```
c
c call meancal(x,mean,nfreq,nattrs,klass,iclass,ndata,
+          xwork,xbar)
```

```
c
c calculate the overall means for all attributes
c
c devnul = Null deviance H0: all classes equally likely
c          and all attributes irrelevant
```

```
class = klass
devnul = two * ndata * dlog(class)
```

```
c
c devpro = Null deviance H0: classes not equally likely
c          and all attributes irrelevant
```

```
c
c devpro = 0.0d0
c fndata = ndata
c do 68 kk=1,klass
c   ffk = nfreq(kk)/fndata
c   devpro = devpro - two * nfreq(kk) * dlog(ffk)
```

```
68 continue
big = dsmall * nattrs * (klass-1)
```

```
c
c With p degrees of freedom, 'big' is not so big
c a difference in deviances to be significant
c
c Starting values for beta, delbeta
c Either start with beta = 0 and deltabeta from linear disc. file
c
c Or start with beta = log(class probs.)
```

```

c and find min. gradient (= linear disc.)
c
do 2 kj=1,nattr+1
do 2 kk=1,klass-1
beta(kj,kk) = 0.0d0
delbeta(kj,kk) = 0.0d0
2 betaold(kj,kk) = 0.0d0

do 25 km=1,klass
probk = nfreq(km) / fndata
if (km.lt.klass) betaold(1,km) = dlog(probk)
do 25 kn=1,ndata
25 allpro(kn,km) = probkm

ifull = 1
call findel(x,iclass,nattr,klass,ndata,mean,xwork,
+ betaold,nfreq,prob,allpro,xwx,resid,nparam,delbeta,
+ dnorm,invxwx,ifull)
call newbeta(betaold,delbeta,beta,dnorm,nattr,klass)

olddev = devnul
call equate(oldinv,invxwx,nparam*nparam)
c
c Now olddev = value of deviance at beta = 0
c delbeta is normalised max. gradient direction
c dnorm = magnitude of step for max. gradient
c
c This is the magnitude of first iteration step
c maximum iterations = niter
c-----begin iteration loop
c
ifull = 1
do 999 iter=1,niter

c
c F(alpha,delta) = deviance(alpha,delta)
c delta = direction of maximum gradient
c alpha = scalar parameter
c
c-----
c
c remember what the last step length was
oldnor = dnorm
c
c Either call golden or devcal (golden calls devcal repeatedly)
c to find the best alpha
c (else just use alpha = dnorm in devcal)
c
c to call golden, set igold = 1
c devcal, 0
c-----
igold = 1
if (igold.eq.1) then

```

```

    call golden(betaold,delbeta,beta,nattrs,klass,
+           ndata,x,iclass,prob,allpro,dnorm,alpha,devs)
    call newbeta(betaold,delbeta,beta,alpha,nattrs,klass)
  else
    alpha = dnorm
    call devcal(betaold,delbeta,beta,nattrs,klass,
+           ndata,x,iclass,prob,allpro,alpha,devs)
  endif
  call ips(allpro,sumpro,nfreq,ndata,klass,nattrs,devs,
+       beta,dchisq)
c
c   take new beta's as OK for now ...
c   (but remember what the previous values of delbeta were ...
c
  do 31 jp=1,nattrs+1
  do 31 jk=1,klass-1
  delbeta(jp,jk) = 0.0d0
31  betaold(jp,jk) = beta(jp,jk)
c
c           If deviance is much less than old values
c           calculate the new delta beta's
c           (otherwise exit)
c
  if (devs.gt.olddev - big) goto 99
c


---


c           Now find new direction of maximum gradient ...
c
  call findel(x,iclass,nattrs,klass,ndata,mean,xwork,
+       beta,nfreq,prob,allpro,xwx,resid,nparam,delbeta,
+       dnorm,invxwx,ifull)
c


---


c   Premature stop if proposed step length is GIGANTIC
  if (dnorm.gt.100.d0 * oldnor) then

    iw=1
    write (6,*) " Evidence of instability"
    write (6,32) dnorm/oldnor
32  format(" Next step length would be ",e12.3, " times first step")
c
c           use the previous inverse
c           and from now on, only calculate residuals
c           (and don't bother finding xwx)
c
    ifull = 0
    call equate(invxwx,oldinv,nparam*nparam)
    call findel(x,iclass,nattrs,klass,ndata,mean,xwork,
+       beta,nfreq,prob,allpro,xwx,resid,nparam,delbeta,
+       dnorm,invxwx,ifull)

    else
    call equate(oldinv,invxwx,nparam*nparam)
  endif
  olddev = devs
999 continue

```

```

c
c           end of iteration loop -----
c
c   write (6,*) " Failed to converge"
c   iw=1
c   goto 9
99  continue
9   continue

c   call zbetax(beta,invxwx,betax,betinv,natrs,klass,
+         betvar,prod,chisq,ndata)

c   return
c   end

c   subroutine process(xin,iclass,natrs,klass,ndata,mean,x,
+         beta,nfreq,prob,allpro,xwx,resid,ifull)
c
c           input beta's
c
c           output nfreq, xwx, resid, devs
c
c           workspace  prob
c
c   implicit real*8 (a-h,o-z)
c   real*8 xwx(natrs+1,klass-1,natrs+1,klass-1)
c   real*8 beta(natrs+1,klass-1), mean(natrs)
c   real*8 resid(natrs+1,klass-1), prob(klass)
c   real*8 x(natrs+1), xin(ndata,natrs), allpro(ndata,klass)
c   integer*4 klass,ndata,natrs,nfreq(klass),iclass(ndata),ifull
c   integer*4 nclass

c   data one/1.d0/
c   x(1) = one
c   do 10 i=1,ndata
c
c           Use the deviations from the overall means to
c           improve numerical accuracy
c
c           Use previously calculated probabilities
c
c   do 44 kp=1,natrs
44  x(kp+1) = xin(i,kp)
c   nclass = iclass(i)
c   do 45 kk=1,klass
45  prob(kk) = allpro(i,kk)
c   if (ifull.eq.1) call design(x,prob,xwx,natrs,klass)
c   call resids(x,nclass,prob,natrs,klass,resid)
10  continue
c   return
c   end

```

```

subroutine matinv(a,ainv,detlog,n)
implicit real*8 (a-h,o-z)
real*8 a(n,n),ainv(n,n)
integer*4 n
c
c      A must be SYMMETRIC
c      T is upper triangular matrix
c      Choleski decomposition  A = T'.T
c      S = INV(T)
c
c      AINV = S.S' = INV(A)
c
do 23 i=1,n
do 23 j=1,n
23  ainv(i,j) = 0.0d0
detlog = dlog(a(1,1))
ainv(1,1) = dsqrt(a(1,1))
do 2 j=2,n
2  ainv(1,j) = a(1,j)/ainv(1,1)
do 3 i=2,n
ainv(i,i) = a(i,i)
i1 = i-1
i2 = i+1
do 4 k=1,i1
4  ainv(i,i) = ainv(i,i) - ainv(k,i)**2
detlog = detlog + dlog(ainv(i,i))
ainv(i,i) = dsqrt(ainv(i,i))

if (i.eq.n) goto 3
do 5 j=i2,n
ainv(i,j) = a(i,j)
do 6 k=1,i1
6  ainv(i,j) = ainv(i,j) - ainv(k,i)*ainv(k,j)
5  ainv(i,j) = ainv(i,j)/ainv(i,i)
3  continue
c
c      AINV is now upper diagonal factor T of A = T'.T
c      detlog is now the logarithm of determinant of A
c      now find inverse S = INV(T)
c
do 7 i=1,n
i2 = i+1
ainv(i,i) = 1.d0/ainv(i,i)
if (i.eq.n) goto 7
do 8 j=i2,n
j1 = j-1
temp = 0.0d0
do 9 k=1,j1
9  temp = temp-ainv(i,k)*ainv(k,j)
8  ainv(i,j) = temp/ainv(j,j)
7  continue
c
c      AINV is now the inverse of T
c
do 10 i=1,n

```

```
do 10 j=i,n
temp = 0.0d0
do 11 k=j,n
11 temp = temp + ainv(i,k)*ainv(j,k)
ainv(i,j) = temp
10 ainv(j,i) = temp
return
end
```

```
subroutine matmul(a,b,prod,n1,n2,n3)
implicit real*8 (a-h,o-z)
real*8 a(n1,n2),b(n2,n3),prod(n1,n3)
integer*4 n1,n2,n3
data zero/ 0.0d0/
do 1 k1=1,n1
do 1 k3=1,n3
temp = 0.0d0
do 2 k2=1,n2
temp = temp + a(k1,k2) * b(k2,k3)
2 continue
prod(k1,k3) = temp
1 continue
return
end
```

```
subroutine inner(a,b,prod,n1,n2)
implicit real*8 (a-h,o-z)
real*8 a(n1,n2),b(n1,n2),prod(n2)
integer*4 n1,n2
data zero/0.0d0/
do 1 k2=1,n2
temp = 0.0d0
do 2 k1=1,n1
temp = temp + a(k1,k2) * b(k1,k2)
2 continue
prod(k2) = temp
1 continue
return
end
```

```
subroutine design(x,prob,xwx,nattrs,klass)
implicit real*8 (a-h,o-z)
real*8 xwx(nattrs+1,klass-1,nattrs+1,klass-1)
real*8 x(nattrs+1),prob(klass)
integer nattrs, klass
do 1 ir=1,klass-1
do 1 it=ir,klass-1
do 1 js=1,nattrs+1
do 1 ju=js,nattrs+1
sum = xwx(js,ir,ju,it)
prodpr = - prob(ir) * prob(it)
if (ir.eq.it) prodpr = prodpr + prob(ir)
```

```

sum = sum + x(js) * x(ju) * prodpr
xwx(js,ir,ju,it) = sum
xwx(js,it,ju,ir) = sum
xwx(ju,ir,js,it) = sum
xwx(ju,it,js,ir) = sum

```

```

1 continue
return
end

```

```

subroutine resid(x,nclass,prob,natrs,klass,resid)

```

```

implicit real*8 (a-h,o-z)
real*8 x(natrs+1),prob(klass)
real*8 resid(natrs+1,klass-1)
integer*4 nclass,natrs,klass
data zero,one /0.0d0, 1.0d0/
do 1 kl=1,klass-1
ydata = 0.0d0
if (kl.eq.nclass) ydata = one
do 1 kp=1,natrs+1
resid(kp,kl) = resid(kp,kl) +
+ x(kp) * (ydata - prob(kl))

```

```

1 continue
return
end

```

```

subroutine newbeta(betaold,delbeta,beta,alpha,natrs,klass)

```

```

implicit real*8 (a-h,o-z)
real*8 beta(natrs+1,klass-1),delbeta(natrs+1,klass-1)
real*8 betaold(natrs+1,klass-1)
do 1 j=1,natrs+1
do 1 k=1,klass-1
beta(j,k) = betaold(j,k) + alpha * delbeta(j,k)

```

```

1 continue
return
end

```

```

subroutine meancal(xout,mean,nfreq,natrs,klass,iclass,ndata,
+ x,xbar)

```

```

c
c          calculate means, class frequencies
c
c          xbar  overall means (without regard to classes)
c          mean  means for individual classes
c

```

```

implicit real*8 (a-h,o-z)
real*8 mean(natrs,klass),xbar(natrs)
real*8 xout(ndata,natrs),x(natrs+1)
integer*4 klass,natrs,ndata,nfreq(klass),iclass(ndata)
data one/1.d0/
do 43 k=1,klass
do 1043 j=1,natrs
mean(j,k)=0.0d0

```

```

xbar(j)=0.0d0
1043 continue
43  nfreq(k) = 0
   do 10 i=1,ndata
   do 1 j=1,nattr
x(j)=xout(i,j)
1  continue
   nclass=iclass(i)
   if (nclass.eq.0) nclass = klass
c      class 0 is always treated as final class
   do 44 kp=1,nattr
   xout(i,kp) = x(kp)
   xbar(kp) = xbar(kp) + x(kp)
44  mean(kp,nclass) = mean(kp,nclass) + x(kp)
   nfreq(nclass) = nfreq(nclass) + 1
   iclass(i)=nclass
10  continue
999 continue
   do 20 kp=1,nattr
   xbar(kp) = xbar(kp) / ndata
   do 21 kk=1,klass
21  mean(kp,kk) = mean(kp,kk) / nfreq(kk)
20  continue
   return
   end

subroutine discpr(xin,iclass,prob,allpro,beta,klass,
+      nattr,ndata,devs)
implicit real*8 (a-h,o-z)
real*8 xin(ndata,nattr)
real*8 prob(klass),allpro(ndata,klass)
real*8 beta(nattr+1,klass-1)
integer*4 klass,nattr,ndata,iclass(ndata)
data zero, one, two, pllim/0.0d0, 1.d0, 2.d0, -60.d0/
data epsiln / 1.d-30/
devs = 0.0d0
do 99 i=1,ndata
  nclass = iclass(i)
  prob(klass) = 0.0d0
  do 11 k=1,klass-1
  prob(k) = beta(1,k)
  do 11 n=1,nattr
  prob(k) = prob(k) + xin(i,n) * beta(n+1,k)
11  continue
c
c      ensure that the maximum is zero and min = pllim
c      so that, when exponentiating, max = 1, min = exp(pllim)
c
c      when pllim = -60,  min(Prob) = exp(-60) = 8.756511e-27
c
  prmax = -1.d40
  do 12 k=1,klass
  if (prmax.lt.prob(k)) prmax = prob(k)
12  continue

```



```

sumpr = 0.0d0
do 13 k=1, klass
  prob(k) = prob(k) - prmax
  if (prob(k).lt.pllim) prob(k) = pllim
  prob(k) = dexp(prob(k))
  sumpr = sumpr + prob(k)
13 continue
do 14 k=1, klass
  prob(k) = prob(k) / sumpr
  allpro(i,k) = prob(k)
14 continue
c
c           probabilities now sum to one
c           conditional probabilities of class given x
c
c           deviance is 2log(prob(observed class))
devs = devs - two * dlog(prob(nclass) + epsilon)
99 continue
return
end

```

```

subroutine nordel(delta, ndim, dnorm)
real*8 delta(ndim), dnorm, zero
data zero/0.0d0/
dnorm = 0.0d0
do 1 k=1, ndim
1  dnorm = dnorm + delta(k)**2
  dnorm = dsqrt(dnorm)
do 2 k=1, ndim
2  delta(k) = delta(k) / dnorm
return
end

```

```

subroutine findel(x, iclass, nattrs, klass, ndata, mean, xwork,
+           beta, nfreq, prob, allpro, xwx, resid, nparam, delbeta,
+           dnorm, invxwx, ifull)
implicit real*8 (a-h, o-z)
real*8 xwx(nattrs+1, klass-1, nattrs+1, klass-1)
real*8 beta(nattrs+1, klass-1), delbeta(nattrs+1, klass-1)
real*8 x(ndata, nattrs)
real*8 allpro(ndata, klass)
real*8 resid(nattrs+1, klass-1), prob(klass)
real*8 invxwx(nattrs+1, klass-1, nattrs+1, klass-1)
real*8 mean(nattrs, klass), xwork(nattrs+1)
integer*4 klass, ndata, nattrs, nfreq(klass), iclass(ndata), ifull
data zero/0.0d0/
c
c
c           reset arrays to zero
c
do 33 kp=1, nattrs+1
do 33 kc=1, klass-1
if (ifull.eq.0) goto 33

```

```

do 34 jp=1,nattrs+1
do 34 jc=1,klass-1
34 xwx(kp,kc,jp,jc) = 0.0d0
33 resid(kp,kc) = 0.0d0
call process(x,iclass,nattrs,klass,ndata,mean,xwork,
+ beta,nfreq,prob,allpro,xwx,resid,ifull)
fndata = ndata
do 44 kp=1,nattrs+1
do 44 kc=1,klass-1
resid(kp,kc) = resid(kp,kc) / fndata
do 45 jp=1,nattrs+1
do 45 jc=1,klass-1
45 xwx(kp,kc,jp,jc) = xwx(kp,kc,jp,jc) / fndata
44 continue
if (ifull.eq.1) call matinv(xwx,invxwx,detlog,nparam)
c
c
c NB resid is now a vector of length nparam
call matmul(invxwx,resid,delbeta,nparam,nparam,1)
c
c The delbeta's must now be 'normalised' to unit length
call nordel(delbeta,nparam,dnorm)
return
end

```

```

subroutine devcal(betaold,delbeta,beta,nattrs,klass,
+ ndata,x,iclass,prob,allpro,alpha,devs)
implicit real*8 (a-h,o-z)
real*8 beta(nattrs+1,klass-1),delbeta(nattrs+1,klass-1)
real*8 betaold(nattrs+1,klass-1),x(ndata,nattrs)
real*8 allpro(ndata,klass),prob(klass)
integer*4 klass,nattrs,ndata,iclass(ndata)
call newbeta(betaold,delbeta,beta,alpha,nattrs,klass)
call discpr(x,iclass,prob,allpro,beta,klass,
+ nattrs,ndata,devs)
return
end

```

```

subroutine equate(vecnew,vecold,ndim)
implicit real*8 (a-h,o-z)
real*8 vecnew(ndim), vecold(ndim)
integer*4 ndim
do 1 j=1,ndim
vecnew(j) =vecold(j)
1 continue
return
end

```

```

subroutine zbetax(beta,invxwx,betax,betinv,nattrs,klass,
+ betvar,prod,chisq,ndata)
implicit real*8 (a-h,o-z)
real*8 invxwx(nattrs+1,klass-1,nattrs+1,klass-1)

```

```

real*8 beta(nattrs+1,klass-1),betax(klass-1),prod(klass-1)
real*8 betax1(klass-1)
real*8 betinv(klass-1,klass-1),betvar(klass-1,klass-1)
integer*4 klass,nattrs
fndata = ndata
do 1 k=1,nattrs
chisq = 0.0d0
do 2 kl=1,klass-1
betax(kl) = beta(k+1,kl)
do 2 kj=1,klass-1
betinv(kl,kj) = invxwx(k+1,kl,k+1,kj) / fndata
2 continue
call matinv(betinv,betvar,detbet,klass-1)
call matmul(betvar,betax,prod,klass-1,klass-1,1)
call inner(betax,prod,chisq,klass-1,1)
do 3 m=1,klass-1
betax1(m) = betax(m) / dsqrt(betinv(m,m))
3 continue
1 continue
return
end

subroutine golden(betaold,delbeta,beta,nattrs,klass,
+      ndata,x,iclass,prob,allpro,dnorm,alpha,devs)
c
c calculate best step length for current direction of search
c
c and probabilities for all data (used in later calculations)
c
implicit real*8 (a-h,o-z)
real*8 beta(nattrs+1,klass-1),delbeta(nattrs+1,klass-1)
real*8 betaold(nattrs+1,klass-1),x(ndata,nattrs)
real*8 allpro(ndata,klass),prob(klass)
integer*4 klass,nattrs,ndata,iclass(ndata)
data eps/ 0.1d0/
data one, two, three, four, five/1.d0, 2.d0,3.d0,4.d0,5.d0/
snorm = dnorm * eps
v1 = (three - dsqrt(five))/two
v2 = (dsqrt(five) - one)/two
ratio = one + v2
c
c now pick length of first step (on the theory that the Newton Raphson
c value is about right), so that three points are taken, straddling
c the Newton value.
c
11 continue
tau = 0.05d0 * dnorm
alpha = dnorm - ratio*tau
a1 = 0.0d0
b = alpha
call devcal(betaold,delbeta,beta,nattrs,klass,
+      ndata,x,iclass,prob,allpro,b,r2)
22 continue
tau = tau * ratio

```

```

a = a1
a1 = b
r1 = r2
b = b + tau
call devcal(betaold,delbeta,beta,natrs,klass,
+          ndata,x,iclass,prob,allpro,b,r2)
if (r2.lt.r1) goto 22
c
c          from here on, minimum is in range (a,b)
c          write (26,*) " a and b", a, b
c
range = b - a
del = a + v1 * range
sig = a + v2 * range
call devcal(betaold,delbeta,beta,natrs,klass,
+          ndata,x,iclass,prob,allpro,del,rdel)
call devcal(betaold,delbeta,beta,natrs,klass,
+          ndata,x,iclass,prob,allpro,sig,rsig)
33 continue
if (rdel.lt.rsig) then
  b = sig
  sig = del
  rsig = rdel
  range = b - a
  del = a + v1 * range
  call devcal(betaold,delbeta,beta,natrs,klass,
+          ndata,x,iclass,prob,allpro,del,rdel)
  else
  a = del
  del = sig
  rdel = rsig
  range = b - a
  sig = a + v2 * range
  call devcal(betaold,delbeta,beta,natrs,klass,
+          ndata,x,iclass,prob,allpro,sig,rsig)
  endif
if (range.gt.snorm) goto 33
c
c ----- loop to find tight range for alpha
c
alpha = del
devs = rdel
if (rsig.lt.rdel) then
  devs = rsig
  alpha = sig
endif
return
end

subroutine ips(allpro,sumpro,nfreq,ndata,klass,natrs,
+          big,beta,dchisq)
implicit real*8 (a-h,o-z)
real*8 allpro(ndata,klass),sumpro(klass)
real*8 beta(natrs+1,klass-1)

```

```

integer*4 klass,ndata,nfreq(klass)
data zero, small/0.0d0, 0.01d0/
iter = 0
1  continue
  iter = iter + 1
  if (iter.eq.24) return
  chisq = 0.0d0
  do 2 k=1,klass
    sumpro(k) = 0.0d0
  do 3 n=1,ndata
    sumpro(k) = sumpro(k) + allpro(n,k)
3  continue
  chisq = chisq + (nfreq(k)-sumpro(k))**2/sumpro(k)
  sumpro(k) = nfreq(k) / sumpro(k)
2  continue
  do 22 k=1,klass-1
22 beta(1,k) = beta(1,k) + dlog(sumpro(k)/sumpro(klass))
  if (iter.eq.1) chisq1 = chisq
  dchisq = chisq1 - chisq
  if (chisq.lt.small*big) return
c
c           then fit is good enough
c
c           otherwise rescale all "probabilities"
c           renormalise over rows, and do another column sweep
c
  do 4 n=1,ndata
    sumrow = 0.0d0
  do 5 k=1,klass
    allpro(n,k) = allpro(n,k) * sumpro(k)
5  sumrow = sumrow + allpro(n,k)
  do 6 k=1,klass
    allpro(n,k) = allpro(n,k) / sumrow
6  continue
4  continue
  goto 1
  return
end

```

PROGRAM 2

C IN THIS PROGRAM A MONTE CARLO SIMULATION STUDY IS DONE TO COMPARE THE
C FOLLOWING VARIABLE SELECTION PROCEDURES IN DISCRIMINANT ANALYSIS

- C
C 1. THE 20% OR 40% HOLDOUT-METHOD PROPOSED BY RUTTER, FLACK AND
C LACHENBRUCH (1991)
C 2. THE NSp* METHOD PROPOSED BY SNAPINN EN KNOKE (1989)
C 3. THE CROSS MODEL VALIDATION TECHNIQUE WITH FORWARD F-BASED
C SELECTION AS INNER CRITERION.

C THE FEATURE VARIABLES ARE ASSUMED TO BE UNCORRELATED, AND TO
C HAVE A LOGNORMAL DISTRIBUTION

C PARAMETERS :

- C IP=THE TOTAL NUMBER OF AVAILABLE FEATURE VARIABLES
C NN=THE SIZE OF THE TRAINING DATA SET FROM GROUP 1
C MM=THE SIZE OF THE TRAINING DATA SET FROM GROUP 2
C NNPMM=NN+MM=THE TOTAL SIZE OF THE TRAINING DATA SET
C NMC=NUMBER OF MONTE CARLO REPETITIONS
C NB=NUMBER OF SIMULATION REPETITIONS USED PER GROUP TO
C ESTIMATE THE POST-SELECTION ACTUAL ERROR RATE

C THE FOLLOWING IMSL-SUBROUTINES ARE USED IN THE MAIN PROGRAM:

- C 1. ERSET : PREVENTS THE PROGRAM FROM TERMINATING IF DRSTEP SELECTS
C NO VARIABLES
C 2. DLINDS: FINDS THE INVERSE OF A GIVEN COVARIANCE MATRIX
C 3. DCHFAC: FINDS THE CHOLSKY DECOMPOSITION OF A GIVEN MATRIX
C 4. DRNMVN: GENERATES VALUES FROM A MULTIVARIATE NORMAL DISTRIBUTION
C 5. DCORVC: COMPUTES A COVARIANCE OR CORRELATION MATRIX
C 6. DRSTEP: BUILDS MULTIPLE LINEAR REGRESSION MODELS USING FORWARD
C SELECTION, BACKWARD SELECTION, OR STEPWISE SELECTION -
C CAN ALSO BE USED FOR THIS PURPOSE IN DA

IMPLICIT DOUBLE PRECISION (A-H,O-Z)

PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1,NMC=500,
& NB=200)

DIMENSION AMU(2,IP),AMUSEL(2,IP)

DIMENSION SIGMAM(IP,IP),SIGINV(IP,IP),RSIG(IP,IP)

DIMENSION RNX1(NN,IP),RNX2(MM,IP),RESP(NNPMM),XX(NNPMM,IPP1)

C DIMENSION XVH(IP),XX1H(NNPMM-1,IPP1),XX2H(NNPMM,IPP1)

DIMENSION ERRORH(NNPMM,IP),ERTOTH(IP)

DIMENSION PSELVARH(IP),PSELNUMH(IP)

DIMENSION MINH(IP),ISELH(IP)

C DIMENSION XMEANH(IPP1),COVH(IPP1,IPP1)

DIMENSION HISTH(IPP1),AOVH(13),COEFH(IP,5)

DIMENSION SCALEH(IPP1),COVSH(IPP1,IPP1)

DIMENSION LEVELH(IPP1)

C DIMENSION XVL(IP)

```

DIMENSION XX1L(NNPMM,IPP1),XX2L(NNPMM,IPP1),XX3L(NNPMM,IPP1)
DIMENSION XX5L(NNPMM,IPP1),XX6L(NNPMM,IPP1)
DIMENSION PSELVARL(IP),PSELNUML(IP)

```

C

```

DIMENSION XMEANL(IPP1),COVL(IPP1,IPP1)
DIMENSION HISTL(IPP1),AOVL(13),COEFL(IP,5)
DIMENSION SCALEL(IPP1),COVSL(IPP1,IPP1)
DIMENSION LEVLL(IPP1)
DIMENSION DISTL(2),SL(IP,IP),SINVL(IP,IP),XM1L(IP),XM2L(IP)

```

C

```

CHARACTER*70 FILEOUT
FILEOUT='/log.d'
CALL ERSET(0,1,0)

```

C

```

C FOR SMALL SAMPLES (NN=25,MM=25) THE HOLDOUT FRACTION IN THE METHOD OF
C RUTTER ET AL. (1991) IS 20%, AND FOR MIXED SAMPLES (NN=75,MM=25) AND
C LARGE SAMPLES (NN=100,MM=100) THE HOLDOUT FRACTION IS 40% -
C FRAC IS THE FRACTION OF THE DATA USED IN THE SELECTION STEP

```

C

```

IF (NN.GT.25) FRAC=0.6D0
IF (NN.LE.25) FRAC=0.8D0

```

C

```

C NONZERO IS THE NUMBER OF NONZERO ELEMENTS OF THE MEAN VECTOR OF THE
C SECOND GROUP - ALL THE ELEMENTS OF THE MEAN VECTOR OF THE FIRST GROUP
C ARE TAKEN EQUAL TO ZERO

```

C

```

NONZERO=10
DO 2 I=1,IP
SIGMAM(I,I)=1.0D0
DO 1 J=1,IP
IF (I.NE.J) SIGMAM(I,J)=0.0D0

```

1 CONTINUE

2 CONTINUE

```

DO 4 I=1,IP
AMU(1,I)=0.0D0

```

4 CONTINUE

```

CALL DLINDS(IP,SIGMAM,IP,SIGINV,IP)
SUMSIG=0.0D0
DO 9 I=1,NONZERO
DO 8 J=1,NONZERO
SUMSIG=SUMSIG+(1.0D0*I)*(1.0D0*J)*SIGINV(I,J)

```

8 CONTINUE

9 CONTINUE

C

```

C THE CONSTANTS NECESSARY FOR THE JOHNSON TRANSFORMATION OF NORMAL
C VARIABLES TO LOGNORMAL VARIABLES ARE DEFINED

```

C

```

E=DEXP(1.0D0)
ALAM=DSQRT(1.0D0/(E*(E-1.0D0)))
EP=-1.0D0*DSQRT(1/(E-1.0D0))

```

C

```

C DIE LOOP UP TO 500 SYSTEMATICALLY INCREASES THE MAHALANOBIS DISTANCE
C BETWEEN THE TWO GROUPS

```

C THE FOLLOWING SIMULATION COUNTERS ARE ALSO INITIALISED:
 C 1. PSELVAR(L/H)(J): THE ESTIMATED PROBABILITY OF CHOOSING VARIABLE J
 C 2. PSELNUM(L/H)(J): THE ESTIMATED PROBABILITY OF CHOOSING A MODEL WITH
 C J VARIABLES
 C 3. ERE(L/S/H): THE AVERAGE ESTIMATED ACTUAL ERROR RATE
 C 4. AMSEOP(L/S/H): THE UMSE FOR ESTIMATION OF THE OPTIMAL ERROR RATE
 C 5. AUMSE(L/S/H): THE UMSE FOR ESTIMATION OF THE ACTUAL ERROR RATE
 C 6. ERACT(L/H): THE AVERAGE POST-SELECTION ACTUAL ERROR RATE
 C 7. EROPT(L/H): THE AVERAGE POST-SELECTION OPTIMAL ERROR RATE
 C

```

DO 500 IS=0,6
IF (IS.LE.4) D2=1.0D0*IS
IF (IS.EQ.5) D2=6.0D0
IF (IS.EQ.6) D2=9.0D0
D1=DSQRT(D2)
DO 12 J=1, NONZERO
AMU(2,J)=DSQRT(D2/SUMSIG)*J
PSELVARL(J)=0.0D0
PSELNUML(J)=0.0D0
PSELVARH(J)=0.0D0
PSELNUMH(J)=0.0D0
12 CONTINUE
IF (NONZERO.LT.IP) THEN
DO 13 J=NONZERO+1,IP
AMU(2,J)=0.0D0
PSELVARL(J)=0.0D0
PSELNUML(J)=0.0D0
PSELVARH(J)=0.0D0
PSELNUMH(J)=0.0D0
13 CONTINUE
ENDIF

```

```

EREL=0.0D0
AMSEOPL=0.0D0
AUMSEL=0.0D0
ERACTL=0.0D0
EROPTL=0.0D0
CPCSL=0.0D0
SELOVERL=0.0D0
SELUNDERL=0.0D0
SELMIXL=0.0D0
ERES=0.0D0
AMSEOPS=0.0D0
AUMSES=0.0D0
EREH=0.0D0
AMSEOPH=0.0D0
AUMSEH=0.0D0
ERACTH=0.0D0
EROPH=0.0D0
CPCSH=0.0D0
SELOVERH=0.0D0
SELUNDERH=0.0D0
SELMIXH=0.0D0

```

```
TOL=1.0D2*DMACH(4)
```



```

CALL DCHFAC(IP,SIGMAM,IP,TOL,IRANK,RSIG,IP)
C
C THE SIMULATION LOOP STARTS - THE NECESSARY TRAINING DATA SET VALUES
C ARE FIRST OF ALL GENERATED FROM THE RELEVANT NORMAL DISTRIBUTIONS
C
MC=0
14 CALL DRNMVN(NN,IP,RSIG,IP,RNX1,NN)
CALL DRNMVN(MM,IP,RSIG,IP,RNX2,MM)
C
C THE NORMAL VALUES ARE TRANSFORMED TO LOGNORMAL VALUES USING THE
C JOHNSON TRANSFORMATION SYSTEM. THE ELEMENTS OF THE MEAN VECTORS ARE
C ALSO ADDED.
C
DO 16 I=1,NN
DO 15 J=1,IP
RNX1(I,J)=(ALAM*DEXP(RNX1(I,J)))+EP+AMU(1,J)
15 CONTINUE
16 CONTINUE
DO 20 I=1,MM
DO 19 J=1,IP
RNX2(I,J)=(ALAM*DEXP(RNX2(I,J)))+EP+AMU(2,J)
19 CONTINUE
20 CONTINUE
C
C THE RESPONSE VECTOR INDICATING GROUP MEMBERSHIP IS SET UP
C
DO 25 I=1,NN
RESP(I)=1.0D0
25 CONTINUE
DO 30 I=NN+1,NNPMM
RESP(I)=2.0D0
30 CONTINUE
C
C A SINGLE DATA MATRIX XX(NNPMM x IP+1) IS FORMED. THE FIRST IP COLUMNS
C CONTAIN THE FEATURE VARIABLE VALUES, WHILE COLUMN (IP+1) CONTAINS
C THE RESPONSE VARIABLE VALUES INDICATING GROUP MEMBERSHIP.
C
DO 45 J=1,IP
DO 35 I=1,NN
XX(I,J)=RNX1(I,J)
35 CONTINUE
DO 40 I=1,MM
XX(NN+I,J)=RNX2(I,J)
40 CONTINUE
45 CONTINUE
DO 50 I=1,NN
XX(I,IP+1)=RESP(I)
50 CONTINUE
DO 55 I=1,MM
XX(NN+I,IP+1)=RESP(NN+I)
55 CONTINUE
C
C THIS IS THE BEGINNING OF THE METHOD OF RUTTER ET AL. (1991).
C FIRSTLY, THE DATA IS SPLIT INTO TWO PARTS. THE ONE PART (IN MATRIX XX2L)
C IS USED TO PERFORM FORWARD STEPWISE SELECTION. THE SECOND PART OF THE

```

C DATA (IN MATRIX XX3L) IS THEN USED TO CALCULATE AN ERROR RATE ESTIMATE

C

```
N1=INT(FRAC*NN)
N2=INT(FRAC*MM)
IROW=N1+N2
CALL HOLDOUT(IPP1,N1,N2,XX,XX2L,XX3L)
```

```
IDO=0
NROW=IROW
NVAR=IPP1
LDX=NNPMM
IFRQ=0
IWT=0
MOPT=0
ICOPT=0
LDCOV=IPP1
LDINCD=1
```

```
CALL DCORVC(IDO,NROW,NVAR,XX2L,LDX,IFRQ,IWT,MOPT,
& ICOPT,XMEANL,COVL,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)
```

C

C A FORWARD STEPWISE DISCRIMINANT ANALYSIS IS NOW PERFORMED

C

```
INVOKE=0
NVAR=IPP1
LDCOV=IPP1
DO 60 I=1,IP
  LEVELL(I)=2
60 CONTINUE
  LEVELL(IPP1)=-1
  NFORCE=1
  NSTEP=-1
  ISTEP=1
  NOBS=IROW
  PIN=0.15D0
  POUT=0.15D0
  TOL=1.0D2*DMACH(4)
  IPRINT=0
  LDCOEF=IP
  LDCOVS=IPP1
```

```
CALL DRSTEP(INVOKE,NVAR,COVL,LDCOV,LEVELL,NFORCE,
& NSTEP,ISTEP,NOBS,PIN,POUT,TOL,IPRINT,
& SCALEL,HISTL,IEND,AOVL,COEFL,LDCOEF,COVSL,
& LDCOVS)
```

C

C THE MATRIX XX1L, CONTAINING THE SELECTED COLUMNS OF XX2L, AND

C THE MATRIX XX6L, CONTAINING THE SELECTED COLUMNS OF XX, ARE NOW SET UP.

C XX5L CONTAINS THE SELECTED COLUMNS OF XX3L (THE ORIGINAL HOLDOUT DATA).

C

```
IT=0
DO 70 J=1,IP
  IF (HISTL(J).GT.0) THEN
    IT=IT+1
```

```

DO 65 I=1,IROW
XX1L(I,IT)=XX2L(I,J)
XX6L(I,IT)=XX(I,J)
65 CONTINUE
DO 66 I=1,NNPMM-IROW
XX5L(I,IT)=XX3L(I,J)
XX6L(IROW+I,IT)=XX(IROW+I,J)
66 CONTINUE
AMUSEL(1,IT)=AMU(1,J)
AMUSEL(2,IT)=AMU(2,J)
ENDIF
70 CONTINUE
DO 75 I=1,IROW
XX1L(I,IT+1)=XX2L(I,IPP1)
75 CONTINUE
DO 76 I=1,NNPMM-IROW
XX5L(I,IT+1)=XX3L(I,IPP1)
76 CONTINUE
IF (IT.EQ.0) GOTO 14
C
C SUBROUTINE ERROR IS CALLED TO CALCULATE THE POST-SELECTION
C OPTIMAL (EROPTL) AND ACTUAL ERROR RATES (ERACTL).
C
C ONLY THE SELECTED VARIABLES ARE TAKEN INTO ACCOUNT, BUT ALL THE
C DATA IS USED (XX6L CONTAINS ALL THE DATA, BUT ONLY FOR THE SELECTED
C VARIABLES). 'IT' IS THE NUMBER OF VARIABLES THAT WERE SELECTED.
C
C SINCE THE SELECTION USED FOR THE METHODS PROPOSED BY RUTTER ET
C AL. AND SNAPINN AND KNOKE IS IDENTICAL (FORWARD F-BASED SELECTION
C WITH ALPHA-TO-ENTER=0.15), EROPTL AND ERACTL ARE THE POST-SELECTION
C OPTIMAL AND ACTUAL ERROR RATES FOR BOTH THESE METHODS.
C
CALL ERROR(ALAM,EP,NB,IT,AMUSEL,RSIG,XX6L,OPT,ACT)
EROPTL=EROPTL+OPT
ERACTL=ERACTL+ACT

AMIS=0.0D0
C
C THE SUBROUTINE AVGVAR3 IS NOW CALLED TO COMPUTE THE GROUP MEANS
C AND POOLED SAMPLE COVARIANCE MATRIX (AND ITS INVERSE) OF THE DATA IN
C XX1L (THE SELECTED DATA EXCLUDING THE HOLDOUT CASES)
C
CALL AVGVAR3(N1,N2,IROW,IT,XX1L,SL,SINVL,XM1L,XM2L)
C
C THE HOLDOUT CASES (USING ONLY THE SELECTED VARIABLES, XX5L) ARE CLASSIFIED
C USING THE LINEAR DISCRIMINANT FUNCTION BASED ON THE SELECTED VARIABLES
C IN XX1L ("NON-HOLDOUT" CASES) TO OBTAIN A POST-SELECTION ERROR RATE
C ESTIMATE (ERRATE) FOR THE METHOD OF RUTTER ET AL. (1991)
C
DO 100 I=1,NNPMM-IROW
DO 80 J=1,IT+1
XVL(J)=XX5L(I,J)
80 CONTINUE
SUM1=0.0D0
SUM2=0.0D0

```

```

DO 95 I1=1,IT
DO 90 I2=1,IT
V1=XVL(I1)-XM1L(I1)
V2=XVL(I2)-XM1L(I2)
SUM1=SUM1+V1*SINVL(I1,I2)*V2
V1=XVL(I1)-XM2L(I1)
V2=XVL(I2)-XM2L(I2)
SUM2=SUM2+V1*SINVL(I1,I2)*V2
90 CONTINUE
95 CONTINUE
DISTL(1)=SUM1
DISTL(2)=SUM2
IF (DISTL(1).LT.DISTL(2)) GROUP=1.0D0
IF (DISTL(1).GE.DISTL(2)) GROUP=2.0D0
IF (DABS(GROUP-XVL(IT+1)).GT.0.1D0) AMIS=AMIS+1.0D0
100 CONTINUE
ERRATE=AMIS/(NNPMM-IROW)
C
C THE ERROR RATES ARE ACCUMULATED (EREL) AND COMPONENTS OF THE MEAN
C SQUARED ERROR FOR ESTIMATING THE ACTUAL ERROR RATE (AUMSEL) AND THE
C OPTIMAL ERROR RATE (AMSEOPL) ARE CALCULATED.
C THE QUANTITIES NEEDED TO CALCULATE THE PROBABILITY OF CORRECT SELECTION,
C THE PROBABILITY OF SELECTING THE CORRECT MODEL DIMENSION, THE CONDITIONAL
C PROBABILITY OF CORRECT SELECTION, THE PROBABILITIES OF OVERSELECTION,
C UNDERSELECTION AND MIXED SELECTION, ARE ALSO CALCULATED AND ACCUMULATED.
C
EREL=EREL+ERRATE
AMSEOPL=AMSEOPL+((ERRATE-OPT)**2.0D0)
AUMSEL=AUMSEL+((ERRATE-ACT)**2.0D0)
NUM=0
DO 110 J=1,IP
IF (HISTL(J).GT.0.0D0) THEN
PSELVARL(J)=PSELVARL(J)+1.0D0
NUM=NUM+1
ENDIF
110 CONTINUE
C
C NUM IS THE NUMBER OF VARIABLES THAT WERE SELECTED BY MEANS OF DRSTEP
C
PSELNUML(NUM)=PSELNUML(NUM)+1.0D0
IF (NUM.EQ.NONZERO) THEN
ISELR=1
DO 120 J=1, NONZERO
IF (HISTL(J).LT.0.1D0) ISELR=0
120 CONTINUE
CPCSL=CPCSL+ISELR
ENDIF
IF (NUM.GT.NONZERO) THEN
ISELR=1
DO 121 J=1, NONZERO
IF (HISTL(J).LT.0.1D0) ISELR=0
121 CONTINUE
IF (ISELR.EQ.1) SELOVERL=SELOVERL+1.0D0
ENDIF
IF (NUM.LT.NONZERO) THEN

```

```

ISELW=0
DO 122 J=NONZERO+1,IP
IF (HISTL(J).GT.0.1D0) ISELW=1
122 CONTINUE
IF (ISELW.EQ.0) SELUNDERL=SELUNDERL+1.0D0
ENDIF

ISELM=0
DO 123 J=NONZERO+1,IP
IF (HISTL(J).GT.0.1D0) ISELM=1
123 CONTINUE
IF (ISELM.EQ.1) THEN
NCOR=0
DO 124 J=1,NONZERO
IF (HISTL(J).GT.0.1D0) NCOR=NCOR+1
124 CONTINUE
IF ((NCOR.GT.0).AND.(NCOR.LT.NONZERO)) SELMIXL=
& SELMIXL+1.0D0
ENDIF

```

```

C
C SUBROUTINE WFSTAR IS CALLED TO CALCULATE THE POST-SELECTION
C ERROR RATE ESTIMATOR (ERSMOOTH) PROPOSED BY SNAPINN AND KNOKE (1989).
C SINCE THIS PROCEDURE USES THE SAME SELECTION STRATEGY AS THAT OF
C RUTTER ET AL., IT IS NOT NECESSARY TO REPEAT ANY OF THE SELECTION
C RELATED CALCULATIONS. ONLY THE ERROR RATE ESTIMATE AND THE COMPONENTS
C NEEDED TO CALCULATE THE UMSE OF THE ESTIMATOR NEED TO BE CALCULATED. ALL
C QUANTITIES RELATED TO SELECTION, INCLUDING THE POST-SELECTION ACTUAL AND
C OPTIMAL ERROR RATES, ARE IDENTICAL TO THOSE CALCULATED ABOVE FOR THE
C PROCEDURE OF RUTTER ET AL. (1991).
C THE MATRIX XX6L, CONTAINING ONLY THE SELECTED VARIABLES BUT ALL THE CASES,
C ARE USED. NUM IS THE NUMBER OF SELECTED VARIABLES.

```

```

C
CALL WFSTAR(NUM,XX6L,ERSMOOTH)
ERES=ERES+ERSMOOTH
AMSEOPS=AMSEOPS+((ERSMOOTH-OPT)**2.0D0)
AUMSES=AUMSES+((ERSMOOTH-ACT)**2.0D0)

```

```

C
C ERSMOOTH IS THE  $NSp^*$ -ESTIMATE FOR THE CURRENT MONTE CARLO REPETITION.
C THE ESTIMATES ARE ACCUMULATED IN ERES.
C COMPONENTS OF THE MEAN SQUARED ERRORS OF ESTIMATING THE OPTIMAL AND
C ACTUAL ERROR RATES RESPECTIVELY, ARE ACCUMULATED IN AMSEOPS AND AUMSES.

```

```

C
C THIS IS THE END OF THE PROCEDURES OF RUTTER ET AL AND SNAPINN EN KNOKE.
C

```

```

C
C THE CROSS MODEL VALIDATION METHOD STARTS HERE
C

```

```

DO 165 I=1,NNPMM
DO 160 J=1,IP
ERRORH(I,J)=0.0D0

```

```

160 CONTINUE
165 CONTINUE

```

```

C

```

```
C THE SUBROUTINE LOO IS CALLED TO OMIT THE ROWS ONE BY ONE -
C THE MATRIX XX1H IS THE MATRIX XX WITH ROW II (II=1,NNPMM)
C DELETED
C
DO 200 II=1,NNPMM
CALL LOO(II,XX,XX1H)
IDO=0
NROW=NNPMM-1
NVAR=IPP1
LDX=NNPMM-1
IFRQ=0
IWT=0
MOPT=0
ICOPT=0
LDCOV=IPP1
LDINCD=1

C
C THE IMSL ROUTINE DCORVC IS CALLED TO CALCULATE THE COVARIANCE MATRIX
C COVH NEEDED AS INPUT FOR DRSTEP
C
CALL DCORVC(IDO,NROW,NVAR,XX1H,LDX,IFRQ,IWT,MOPT,
& ICOPT,XMEANH,COVH,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)

DO 195 ID=1,IP
INVOKE=0
NVAR=IPP1
LDCOV=IPP1
DO 167 I=1,IP
LEVELH(I)=2
167 CONTINUE
LEVELH(IPP1)=-1
NFORCE=1
NSTEP=ID
ISTEP=1
NOBS=NNPMM-1
PIN=0.9999999D0
POUT=0.9999999D0
TOL=1.0D2*DMACH(4)
IPRINT=0
LDCOEF=IP
LDCOVS=IPP1

C
C THE IMSL ROUTINE DRSTEP IS USED TO CALCULATE THE BEST MODEL OF DIMENSION
C 1,2,...,IP
C
CALL DRSTEP(INVOKE,NVAR,COVH,LDCOV,LEVELH,NFORCE,
& NSTEP,ISTEP,NOBS,PIN,POUT,TOL,IPRINT,
& SCALEH,HISTH,IEND,AOVH,COEFH,LDCOEF,COVSH,
& LDCOVS)
IT=0
DO 170 J=1,IP
MINH(J)=0
IF (HISTH(J).GT.0) THEN
MINH(J)=1
```

```

      IT=IT+1
      XVH(IT)=XX(IL,J)
    ENDIF
170 CONTINUE
C
C   A SMOOTHED LOSS FOR THE OMITTED CASE IS CALCULATED
C   THIS IS DONE FOR MODEL DIMENSION ID (ID=1,...,IP)
C
    CALL WF(MINH,IL,XX1H,XVH,WW,AMAH)
    IF (IL.LE.NN) THEN
      BKON=((IP+2)*(NN-2.0D0)+MM-1.0D0)/((NN-1.0D0)*
& (NN+MM-IP-4.0D0))
      BKON=DSQRT(BKON)
      ARG=-WW/(BKON*AMAH)
      ERRORH(IL,ID)=DNORDF(ARG)
    ENDIF
    IF (IL.GT.NN) THEN
      BKON=((IP+2)*(MM-2.0D0)+NN-1.0D0)/((MM-1.0D0)*
& (NN+MM-IP-4.0D0))
      BKON=DSQRT(BKON)
      ARG=WW/(BKON*AMAH)
      ERRORH(IL,ID)=DNORDF(ARG)
    ENDIF
195 CONTINUE
200 CONTINUE
C
C   THIS IS THE END OF THE LOOP WHERE THE CASES ARE OMITTED ONE BY ONE
C
C   THE SUM OF THE SMOOTHED ERRORS FOR EACH MODEL DIMENSION (1,...,IP)
C   IS NOW CALCULATED. THIS IS THE CMV-CRITERION ASSOCIATED WITH EACH
C   MODEL DIMENSION
C
    DO 220 J=1,IP
      ERTOTH(J)=0.0D0
      DO 210 I=1,NNPMM
        ERTOTH(J)=ERTOTH(J)+ERRORH(I,J)
      210 CONTINUE
      ERTOTH(J)=ERTOTH(J)/NNPMM
    220 CONTINUE
C
C   THE OPTIMAL MODEL DIMENSION IS IDENTIFIED USING THE STRATEGY
C   INVOLVING PHI
C
    AMIN=ERTOTH(1)
    IMIN=1
    PHI=0.025D0*AMIN
    DO 223 J=2,IP
      IF (ERTOTH(J).LT.AMIN-PHI) THEN
        AMIN=ERTOTH(J)
        IMIN=J
        PHI=0.025D0*AMIN
      ENDIF
    223 CONTINUE
C
C   THE MODEL OPTIMAL MODEL DIMENSION HAS NOW BEEN DETERMINED (IMIN).

```

```
C  IMSL SUBROUTINE DCORVC IS USED TO CALCULATE THE COVARIANCE MATRIX
C  (USING ALL THE DATA) REQUIRED AS INPUT FOR DRSTEP.
```

```
C
C  IDO=0
C  NROW=NNPMM
C  NVAR=IPP1
C  LDX=NNPMM
C  IFRQ=0
C  IWT=0
C  MOPT=0
C  ICOPT=0
C  LDICOV=IPP1
C  LDINCD=1
C  CALL DCORVC(IDO,NROW,NVAR,XX,LDX,IFRQ,IWT,MOPT,
C  &          ICOPT,XMEANH,COVH,LDICOV,INCD,LDINCD,NOBS,
C  &          NMIS,SUMWT)
```

```
C
C  IMSL ROUTINE DRSTEP IS NOW USED TO SELECT THE FINAL OPTIMAL MODEL OF
C  DIMENSION IMIN
```

```
C
C  INVOKE=0
C  NVAR=IPP1
C  LDICOV=IPP1
C  DO 227 I=1,IP
C  LEVELH(I)=2
227 CONTINUE
C  LEVELH(IPP1)=-1
C  NFORCE=1
C  NSTEP=IMIN
C  ISTEP=1
C  NOBS=NNPMM
C  PIN=0.9999999D0
C  POUT=0.9999999D0
C  TOL=1.0D2*DMACH(4)
C  IPRINT=0
C  LDICOEF=IP
C  LDICOVS=IPP1
C  CALL DRSTEP(INVOKE,NVAR,COVH,LDICOV,LEVELH,NFORCE,
C  &          NSTEP,ISTEP,NOBS,PIN,POUT,TOL,IPRINT,
C  &          SCALEH,HISTH,IEND,AOVL,COEFH,LDICOEF,COVSH,
C  &          LDICOVS)
C  IT=0
C  DO 238 J=1,IP
C  ISELH(J)=0
C  IF (HISTH(J).GT.0) THEN
C  IT=IT+1
C  ISELH(IT)=J
C  ENDIF
238 CONTINUE
C  DO 240 J=1,IMIN
C  DO 239 I=1,NNPMM
C  XX2H(I,J)=XX(I,ISELH(J))
239 CONTINUE
240 CONTINUE
C  DO 245 J=1,IMIN
```



```

AMUSEL(1,J)=AMU(1,ISELH(J))
AMUSEL(2,J)=AMU(2,ISELH(J))
245 CONTINUE
C
C SUBROUTINE ERROR IS CALLED TO CALCULATE THE POST-SELECTION
C OPTIMAL AND ACTUAL ERROR RATES
C
CALL ERROR(ALAM,EP,NB,IMIN,AMUSEL,RSIG,XX2H,OPT,ACT)
IF (IS.EQ.0) OPT=0.5D0
EROPH=EROPH+OPT
ERACTH=ERACTH+ACT
EREH=EREH+AMIN
AMSEOPH=AMSEOPH+((AMIN-OPT)**2.0D0)
AUMSEH=AUMSEH+((AMIN-ACT)**2.0D0)
NUM=0
DO 250 J=1,IP
JJ=ISELH(J)
IF (JJ.NE.0) THEN
PSELVARH(JJ)=PSELVARH(JJ)+1.0D0
NUM=NUM+1
ENDIF
250 CONTINUE
PSELNUMH(NUM)=PSELNUMH(NUM)+1.0D0
IF (NUM.EQ.NONZERO) THEN
ISELR=1
DO 251 J=1, NONZERO
IF (HISTH(J).LT.0.1D0) ISELR=0
251 CONTINUE
CPCSH=CPCSH+ISELR
ENDIF
IF (NUM.GT.NONZERO) THEN
ISELR=1
DO 252 J=1, NONZERO
IF (HISTH(J).LT.0.1D0) ISELR=0
252 CONTINUE
IF (ISELR.EQ.1) SELOVERH=SELOVERH+1.0D0
ENDIF
IF (NUM.LT.NONZERO) THEN
ISELW=0
DO 253 J=NONZERO+1,IP
IF (HISTH(J).GT.0.1D0) ISELW=1
253 CONTINUE
IF (ISELW.EQ.0) SELUNDERH=SELUNDERH+1.0D0
ENDIF

ISELM=0
DO 254 J=NONZERO+1,IP
IF (HISTH(J).GT.0.1D0) ISELM=1
254 CONTINUE
IF (ISELM.EQ.1) THEN
NCOR=0
DO 255 J=1, NONZERO
IF (HISTH(J).GT.0.1D0) NCOR=NCOR+1
255 CONTINUE
IF ((NCOR.GT.0).AND.(NCOR.LT.NONZERO)) SELMIXH=

```

```
& SELMIXH+1.0D0
ENDIF
C
C THIS IS THE END OF THE CMV PROCEDURE ...
C
C AND ALSO THE MONTE CARLO LOOP
C
MC=MC+1
IF (MC.LT.NMC) GOTO 14

400 IF (PSELNUML(NONZERO).LT.0.5D0) PSELNUML(NONZERO)=-1.0D0
IF (PSELNUMH(NONZERO).LT.0.5D0) PSELNUMH(NONZERO)=-1.0D0
C
C DIVIDE THE SIMULATION COUNTERS BY THE NUMBER OF MC REPETITIONS
C
EREL=EREL/NMC
ERACTL=ERACTL/NMC
EROPTL=EROPTL/NMC
BIASL1=EREL-EROPTL
BIASL2=EREL-ERACTL
AMSEL1=AMSEOPL/NMC
AMSEL2=AUMSEL/NMC
CPCSL=CPCSL/PSELNUML(NONZERO)
PCSL=(CPCSL*PSELNUML(NONZERO))/NMC
SELOVERL=SELOVERL/NMC
SELUNDERL=SELUNDERL/NMC
SELMIXL=SELMIXL/NMC

ERES=ERES/NMC
BIASS1=ERES-EROPTL
BIASS2=ERES-ERACTL
AMSES1=AMSEOPS/NMC
AMSES2=AUMSES/NMC

EREH=EREH/NMC
ERACTH=ERACTH/NMC
EROPH=EROPH/NMC
BIASH1=EREH-EROPH
BIASH2=EREH-ERACTH
AMSEH1=AMSEOPH/NMC
AMSEH2=AUMSEH/NMC
CPCSH=CPCSH/PSELNUMH(NONZERO)
PCSH=(CPCSH*PSELNUMH(NONZERO))/NMC
SELOVERH=SELOVERH/NMC
SELUNDERH=SELUNDERH/NMC
SELMIXH=SELMIXH/NMC

DO 410 J=1,IP
PSELNUML(J)=PSELNUML(J)/NMC
PSELVARL(J)=PSELVARL(J)/NMC
PSELNUMH(J)=PSELNUMH(J)/NMC
PSELVARH(J)=PSELVARH(J)/NMC
410 CONTINUE

OPEN(1,FILE=FILEOUT,ACCESS='APPEND')
```

```

WRITE(1,600) IS,(AMU(2,J),J=1,IP)
WRITE(1,600)
WRITE(1,610) EROPTL,ERACTL
WRITE(1,610) BIASL1,AMSEL1
WRITE(1,610) BIASL2,AMSEL2
WRITE(1,620) (PSELVARL(J),J=1,IP)
WRITE(1,620) (PSELNUML(J),J=1,IP)
WRITE(1,620) CPCSL,PCSL,SELOVERL,SELUNDERL,SELMIXL
WRITE(1,600)
WRITE(1,610) EROPTL,ERACTL
WRITE(1,610) BIAS1,AMSES1
WRITE(1,610) BIAS2,AMSES2
WRITE(1,620) (PSELVARL(J),J=1,IP)
WRITE(1,620) (PSELNUML(J),J=1,IP)
WRITE(1,620) CPCSL,PCSL,SELOVERL,SELUNDERL,SELMIXL
WRITE(1,*)
WRITE(1,610) EROPTH,ERACTH
WRITE(1,610) BIASH1,AMSEH1
WRITE(1,610) BIASH2,AMSEH2
WRITE(1,620) (PSELVARH(J),J=1,IP)
WRITE(1,620) (PSELNUMH(J),J=1,IP)
WRITE(1,620) CPCSH,PCSH,SELOVERH,SELUNDERH,SELMIXH
WRITE(1,*)
CLOSE(1)
500 CONTINUE

600 FORMAT(I4,2X,5(F10.5,2X))
610 FORMAT(F12.6,2X,F12.6,2X,F12.6)
620 FORMAT(10(F10.5,2X))

1000 STOP
END

```

```

SUBROUTINE HOLDOUT(ICOL,IROW1,IROW2,XX1,XX2,XX3)

```

```

C
C THIS SUBROUTINE SPLITS THE DATA MATRIX XX1 INTO TWO SUBMATRICES
C INPUT : ICOL=NUMBER OF COLUMNS OF XX1 TO BE USED
C IROW1=THE NUMBER OF ROWS (FROM GROUP 1) OF XX1 TO BE
C WRITTEN IN XX2
C IROW2=THE NUMBER OF ROWS (FROM GROUP 2) OF XX1 TO BE
C WRITTEN IN XX2
C XX1=THE INPUT MATRIX
C OUTPUT : XX2=A SUB-MATRIX CONTAINING IROW=IROW1+IROW2 ROWS OF XX1
C XX3=A SUB-MATRIX CONTAINING THE REMAINING
C (NNPMM-IROW) ROWS OF XX1
C NOTE THAT THE ROWS OF XX1 ARE RANDOMLY ASSIGNED TO EITHER XX2 OR XX3
C

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)

```

```

PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
DIMENSION XX1(NNPMM,IPP1),XX2(NNPMM,IPP1),XX3(NNPMM,IPP1)
DIMENSION IPER1(NN),IPER2(MM)
IROW=IROW1+IROW2
CALL RNPER(NN,IPER1)
CALL RNPER(MM,IPER2)

```

```

DO 5 J=1,ICOL
DO 1 I=1,IROW1
XX2(I,J)=XX1(IPER1(I),J)
1 CONTINUE
DO 2 I=1,IROW2
XX2(IROW1+I,J)=XX1(NN+IPER2(I),J)
2 CONTINUE
DO 3 I=IROW1+1,NN
XX3(I-IROW1,J)=XX1(IPER1(I),J)
3 CONTINUE
DO 4 I=IROW2+1,MM
XX3(NN-IROW+I,J)=XX1(NN+IPER2(I),J)
4 CONTINUE
5 CONTINUE
RETURN
END

```

SUBROUTINE ERROR(ALAM,EP,NB,IT,AMU,RSIG,XX,OPT,ACT)

```

C
C THIS SUBROUTINE USES SIMULATION TO ESTIMATE THE POST-SELECTION
C ACTUAL AND OPTIMAL ERROR RATES
C
C INPUT : ALAM,EP=THE CONSTANTS USED IN THE JOHNSON TRANSFORMATION
C NB=THE NUMBER OF CASES TO BE GENERATED FROM EACH GROUP
C IT=THE NUMBER OF COLUMNS OF XX TO BE TAKEN INTO ACCOUNT
C AMU=THE MATRIX CONTAINING THE GROUP MEANS
C RSIG=THE MATRIX OBTAINED FROM THE CHOLESKY DECOMPOSITION
C OF THE COVARIANCE MATRIX
C XX=THE DATA MATRIX
C OUTPUT : OPT=THE OPTIMAL ERROR RATE
C ACT=THE ACTUAL ERROR RATE
C

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)

```

PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
DIMENSION XX(NNPMM,IP+1),S(IP,IP),SINV(IP,IP),XM1(IP),XM2(IP)
DIMENSION AMU(2,IP),RNX1(1,IP),XB(IP),RSIG(IP,IP),AMUN(2,IP)

```

CALL AVGVARV(IT,XX,S,SINV,XM1,XM2)

SUMO1=0.0D0

SUMO2=0.0D0

SUMA1=0.0D0

SUMA2=0.0D0

DO 3 I=1,2

DO 2 J=1,IT

AMUN(I,J)=ALAM+EP+AMU(I,J)

2 CONTINUE

3 CONTINUE

DO 100 IB=1,NB

CALL DRNMVN(1,IP,RSIG,IP,RNX1,1)

DO 5 J=1,IT

XB(J)=(ALAM*DEXP(RNX1(1,J)))+EP+AMU(1,J)

5 CONTINUE

```

SUM1=0.0D0
SUM2=0.0D0
DO 15 I1=1,IT
V1=XB(I1)-(AMUN(1,I1)+AMUN(2,I1))/2.0D0
V2=AMUN(1,I1)-AMUN(2,I1)
DO 10 I2=1,IT
V3=XB(I1)-(XM1(I1)+XM2(I1))/2.0D0
V4=XM1(I2)-XM2(I2)
SUM1=SUM1+V1*V2
SUM2=SUM2+V3*SINV(I1,I2)*V4
10 CONTINUE
15 CONTINUE
DTXB=SUM1
DSXB=SUM2
IF (DTXB.LE.0.0D0) SUMO1=SUMO1+1.0D0
IF (DSXB.LE.0.0D0) SUMA1=SUMA1+1.0D0
CALL DRNMVN(1,IP,RSIG,IP,RNX1,1)
DO 25 J=1,IT
XB(J)=(ALAM*DEXP(RNX1(1,J)))+EP+AMU(2,J)
25 CONTINUE

```

```

SUM1=0.0D0
SUM2=0.0D0
DO 35 I1=1,IT
V1=XB(I1)-(AMUN(1,I1)+AMUN(2,I1))/2.0D0
V2=AMUN(1,I1)-AMUN(2,I1)
DO 30 I2=1,IT
V3=XB(I1)-(XM1(I1)+XM2(I1))/2.0D0
V4=XM1(I2)-XM2(I2)
SUM1=SUM1+V1*V2
SUM2=SUM2+V3*SINV(I1,I2)*V4
30 CONTINUE
35 CONTINUE
DTXB=SUM1
DSXB=SUM2
IF (DTXB.GT.0.0D0) SUMO2=SUMO2+1.0D0
IF (DSXB.GT.0.0D0) SUMA2=SUMA2+1.0D0
100 CONTINUE
OPT=(SUMO1+SUMO2)/(2.0D0*NB)
ACT=(SUMA1+SUMA2)/(2.0D0*NB)
RETURN
END

```

SUBROUTINE AVGVARV(IT,XX,S,SINV,XM1,XM2)

```

C
C THIS SUBROUTINE CALCULATES THE MEAN VECTORS OF THE TWO GROUPS (XM1
C AND XM2) AS WELL AS THE POOLED COVARIANCE MATRIX (S) AND ITS
C INVERSE (SINV). THIS ROUTINE IS USED FOR THE MATRIX CONTAINING THE
C ORIGINAL NUMBER OF ROWS .
C INPUT: THE MATRIX XX(NNPM,IPP1) - THE FIRST NN ROWS OF XX CONTAIN THE
C OBSERVATIONS FROM GROUP1 AND THE NEXT MM ROWS CONTAIN THE
C OBSERVATIONS FROM GROUP2. ONLY THE FIRST IT COLUMNS ARE
C TAKEN INTO ACCOUNT
C

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
DIMENSION XX(NNPMM,IP+1),XX1(NN,IP),XX2(MM,IP)
DIMENSION XM1(IP),XM2(IP)
DIMENSION S(IP,IP),SINV(IP,IP),S1(IP,IP),S2(IP,IP)
EXTERNAL DCORVC,DLINDS
DO 10 I=1,NN
DO 5 J=1,IT
XX1(I,J)=XX(I,J)
5 CONTINUE
10 CONTINUE
DO 20 I=1,MM
DO 15 J=1,IT
XX2(I,J)=XX(NN+I,J)
15 CONTINUE
20 CONTINUE
IDO=0
NVAR=IT
IFRQ=0
IWT=0
MOPT=0
ICOPT=0
LDCOV=IP
LDINCD=1
NROW=NN
LDX=NN
CALL DCORVC(IDO,NROW,NVAR,XX1,LDX,IFRQ,IWT,MOPT,
&      ICOPT,XM1,S1,LDCOV,INCD,LDINCD,NOBS,
&      NMISS,SUMWT)
NROW=MM
LDX=MM
CALL DCORVC(IDO,NROW,NVAR,XX2,LDX,IFRQ,IWT,MOPT,
&      ICOPT,XM2,S2,LDCOV,INCD,LDINCD,NOBS,
&      NMISS,SUMWT)
NNPMMM2=NNPMM-2
DO 30 I=1,IT
DO 25 J=1,IT
S(I,J)=(NN-1)*S1(I,J)+(MM-1)*S2(I,J)/NNPMMM2
25 CONTINUE
30 CONTINUE
CALL DLINDS(IT,S,IP,SINV,IP)
RETURN
END

```

SUBROUTINE AVGVAR3(N,M,NPM,IT,XX,S,SINV,XM1,XM2)

```

C
C THIS SUBROUTINE CALCULATES THE MEAN VECTORS OF THE TWO GROUPS (XM1
C AND XM2) AS WELL AS THE POOLED COVARIANCE MATRIX (S) AND ITS
C INVERSE (SINV). THIS ROUTINE IS USED FOR THE MATRIX CONTAINING
C ONLY A SUBSET OF THE ORIGINAL NUMBER OF ROWS.
C INPUT: THE MATRIX XX(NPM,IPP1) - THE FIRST N ROWS OF XX CONTAIN THE
C OBSERVATIONS FROM GROUP1 AND THE NEXT M ROWS CONTAIN THE
C OBSERVATIONS FROM GROUP2. ONLY THE FIRST IT COLUMNS ARE
C TAKEN INTO ACCOUNT

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
DIMENSION XX(NNPMM,IP+1),XX1(N,IP),XX2(M,IP)
DIMENSION XM1(IP),XM2(IP)
DIMENSION S(IP,IP),SINV(IP,IP),S1(IP,IP),S2(IP,IP)
EXTERNAL DCORVC,DLINDS
DO 10 I=1,N
DO 5 J=1,IT
XX1(I,J)=XX(I,J)
5 CONTINUE
10 CONTINUE
DO 20 I=1,M
DO 15 J=1,IT
XX2(I,J)=XX(N+I,J)
15 CONTINUE
20 CONTINUE
IDO=0
NVAR=IT
IFRQ=0
IWT=0
MOPT=0
ICOPT=0
LDCOV=IP
LDINCD=1
NROW=N
LDX=N
CALL DCORVC(IDO,NROW,NVAR,XX1,LDX,IFRQ,IWT,MOPT,
& ICOPT,XM1,S1,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)
NROW=M
LDX=M
CALL DCORVC(IDO,NROW,NVAR,XX2,LDX,IFRQ,IWT,MOPT,
& ICOPT,XM2,S2,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)
NPMM2=NPM-2
DO 30 I=1,IT
DO 25 J=1,IT
S(I,J)=((N-1)*S1(I,J)+(M-1)*S2(I,J))/NPMM2
25 CONTINUE
30 CONTINUE
CALL DLINDS(IT,S,IP,SINV,IP)
RETURN
END

```

SUBROUTINE WFSTAR(IT,XXSEL,ERSMOOTH)

```

C
C THIS SUBROUTINE CALCULATES THE POST-SELECTION NSp* ERROR RATE ESTIMATOR
C SUGGESTED BY SNAPINN AND KNOKE (1989)
C INPUT : THE MATRIX XXSEL CONTAINS ALL THE DATA, BUT ONLY THE SELECTED
C VARIABLES
C IT IS THE NUMBER OF SELECTED VARIABLES
C OUTPUT : ERSMOOTH IS THE NSp*-ESTIMATE (SNAPINN AND KNOKE, 1989)
C

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)

```

PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM)
DIMENSION XXSEL(NNPMM,IP+1),XV(IP)
DIMENSION S(IP,IP),SINV(IP,IP),XM1(IP),XM2(IP)
C1=1.0D0*(NNPMM-IP-3.0D0)/(NNPMM-2.0D0)
C2=1.0D0*IP*NNPMM/(NN*MM)
C2DC1=C2/C1
CALL AVGVARV(IT,XXSEL,S,SINV,XM1,XM2)
SUM=0.0D0
DO 50 IUIT=1,NNPMM
DO 15 J=1,IT
XV(J)=XXSEL(IUIT,J)
15 CONTINUE
SUM1=0.0D0
SUM2=0.0D0
DO 25 I1=1,IT
DO 20 I2=1,IT
V1=XV(I1)-(XM1(I1)+XM2(I1))/2.0D0
V2=XM1(I2)-XM2(I2)
V3=XM1(I1)-XM2(I1)
SUM1=SUM1+V1*SINV(I1,I2)*V2
SUM2=SUM2+V3*SINV(I1,I2)*V2
20 CONTINUE
25 CONTINUE
WW=SUM1
AMAH2=SUM2
AMAH=DSQRT(SUM2)
IF (IUIT.LE.NN) THEN
  BKON=AMAH2/(C1*AMAH2-C2)-(NN-1.0D0)/NN
  BKON=DSQRT(BKON)
  ARG=-WW/(BKON*AMAH)
  IF (AMAH2.GT.C2DC1) SUM=SUM+DNORDF(ARG)
  IF (AMAH2.LE.C2DC1) SUM=SUM+0.5D0
ENDIF
IF (IUIT.GT.NN) THEN
  BKON=AMAH2/(C1*AMAH2-C2)-(MM-1.0D0)/MM
  BKON=DSQRT(BKON)
  ARG=WW/(BKON*AMAH)
  IF (AMAH2.GT.C2DC1) SUM=SUM+DNORDF(ARG)
  IF (AMAH2.LE.C2DC1) SUM=SUM+0.5D0
ENDIF
50 CONTINUE
ERSMOOTH=SUM/NNPMM
RETURN
END

```

SUBROUTINE LOO(II,X,X1)

```

C
C THIS SUBROUTINE OMITTS ROW II OF THE MATRIX X. X1 IS THE X-MATRIX
C WITH ROW II DELETED
C INPUT : X(NNPMM,IP+1)=THE DATA MATRIX WITH ALL THE ROWS
C II=THE NUMBER OF THE ROW TO BE DELETED
C OUTPUT : X1(NNPMM-1,IP+1)=THE DATA MATRIX WITH ROW II DELETED
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)

```



```

DIMENSION X(NNPMM,IPP1),X1(NNPMM-1,IPP1)
N=NNPMM
IF (II.EQ.1) THEN
  DO 5 I=1,N-1
  DO 1 J=1,IPP1
  X1(I,J)=X(I+1,J)
1   CONTINUE
5   CONTINUE
ENDIF
IF ((II.GT.1).AND.(II.LT.N)) THEN
  DO 15 I=1,II-1
  DO 10 J=1,IPP1
  X1(I,J)=X(I,J)
10  CONTINUE
15  CONTINUE
  DO 25 I=II,N-1
  DO 20 J=1,IPP1
  X1(I,J)=X(I+1,J)
20  CONTINUE
25  CONTINUE
ENDIF
IF (II.EQ.N) THEN
  DO 35 I=1,N-1
  DO 30 J=1,IPP1
  X1(I,J)=X(I,J)
30  CONTINUE
35  CONTINUE
ENDIF
RETURN
END

```

SUBROUTINE WF(MIN,II,X1,XV,WW,AMAH)

```

C
C THIS SUBROUTINE CALCULATES WW, THE VALUE OF THE ANDERSON CLASSIFICATION
C STATISTIC BASED ON THE DATA IN X1 FOR THE OMITTED CASE XV.
C IT ALSO CALCULATES THE SAMPLE MAHALANOBIS DISTANCE BETWEEN
C THE GROUPS BASED ON THE DATA IN X1.
C INPUT : MIN=INDICATOR VECTOR OF DIMENSION IP TO IDENTIFY SELECTED
C         VARIABLES
C         II=NUMBER OF THE DELETED ROW
C         X1=MATRIX CONTAINING ALL THE DATA WITH ROW II OMITTED
C         XV=THE OMITTED CASE
C OUTPUT : WW=THE VALUE OF THE ANDERSON CLASSIFICATION STATISTIC FOR
C           CASE WV
C           AMAH=THE SAMPLE MAHALANOBIS DISTANCE
C
C IMPLICIT DOUBLE PRECISION (A-H,O-Z)
C PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM)
C DIMENSION X1(NNPMM-1,IP+1),XX(NNPMM-1,IP+1),XV(IP)
C DIMENSION S(IP,IP),SINV(IP,IP),XM1(IP),XM2(IP)
C DIMENSION MIN(IP)
C
C THE INDICATOR VECTOR MIN IS NOW USED TO FORM A MATRIX XX, CONTAINING ONLY
C THE SELECTED VARIABLES

```

```

C
  IF (II.LE.NN) THEN
    N1=NN-1
    N2=MM
  ENDIF
  IF (II.GT.NN) THEN
    N1=NN
    N2=MM-1
  ENDIF
  DO 10 I=1,N1
    IT=0
    DO 5 J=1,IP
      IF (MIN(J).EQ.1) THEN
        IT=IT+1
        XX(I,IT)=X1(I,J)
      ENDIF
5    CONTINUE
10   CONTINUE
    DO 20 I=1,N2
      IT=0
      DO 15 J=1,IP
        IF (MIN(J).EQ.1) THEN
          IT=IT+1
          XX(N1+I,IT)=X1(N1+I,J)
        ENDIF
15   CONTINUE
20   CONTINUE
C
C   SUBROUTINE AVGVARD IS USED TO CALCULATE THE MEANS OF THE TWO GROUPS
C   AS WELL AS THE POOLED COVARIANCE MATRIX AND ITS INVERSE (ONLY THE SELEC-
C   TED VARIABLES ARE TAKEN INTO ACCOUNT)
C
N1PN2=N1+N2
CALL AVGVARD(N1,N2,N1PN2,IT,XX,S,SINV,XM1,XM2)
C
C   THE SAMPLE MAHALANOBIS DISTANCE BASED ONLY ON THE SELECTED VARIABLES IS
C   NOW CALCULATED. THE ANDERSON CLASSIFICATION STATISTIC FOR THE OMITTED
C   CASE (ALSO BASED ONLY ON THE SELECTED VARIABLES) IS ALSO CALCULATED.
C
SUM1=0.0D0
SUM2=0.0D0
DO 95 I1=1,IT
DO 90 I2=1,IT
V1=XV(I1)-(XM1(I1)+XM2(I1))/2.0D0
V2=XM1(I2)-XM2(I2)
V3=XM1(I1)-XM2(I1)
SUM1=SUM1+V1*SINV(I1,I2)*V2
SUM2=SUM2+V3*SINV(I1,I2)*V2
90  CONTINUE
95  CONTINUE
WW=SUM1
AMAH=DSQRT(SUM2)
RETURN
END

```

SUBROUTINE AVGVARD(N,M,NPM,IT,XX,S,SINV,XM1,XM2)

```

C
C THIS SUBROUTINE CALCULATES THE MEAN VECTORS OF THE TWO GROUPS (XM1
C AND XM2) AS WELL AS THE POOLED COVARIANCE MATRIX (S) AND ITS
C INVERSE (SINV). THIS ROUTINE IS USED FOR THE MATRIX WITH ONE ROW
C OMITTED
C INPUT: THE MATRIX XX(NPM,IP+1) - THE FIRST N ROWS OF XX CONTAIN THE
C OBSERVATIONS FROM GROUP1 AND THE NEXT M ROWS CONTAIN THE
C OBSERVATIONS FROM GROUP2. ONLY THE FIRST IT COLUMNS ARE
C TAKEN INTO ACCOUNT
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10)
DIMENSION XX(NPM,IP+1),XX1(N,IP),XX2(M,IP)
DIMENSION XM1(IP),XM2(IP)
DIMENSION S(IP,IP),SINV(IP,IP),S1(IP,IP),S2(IP,IP)
EXTERNAL DCORVC,DLINDS
DO 10 I=1,N
DO 5 J=1,IT
XX1(I,J)=XX(I,J)
5 CONTINUE
10 CONTINUE
DO 20 I=1,M
DO 15 J=1,IT
XX2(I,J)=XX(N+I,J)
15 CONTINUE
20 CONTINUE
IDO=0
NVAR=IT
IFRQ=0
IWT=0
MOPT=0
ICOPT=0
LDCOV=IP
LDINCD=1
NROW=N
LDX=N
CALL DCORVC(IDO,NROW,NVAR,XX1,LDX,IFRQ,IWT,MOPT,
& ICOPT,XM1,S1,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)
NROW=M
LDX=M
CALL DCORVC(IDO,NROW,NVAR,XX2,LDX,IFRQ,IWT,MOPT,
& ICOPT,XM2,S2,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)
NPMM2=NPM-2
DO 30 I=1,IT
DO 25 J=1,IT
S(I,J)=((N-1)*S1(I,J)+(M-1)*S2(I,J))/NPMM2
25 CONTINUE
30 CONTINUE
CALL DLINDS(IT,S,IP,SINV,IP)
RETURN
END

```

PROGRAM 3

```

C NOTE THAT THIS PROGRAM IS NOT GIVEN IN ITS ENTIRETY HERE. MISSING IS
C SUBROUTINE POLY, THE ROUTINE USED TO PERFORM A LOGISTIC REGRESSION
C ANALYSIS. THIS ROUTINE IS GIVEN AS PART OF PROGRAM 1, AND IT IS
C THEREFORE NOT REPEATED HERE.
C
C IN THIS PROGRAM A MONTE CARLO SIMULATION STUDY IS DONE TO COMPARE THE
C FOLLOWING POST-SELECTION ERROR RATE ESTIMATORS IN LOGISTIC REGRESSION:
C
C 1. THE CROSS MODEL VALIDATION TECHNIQUE WITH AN ALL POSSIBLE SUBSETS
C APPROACH BASED ON  $C_p$  AS INNER CRITERION
C 2. THE BOOTSTRAP METHOD PROPOSED BY EFRON AND GONG (1983)
C AND GONG (1986)
C
C IN THIS PROGRAM IT IS ASSUMED THAT THE FEATURE VARIABLES ARE
C EQUI-CORRELATED (COMMON CORRELATION =  $\rho$ ) AND NORMALLY DISTRIBUTED
C
C PARAMETERS :
C IP=THE TOTAL NUMBER OF AVAILABLE FEATURE VARIABLES
C NN=THE SIZE OF THE TRAINING DATA SET FROM GROUP 1
C MM=THE SIZE OF THE TRAINING DATA SET FROM GROUP 2
C NNPMM=NN+MM=THE TOTAL SIZE OF THE TRAINING DATA SET
C NMC=NUMBER OF MONTE CARLO REPETITIONS
C KLASS=THE NUMBER OF GROUPS
C NB=NUMBER OF SIMULATION REPETITIONS USED PER GROUP TO
C ESTIMATE THE POST-SELECTION ACTUAL ERROR RATE
C KB=THE NUMBER OF BOOTSTRAP REPETITIONS USED TO OBTAIN THE BOOTSTRAP
C ESTIMATE
C
C THE FOLLOWING IMSL-SUBROUTINES ARE USED IN THE MAIN PROGRAM:
C 1. DLINDS: FINDS THE INVERSE OF A GIVEN COVARIANCE MATRIX
C 2. DCHFAC: FINDS THE CHOLSKY DECOMPOSITION OF A GIVEN MATRIX
C 3. DRNMVN: GENERATES VALUES FROM A MULTIVARIATE NORMAL DISTRIBUTION
C 4. DCORVC: COMPUTES A COVARIANCE OR CORRELATION MATRIX
C 5. DRBEST: SELECTS THE BEST MULTIPLE LINEAR REGRESSION MODELS -
C CAN ALSO BE ADAPTED AND APPLIED FOR THIS PURPOSE IN DA
C
C IMPLICIT DOUBLE PRECISION (A-H,O-Z)
C LOGICAL DIFNAN
C PARAMETER (IP=10,NN=50,MM=50,NNPMM=NN+MM,IPP1=IP+1,NMC=200,
C &KLASS=2,NB=500,KB=200,RHO=0.9D0)
C
C THE FOLLOWING STATEMENT INITIALISES ARGUMENTS FOR SUBROUTINE DRBEST
C
C PARAMETER(NVAR1=IPP1,LDCOV1=NVAR1,NOBS1=NNPMM-1,ICRIT1=3,
C &NBEST1=1,NGOOD1=1,IPRINT1=0,LDCOEF1=NBEST1*IP,
C &NSIZE1=IP,LINDVAR1=NGOOD1*NSIZE1*(NSIZE1+1)/2,NTBEST1=NBEST1)
C
C DIMENSION AMU(2,IP),SIGMAM(IP,IP),SIGINV(IP,IP),RSIG(IP,IP)
C DIMENSION RNX1(NN,IP),RNX2(MM,IP),RESP(NNPMM)
C DIMENSION XX(NNPMM,IPP1),XMEAN(IPP1)
C DIMENSION XBOOT(NNPMM,IP),XXB(NNPMM,IP+2)

```

```

C
DIMENSION CRIT1(NGOOD1*NSIZE1),COEF1(LDCOEF1,5)
DIMENSION ICRITX1(NSIZE1+1),IVARX1(NSIZE1+1),INDVAR1(LINDVAR1)
DIMENSION ICOEFX1(NTBEST1+1),IND01(IP,IP)
C
DIMENSION XVH(IP),XX1H(NNPMM-1,IPP1)
DIMENSION ERRORH(NNPMM,IP),ERTOTH(IP)
DIMENSION PSELVARH(IP),PSELNUMH(IP),PSELVARB(IP),PSELNUMB(IP)
DIMENSION HISTH(IP),HISTB(IP)
DIMENSION MINH(IP),ISELH(IP),ISELB(IP)
C
DIMENSION XPOLY(NNPMM-1,IP),BETA1(IPP1,1)
DIMENSION XPOLYF(NNPMM,IP),BETA1F(IPP1,1)
DIMENSION ICLASS(NNPMM-1),ICLASSF(NNPMM)
C
DIMENSION COVW(IPP1,IPP1),COVWF(IPP1,IPP1)
DIMENSION V(NNPMM-1),Z(NNPMM-1),VF(NNPMM),ZF(NNPMM)
DIMENSION XXW(NNPMM-1,IP+2),XXWF(NNPMM,IP+2)
C
DIMENSION NOCONV(0:10),INCD(1,1)

CHARACTER*70 FILEOUT
EXTERNAL DIFNAN

NITER=100
DSMALL=0.1D0
FILEOUT='/nor.d'
C
C NONZERO IS THE NUMBER OF NONZERO ELEMENTS OF THE MEAN VECTOR OF THE
C SECOND GROUP - ALL THE ELEMENTS OF THE MEAN VECTOR OF THE FIRST
C GROUP ARE TAKEN EQUAL TO ZERO
C
NONZERO=1
DO 2 I=1,IP
SIGMAM(I,I)=1.0D0
DO 1 J=1,IP
SIGMAM(I,J)=RHO
1 CONTINUE
2 CONTINUE
DO 7 I=1,IP
AMU(1,I)=0.0D0
7 CONTINUE

CALL DLINDS(IP,SIGMAM,IP,SIGINV,IP)
SUMSIG=0.0D0
DO 9 I=1,NONZERO
DO 8 J=1,NONZERO
SUMSIG=SUMSIG+SIGINV(I,J)
8 CONTINUE
9 CONTINUE
C
C THE LOOP UP TO 500 SYSTEMATICALLY INCREASES THE MAHALANOBIS DISTANCE
C BETWEEN THE TWO GROUPS
C THE FOLLOWING SIMULATION COUNTERS ARE ALSO INITIALISED:
C 1. PSELVAR(H/B)(J): THE ESTIMATED PROBABILITY OF CHOOSING VARIABLE J

```

- C 2. PSELNUM(H/B)(J): THE ESTIMATED PROBABILITY OF CHOOSING A MODEL WITH
 C J VARIABLES
 C 3. ERE(H/B): THE AVERAGE ESTIMATED ACTUAL ERROR RATE
 C 4. AUMSE(H/B): THE UMSE FOR ESTIMATION OF THE ACTUAL ERROR RATE
 C 5. ERACT(H/B): THE AVERAGE POST-SELECTION ACTUAL ERROR RATE
 C

```
DO 500 IS=0,4
D2=1.0D0*IS
DO 12 J=1, NONZERO
AMU(2,J)=DSQRT(D2/SUMSIG)
PSELVARH(J)=0.0D0
PSELNUMH(J)=0.0D0
PSELVARB(J)=0.0D0
PSELNUMB(J)=0.0D0
```

- 12 CONTINUE
 IF (NONZERO.LT.IP) THEN
 DO 13 J=NONZERO+1,IP
 AMU(2,J)=0.0D0
 PSELVARH(J)=0.0D0
 PSELNUMH(J)=0.0D0
 PSELVARB(J)=0.0D0
 PSELNUMB(J)=0.0D0

- 13 CONTINUE
 ENDIF

```
EREH=0.0D0
AUMSEH=0.0D0
ERACTH=0.0D0
CPCSH=0.0D0
SELOVERH=0.0D0
SELUNDERH=0.0D0
SELMIXH=0.0D0
EREB=0.0D0
AUMSEB=0.0D0
ERACTB=0.0D0
CPCSB=0.0D0
SELOVERB=0.0D0
SELUNDERB=0.0D0
SELMIXB=0.0D0
```

```
TOL=1.0D2*DMACH(4)
CALL DCHFAC(IP, SIGMAM, IP, TOL, IRANK, RSIG, IP)
```

- C
 C THE SIMULATION LOOP BEGINS, AND THE FIRST STEP IS TO GENERATE THE
 C REQUIRED TRAINING DATA SETS FROM THE RELEVANT MULTIVARIATE NORMAL
 C DISTRIBUTIONS - NOTE THAT THE MEAN VALUES ARE ADDED SEPARATELY
 C

```
MC=0
```

- 14 CALL DRNMVN(NN, IP, RSIG, IP, RNX1, NN)
 CALL DRNMVN(MM, IP, RSIG, IP, RNX2, MM)
 DO 16 I=1, NN
 DO 15 J=1, IP
 RNX1(I, J)=RNX1(I, J)+AMU(1, J)

- 15 CONTINUE
 16 CONTINUE

```
DO 20 I=1,MM
DO 19 J=1,IP
RNX2(I,J)=RNX2(I,J)+AMU(2,J)
19 CONTINUE
20 CONTINUE
C
C ICLASSF AND RESP BOTH CONTAIN THE RESPONSE VARIABLE VALUES
C INDICATING GROUP MEMBERSHIP
C
DO 25 I=1,NN
ICLASSF(I)=0
RESP(I)=0.0D0
25 CONTINUE
DO 30 I=NN+1,NNPMM
ICLASSF(I)=1
RESP(I)=1.0D0
30 CONTINUE
C
C A SINGLE DATA MATRIX XX (NNPMM x IP+1) IS FORMED. THE FIRST IP COLUMNS
C CONTAIN THE FEATURE VARIABLES, WHILE COLUMN (IP+1) CONTAINS THE RESPONSE
C VARIABLE VALUES INDICATING GROUP MEMBERSHIP.
C
DO 45 J=1,IP
DO 35 I=1,NN
XX(I,J)=RNX1(I,J)
35 CONTINUE
DO 40 I=1,MM
XX(NN+I,J)=RNX2(I,J)
40 CONTINUE
45 CONTINUE
DO 50 I=1,NN
XX(I,IP+1)=RESP(I)
50 CONTINUE
DO 55 I=1,MM
XX(NN+I,IP+1)=RESP(NN+I)
55 CONTINUE
C
C THE CMV METHOD STARTS HERE
C
DO 65 I=1,NNPMM
DO 60 J=1,IP
ERRORH(I,J)=0.0D0
60 CONTINUE
65 CONTINUE
C
C SUBROUTINE LOO IS CALLED TO OMIT THE ROWS ONE BY ONE
C
DO 200 II=1,NNPMM
CALL LOO(II,XX,XX1H)
IF (II.LE.NN) THEN
  NN1=NN-1
  MM1=MM
ENDIF
IF (II.GT.NN) THEN
  NN1=NN
```

```
MM1=MM-1
ENDIF
DO 66 I=1,NN1
ICLASS(I)=0
66 CONTINUE
DO 67 I=NN1+1,NNPMM-1
ICLASS(I)=1
67 CONTINUE
DO 70 I=1,NNPMM-1
DO 69 J=1,IP
XPOLY(I,J)=XX1H(I,J)
69 CONTINUE
70 CONTINUE

IW=0
NITER=100
DSMALL=0.1D0
NPMM1=NNPMM-1

C
C SUBROUTINE POLY IS CALLED TO CALCULATE LOGISTIC REGRESSION
C COEFFICIENTS FROM THE DATA CONTAINING ALL THE VARIABLES BUT WITH
C ROW II DELETED. THIS IS DONE TO OBTAIN THE INITIAL
C BETA-ESTIMATES TO BE USED TO CALCULATE THE Z(I) (DEPENDENT
C VARIABLE) AND THE V(I) (WEIGHTS) TO BE USED AS INPUT IN A
C LINEAR REGRESSION SELECTION PROGRAM (DRBEST).
C
CALL POLY(IW,ICLASS,NITER,NPMM1,KLASS,IP,DSMALL,XPOLY,BETA1)
C
C RESET THE VALUES OF ICLASS (IT IS CHANGED BY SUBROUTINE POLY) AND
C TEST FOR CONVERGENCE
C
DO 71 I=1,NN1
ICLASS(I)=0
71 CONTINUE
DO 72 I=NN1+1,NNPMM-1
ICLASS(I)=1
72 CONTINUE

IF (IW.EQ.1) THEN
NOCONV(IS+1)=NOCONV(IS+1)+1
GOTO 14
ENDIF

C
C USE THE BETA1 COEFFICIENTS TO CALCULATE THE WEIGHTS AND DEPENDENT
C VARIABLE VALUES TO BE USED AS INPUT IN DRBEST. NOTE THAT WE ONCE
C MORE TEST WHETHER THE ITERATIVE PROCEDURE DID IN FACT CONVERGE TO
C STABLE VALUES.
C
DO 75 I=1,NNPMM-1
SUM1=BETA1(1,I)
DO 74 J=1,IP
SUM1=SUM1+BETA1(J+1,I)*XX1H(I,J)
74 CONTINUE
ESUM1=DEXP(SUM1)
PI1=ESUM1/(1.0D0+ESUM1)
```



```

V(I)=PI1*(1.0D0-PI1)
IF (DIFNAN(V(I))) GOTO 14
Z(I)=SUM1+(1.0D0*ICLASS(I)-PI1)/V(I)
IF (DIFNAN(Z(I))) GOTO 14
75 CONTINUE

DO 85 I=1,NNPMM-1
DO 80 J=1,IP
XXW(I,J)=XX1H(I,J)
80 CONTINUE
XXW(I,IPP1)=Z(I)
XXW(I,IPP1+1)=V(I)
85 CONTINUE
C
C IMSL ROUTINE DCORVC IS USED TO CALCULATE THE COVARIANCE MATRIX REQUIRED
C AS INPUT FOR DRBEST
C
IDO=0
NROW=NNPMM-1
NVAR=IPP1
LDX=NNPMM-1
IFRQ=0
IWT=IPP1+1
MOPT=0
ICOPT=0
LDCOV=IPP1
LDINCD=1
NOBS=NNPMM-1

CALL DCORVC(IDO,NROW,NVAR,XXW,LDX,IFRQ,IWT,MOPT,
&ICOPT,XMEAN,COVW,LDCOV,INCD,LDINCD,NOBS,NMISS,SUMWT)
C
C IMSL SUBROUTINE DRBEST IS USED TO IDENTIFY THE BEST MODEL OF EACH
C DIMENSION 1,....,IP
C
CALL DRBEST(NVAR1,COVW,LDCOV1,NOBS,ICRIT1,NBEST1,NGOOD1,IPRINT1,
&ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)

DO 169 IK=1,IP
DO 166 J=1,IP
IND01(IK,J)=0
166 CONTINUE
IF (IK.EQ.1) IB=1
IF (IK.EQ.2) IB=ICRITX1(2)-ICRITX1(1)+1
IF (IK.GT.2) THEN
IB=ICRITX1(2)-ICRITX1(1)+1
DO 167 J=2,IK-1
IB=IB+(ICRITX1(J+1)-ICRITX1(J))*J
167 CONTINUE
ENDIF
DO 168 I=0,IK-1
III=INDVAR1(IB+I)
IND01(IK,III)=1
168 CONTINUE
169 CONTINUE

```

```
DO 195 ID=1,IP
ITEL=0
DO 170 J=1,IP
MINH(J)=0
IF (IND01(ID,J).GT.0) THEN
  MINH(J)=1
  ITEL=ITEL+1
  XVH(ITEL)=XX(I,I)
ENDIF
170 CONTINUE
C
C THE SMOOTHED LOSS (SMLOSS) ASSOCIATED WITH THE OMITTED CASE IS
C CALCULATED FOR THE BEST ID-DIMENSIONAL MODEL
C
IDEM=ID
CALL WF(IDEM,MINH,I,XX1H,XVH,SMLOSS)
ERRORH(I,ID)=SMLOSS
195 CONTINUE
200 CONTINUE
C
C THIS IS THE END OF THE LOOP WHERE THE CASES ARE OMITTED ONE BY ONE
C
C THE AVERAGE LOSS ASSOCIATED WITH EACH MODEL DIMENSION IS CALCULATED
C AND THE OPTIMAL MODEL DIMENSION IS IDENTIFIED BY FINDING THE
C MINIMUM AVERAGE LOSS
C
DO 220 J=1,IP
ERTOTH(J)=0.0D0
DO 210 I=1,NNPMM
ERTOTH(J)=ERTOTH(J)+ERRORH(I,J)
210 CONTINUE
ERTOTH(J)=ERTOTH(J)/NNPMM
220 CONTINUE
AMIN=ERTOTH(1)
IMIN=1
DO 221 J=2,IP
IF (ERTOTH(J).LT.AMIN) THEN
  AMIN=ERTOTH(J)
  IMIN=J
ENDIF
221 CONTINUE
C
C IMIN IS THE OPTIMAL MODEL DIMENSION
C
IWF=0
NNPMMF=NNPMM
KLASSF=2
IPF=IP
DO 225 I=1,NNPMM
DO 224 J=1,IP
XPOLYF(I,J)=XX(I,J)
224 CONTINUE
225 CONTINUE
NITERF=100
```

DSMALLF=0.1D0

C
 C SUBROUTINE POLY IS CALLED TO CALCULATE THE LOGISTIC REGRESSION
 C COEFFICIENTS. THE DATA CONTAINING ALL THE VARIABLES AND THE
 C DATA ON ALL THE CASES ARE USED. THIS IS DONE TO OBTAIN THE INITIAL
 C BETA-ESTIMATES TO BE USED TO CALCULATE THE Z(I) (DEPENDENT
 C VARIABLE) AND THE V(I) (WEIGHTS) TO BE USED AS INPUT IN AN
 C LINEAR REGRESSION SELECTION PROGRAM (DRBEST) TO SELECT THE FINAL
 C MODEL OF THE OPTIMAL DIMENSION (IMIN) IDENTIFIED BY MINIMISING THE
 C CMV-CRITERION (AVERAGE LOSS).

C
 CALL POLY(IWF,ICLASSF,NITERF,NNPMMF,KLASSF,IPF,DSMALLF,
 &XPOLYF,BETA1F)

DO 226 I=1,NN

ICLASSF(I)=0

226 CONTINUE

DO 227 I=NN+1,NNPMM

ICLASSF(I)=1

227 CONTINUE

IF (IWF.EQ.1) THEN

NOCONV(IS+1)=NOCONV(IS+1)+1

GOTO 14

ENDIF

DO 230 I=1,NNPMMF

SUM1=BETA1F(1,1)

DO 229 J=1,IP

SUM1=SUM1+BETA1F(J+1,1)*XPOLYF(I,J)

229 CONTINUE

ESUM1=DEXP(SUM1)

PI1=ESUM1/(1.0D0+ESUM1)

VF(I)=PI1*(1.0D0-PI1)

IF (DIFNAN(VF(I))) GOTO 14

ZF(I)=SUM1+(1.0D0*ICLASSF(I)-PI1)/VF(I)

IF (DIFNAN(ZF(I))) GOTO 14

230 CONTINUE

DO 235 I=1,NNPMMF

DO 234 J=1,IP

XXWF(I,J)=XPOLYF(I,J)

234 CONTINUE

XXWF(I,IPP1)=ZF(I)

XXWF(I,IPP1+1)=VF(I)

235 CONTINUE

IDO=0

NROW=NNPMMF

NVAR=IPP1

LDX=NNPMMF

IFRQ=0

IWT=IPP1+1

MOPT=0

ICOPT=0

LDCOV=IPP1

```
LDINCD=1
NOBS=NNPMMF
C
C   IMSL ROUTINE DCORVC IS USED ON ALL THE DATA TO CALCULATE THE
C   COVARIANCE MATRIX REQUIRED AS INPUT FOR DRBEST
C
CALL DCORVC(IDO,NROW,NVAR,XXWF,LDX,IFRQ,IWT,MOPT,
&   ICOPT,XMEAN,COVWF,LDCOV,INCD,LDINCD,NOBS,
&   NMISS,SUMWT)
C
C   IMSL ROUTINE DRBEST IS USED TO IDENTIFY THE BEST MODEL OF DIMENSION
C   IMIN (THE OPTIMAL DIMENSION DETERMINED BY MINIMISING THE CMV-CRITERION)
C
CALL DRBEST(NVAR1,COVWF,LDCOV1,NOBS,ICRIT1,NBEST1,NGOOD1,IPRINT1,
&ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)

DO 245 J=1,IP
HISTH(J)=0
245 CONTINUE
IF (IMIN.EQ.1) IB=1
IF (IMIN.EQ.2) IB=ICRITX1(2)-ICRITX1(1)+1
IF (IMIN.GT.2) THEN
  IB=ICRITX1(2)-ICRITX1(1)+1
  DO 246 J=2,IMIN-1
  IB=IB+(ICRITX1(J+1)-ICRITX1(J))*J
246 CONTINUE
ENDIF
DO 250 I=0,IMIN-1
III=INDVAR1(IB+I)
HISTH(III)=1
250 CONTINUE

ITEL=0
DO 258 J=1,IP
ISELH(J)=0
IF (HISTH(J).GT.0) THEN
  ITEL=ITEL+1
  ISELH(ITEL)=J
ENDIF
258 CONTINUE
C
C   SUBROUTINE ERROR IS CALLED TO CALCULATE THE POST-SELECTION ACTUAL
C   ERROR RATE OF THE MODEL SELECTED BY MEANS OF THE CMV TECHNIQUE
C
IW=0
CALL ERROR(AMU,RSIG,XX,IMIN,HISTH,ACTH,IW)
IF (IW.NE.0) GOTO 14
C
C   THIS IS THE END OF THE CMV PROCEDURE
C
C
C   THE BOOTSTRAP METHOD STARTS HERE
C
C   THE BEST MODEL (USING ALL THE DATA) IS IDENTIFIED BY USING IMSL ROUTINE
```

```
C DRBEST ON ALL THE DATA (WITH THE NECESSARY TRANSFORMATION INVOLVING Z(I)
C AND V(I)). THE MODEL THAT MINIMISES THE Cp CRITERION IS CHOSEN AS THE
C BEST MODEL. THE BOOTSTRAP METHOD PROPOSED BY EFRON AND GONG (1983)
C AND GONG (1986) WILL BE USED TO ESTIMATE THE POST-SELECTION ACTUAL ERROR
C RATE OF THIS MODEL.
```

```
C
RMIN=CRIT1(1)
IBOOT=1
DO 270 I=2,IP
IF (CRIT1(I).LT.RMIN) THEN
  RMIN=CRIT1(I)
  IBOOT=I
ENDIF
270 CONTINUE
C
C IBOOT IS THE DIMENSION OF THE OPTIMAL MODEL
```

```
C
DO 271 J=1,IP
HISTB(J)=0
271 CONTINUE
IF (IBOOT.EQ.1) IB=1
IF (IBOOT.EQ.2) IB=ICRITX1(2)-ICRITX1(1)+1
IF (IBOOT.GT.2) THEN
  IB=ICRITX1(2)-ICRITX1(1)+1
  DO 272 J=2,IBOOT-1
  IB=IB+(ICRITX1(J+1)-ICRITX1(J))*J
272 CONTINUE
ENDIF
DO 275 I=0,IBOOT-1
III=INDVAR1(IB+I)
HISTB(III)=1
275 CONTINUE
```

```
ITELB=0
DO 278 J=1,IP
ISELB(J)=0
IF (HISTB(J).GT.0) THEN
  ITELB=ITELB+1
  ISELB(ITELB)=J
ENDIF
278 CONTINUE
```

```
C
C SUBROUTINE ERROR IS USED TO CALCULATE THE ACTUAL ERROR RATE OF THE
C LOGISTIC DISCRIMINANT FUNCTION BASED ON THE SELECTED VARIABLES
```

```
C
IW=0
CALL ERROR(AMU,RSIG,XX,IBOOT,HISTB,ACTB,IW)
IF (IW.NE.0) GOTO 14
```

```
C
C SUBROUTINE APPERR IS USED TO CALCULATE THE APPARENT (RESUBSTITUTION)
C ERROR RATE OF THE LOGISTIC DISCRIMINANT FUNCTION BASED ON THE SELECTED
C VARIABLES
```

```
C
CALL APPERR(IBOOT,HISTB,XPOLYF,BETA1,APERR)
ERRDIF=0.0D0
```

```
C
C THE BOOTSTRAP LOOP STARTS HERE.
C THE OPTIMISM OF THE APPARENT ERROR RATE WILL BE ESTIMATED BY MEANS OF
C THE BOOTSTRAP. THIS OPTIMISM WILL THEN BE USED TO ADJUST THE APPARENT
C ERROR RATE (APERR) FOR BIAS.
C
DO 350 IK=1,KB
C
C SUBROUTINE BOOTSAM IS USED TO DRAW A BOOTSTRAP SAMPLE FROM THE TRAINING
C DATA
C
CALL BOOTSAM(XX,XBOOT)
C
C THE LOGISTIC DISCRIMINANT FUNCTION IS CALCULATED ON THE BOOTSTRAP
C SAMPLE
C
IW=0
NITER=100
DSMALL=0.1D0
DO 286 I=1,NN
ICLASSF(I)=0
286 CONTINUE
DO 287 I=NN+1,NNPMM
ICLASSF(I)=1
287 CONTINUE
CALL POLY(IW,ICLASSF,NITER,NNPMM,KLASS,IP,DSMALL,XBOOT,BETA1)
DO 288 I=1,NN
ICLASSF(I)=0
288 CONTINUE
DO 289 I=NN+1,NNPMM
ICLASSF(I)=1
289 CONTINUE
IF (IW.EQ.1) THEN
NOCONV(IS+1)=NOCONV(IS+1)+1
GOTO 14
ENDIF
C
C VARIABLE SELECTION IS PERFORMED ON THE BOOTSTRAP DATA SET
C
DO 295 I=1,NNPMM
SUM1=BETA1(1,1)
DO 294 J=1,IP
SUM1=SUM1+BETA1(J+1,1)*XBOOT(I,J)
294 CONTINUE
ESUM1=DEXP(SUM1)
PI1=ESUM1/(1.0D0+ESUM1)
VF(I)=PI1*(1.0D0-PI1)
IF (DIFNAN(VF(I))) GOTO 14
ZF(I)=SUM1+(1.0D0*ICLASSF(I)-PI1)/VF(I)
IF (DIFNAN(ZF(I))) GOTO 14
295 CONTINUE
DO 305 I=1,NNPMM
DO 300 J=1,IP
XXB(I,J)=XBOOT(I,J)
300 CONTINUE
```

```
XXB(I,IPP1)=ZF(I)
XXB(I,IPP1+1)=VF(I)
305 CONTINUE

IDO=0
NROW=NNPMM
NVAR=IPP1
LDX=NNPMM
IFRQ=0
IWT=IPP1+1
MOPT=0
ICOPT=0
LDCOV=IPP1
LDINCD=1
NOBS=NNPMM

CALL DCORVC(IDO,NROW,NVAR,XXB,LDX,IFRQ,IWT,MOPT,
&ICOPT,XMEAN,COVW,LDCOV,INCD,LDINCD,NOBS,NMISS,SUMWT)
CALL DRBEST(NVAR1,COVW,LDCOV1,NOBS,ICRIT1,NBEST1,NGOOD1,IPRINT1,
&ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)
C
C THE VARIABLES SELECTED FOR THE BOOTSTRAP DATA SET ARE IDENTIFIED
C
RMIN=CRIT1(1)
IBOOT=1
DO 310 I=2,IP
IF (CRIT1(I).LT.RMIN) THEN
RMIN=CRIT1(I)
IBOOT=I
ENDIF
310 CONTINUE
DO 311 J=1,IP
HISTB(J)=0
311 CONTINUE
IF (IBOOT.EQ.1) IB=1
IF (IBOOT.EQ.2) IB=ICRITX1(2)-ICRITX1(1)+1
IF (IBOOT.GT.2) THEN
IB=ICRITX1(2)-ICRITX1(1)+1
DO 312 J=2,IBOOT-1
IB=IB+(ICRITX1(J+1)-ICRITX1(J))*J
312 CONTINUE
ENDIF
DO 315 I=0,IBOOT-1
III=INDVAR1(IB+I)
HISTB(III)=1
315 CONTINUE
C
C THE LOGISTIC CLASSIFICATION FUNCTION BASED ON THE SELECTED
C VARIABLES IS CALCULATED AND USED TO CLASSIFY :
C 1. THE CASES IN THE BOOTSTRAP DATA SET TO OBTAIN THE APPARENT ERROR RATE
C OF THE BOOTSTAP DATA SET, APERRB
C 2. THE CASES IN THE ORIGINAL DATA SET TO OBTAIN THE ERROR RATE APERRV.
C CALCULATE THE DIFFERENCE BETWEEN THESE TWO ERROR RATES (ERRDIF) AND
C ACCUMULATE THESE DIFFERENCES.
C
```

```
CALL APPERR(IBOOT,HISTB,XBOOT,BETA1,APERRB)
CALL APPERRV(IBOOT,HISTB,XX,BETA1,APERRV)
ERRDIF=ERRDIF+(APERRV-APERRB)
350 CONTINUE
C
C THIS IS THE END OF THE BOOTSTRAP-LOOP
C
C CALCULATE THE AVERAGE OF ERRDIF OVER ALL BOOTSTRAP SAMPLES AND ADD THIS
C TO THE APPARENT ERROR RATE FOR THE ORIGINAL DATA SET TO CORRECT FOR BIAS.
C THIS GIVES THE BOOTSTRAP ERROR RATE ESTIMATE, ERRBOOT.
C
ERRDIF=ERRDIF/KB
ERRBOOT=APERR+ERRDIF
EREB=EREB+ERRBOOT
AUMSEB=AUMSEB+((ERRBOOT-ACTB)**2.0D0)

ERACTH=ERACTH+ACTH
EREH=EREH+AMIN
AUMSEH=AUMSEH+((AMIN-ACTH)**2.0D0)

NUM=0
DO 360 J=1,IP
JJ=ISELH(J)
IF (JJ.NE.0) THEN
PSELVARH(JJ)=PSELVARH(JJ)+1.0D0
NUM=NUM+1
ENDIF
360 CONTINUE
PSELNUMH(NUM)=PSELNUMH(NUM)+1.0D0
IF (NUM.EQ.NONZERO) THEN
ISELR=1
DO 361 J=1, NONZERO
IF (HISTH(J).LT.0.1D0) ISELR=0
361 CONTINUE
CPCSH=CPCSH+ISELR
ENDIF
IF (NUM.GT.NONZERO) THEN
ISELR=1
DO 362 J=1, NONZERO
IF (HISTH(J).LT.0.1D0) ISELR=0
362 CONTINUE
IF (ISELR.EQ.1) SELOVERH=SELOVERH+1.0D0
ENDIF
IF (NUM.LT.NONZERO) THEN
ISELW=0
DO 363 J=NONZERO+1,IP
IF (HISTH(J).GT.0.1D0) ISELW=1
363 CONTINUE
IF (ISELW.EQ.0) SELUNDERH=SELUNDERH+1.0D0
ENDIF
ISELM=0
DO 364 J=NONZERO+1,IP
IF (HISTH(J).GT.0.1D0) ISELM=1
364 CONTINUE
IF (ISELM.EQ.1) THEN
```



```
    NCOR=0
    DO 365 J=1, NONZERO
    IF (HISTH(J).GT.0.1D0) NCOR=NCOR+1
365    CONTINUE
    IF ((NCOR.GT.0).AND.(NCOR.LT.NONZERO)) SELMIXH=
&    SELMIXH+1.0D0
    ENDIF
    ERACTB=ERACTB+ACTB
    NUM=0
    DO 380 J=1, IP
    JJ=ISELB(J)
    IF (JJ.NE.0) THEN
        PSELVARB(JJ)=PSELVARB(JJ)+1.0D0
        NUM=NUM+1
    ENDIF
380    CONTINUE
    PSELNUMB(NUM)=PSELNUMB(NUM)+1.0D0
    IF (NUM.EQ.NONZERO) THEN
        ISELR=1
        DO 381 J=1, NONZERO
        IF (HISTB(J).LT.0.1D0) ISELR=0
381    CONTINUE
        CPCSB=CPCSB+ISELR
    ENDIF
    IF (NUM.GT.NONZERO) THEN
        ISELR=1
        DO 382 J=1, NONZERO
        IF (HISTB(J).LT.0.1D0) ISELR=0
382    CONTINUE
        IF (ISELR.EQ.1) SELOVERB=SELOVERB+1.0D0
    ENDIF
    IF (NUM.LT.NONZERO) THEN
        ISELW=0
        DO 383 J=NONZERO+1, IP
        IF (HISTB(J).GT.0.1D0) ISELW=1
383    CONTINUE
        IF (ISELW.EQ.0) SELUNDERB=SELUNDERB+1.0D0
    ENDIF

    ISELM=0
    DO 384 J=NONZERO+1, IP
    IF (HISTB(J).GT.0.1D0) ISELM=1
384    CONTINUE
    IF (ISELM.EQ.1) THEN
        NCOR=0
        DO 385 J=1, NONZERO
        IF (HISTB(J).GT.0.1D0) NCOR=NCOR+1
385    CONTINUE
        IF ((NCOR.GT.0).AND.(NCOR.LT.NONZERO)) SELMIXB=
&    SELMIXB+1.0D0
    ENDIF

    MC=MC+1
    IF (MC.LT.NMC) GOTO 14
```

C

```
C THE MONTE CARLO LOOP STOPS HERE. THE SIMULATION COUNTERS ARE NOW
C DIVIDED BY THE NUMBER OF MC REPETITIONS.
C
400 IF (PSELNUMH(NONZERO).LT.0.5D0) PSELNUMH(NONZERO)=-1.0D0
EREH=EREH/NMC
ERACTH=ERACTH/NMC
BIASH=EREH-ERACTH
AUMSEH=AUMSEH/NMC
CPCSH=CPCSH/PSELNUMH(NONZERO)
PCSH=(CPCSH*PSELNUMH(NONZERO))/NMC
SELOVERH=SELOVERH/NMC
SELUNDERH=SELUNDERH/NMC
SELMIXH=SELMIXH/NMC
DO 410 J=1,IP
PSELNUMH(J)=PSELNUMH(J)/NMC
PSELVARH(J)=PSELVARH(J)/NMC
410 CONTINUE
IF (PSELNUMB(NONZERO).LT.0.5D0) PSELNUMB(NONZERO)=-1.0D0
EREB=EREB/NMC
ERACTB=ERACTB/NMC
BIASB=EREB-ERACTB
AUMSEB=AUMSEB/NMC
CPCSB=CPCSB/PSELNUMB(NONZERO)
PCSB=(CPCSB*PSELNUMB(NONZERO))/NMC
SELOVERB=SELOVERB/NMC
SELUNDERB=SELUNDERB/NMC
SELMIXB=SELMIXB/NMC
DO 420 J=1,IP
PSELNUMB(J)=PSELNUMB(J)/NMC
PSELVARB(J)=PSELVARB(J)/NMC
420 CONTINUE
C
C RESULTS FOR THIS SEPARATION BETWEEN THE TWO GROUPS ARE WRITTEN TO FILE
C
OPEN(1,FILE=FILEOUT,ACCESS='APPEND')
WRITE(1,600) IS,(AMU(2,J),J=1,IP)
WRITE(1,600)
WRITE(1,620) ERACTH
WRITE(1,620) BIASH,AUMSEH
WRITE(1,620) (PSELVARH(J),J=1,IP)
WRITE(1,620) (PSELNUMH(J),J=1,IP)
WRITE(1,620) CPCSH,PCSH,SELOVERH,SELUNDERH,SELMIXH
WRITE(1,*)
WRITE(1,610) ERACTB
WRITE(1,610) BIASB,AUMSEB
WRITE(1,620) (PSELVARB(J),J=1,IP)
WRITE(1,620) (PSELNUMB(J),J=1,IP)
WRITE(1,620) CPCSB,PCSB,SELOVERB,SELUNDERB,SELMIXB
WRITE(1,*)
CLOSE(1)
500 CONTINUE
C
C GO BACK AND REPEAT FOR ANOTHER VALUE OF THE MAHALANOBIS DISTANCE
C BETWEEN THE TWO GROUPS
C
```

```

600 FORMAT(I4,2X,5(F10.5,2X))
610 FORMAT(F12.6,2X,F12.6,2X,F12.6)
620 FORMAT(10(F10.5,2X))
621 FORMAT(12F7.4)
630 FORMAT(10I5)
640 FORMAT(10(F6.2,1X))

```

```

1000 STOP
      END

```

```

SUBROUTINE LOO(II,X,X1)
  IMPLICIT DOUBLE PRECISION (A-H,O-Z)
  PARAMETER (IP=10,NN=50,MM=50,NNPMM=NN+MM,IPP1=IP+1,NMC=200,
&KCLASS=2,NB=500,KB=200,RHO=0.9D0)
  DIMENSION X(NNPMM,IPP1),X1(NNPMM-1,IPP1)
  N=NNPMM
  IF (II.EQ.1) THEN
    DO 5 I=1,N-1
      DO 1 J=1,IPP1
        X1(I,J)=X(I+1,J)
1      CONTINUE
5      CONTINUE
  ENDIF
  IF ((II.GT.1).AND.(II.LT.N)) THEN
    DO 15 I=1,II-1
      DO 10 J=1,IPP1
        X1(I,J)=X(I,J)
10     CONTINUE
15     CONTINUE
      DO 25 I=II,N-1
        DO 20 J=1,IPP1
          X1(I,J)=X(I+1,J)
20     CONTINUE
25     CONTINUE
  ENDIF
  IF (II.EQ.N) THEN
    DO 35 I=1,N-1
      DO 30 J=1,IPP1
        X1(I,J)=X(I,J)
30     CONTINUE
35     CONTINUE
  ENDIF
  RETURN
  END

```

```

SUBROUTINE BOOTSAM(XX,XBOOT)
C THIS SUBROUTINE DRAWS A RANDOM SAMPLE WITH REPLACEMENT FROM THE
C TRAINING DATA. A RANDOM SAMPLE OF SIZE NN IS DRAWN FROM THE
C FIRST GROUP AND A RANDOM SAMPLE OF SIZE MM IS DRAWN FROM THE
C SECOND GROUP.
  IMPLICIT DOUBLE PRECISION (A-H,O-Z)
  PARAMETER (IP=10,NN=50,MM=50,NNPMM=NN+MM,IPP1=IP+1,NMC=200,
&KCLASS=2,NB=500,KB=200,RHO=0.9D0)

```

```

DIMENSION XX(NNPMM,IPP1),XBOOT(NNPMM,IP)
DIMENSION IRN(NN),IRM(MM)
N=NN
CALL RNUND(N,N,IRN)
DO 5 I=1,NN
DO 4 J=1,IP
XBOOT(I,J)=XX(IRN(I),J)
4 CONTINUE
5 CONTINUE
M=MM
CALL RNUND(M,M,IRM)
DO 15 I=1,MM
DO 14 J=1,IP
XBOOT(NN+I,J)=XX(NN+IRM(I),J)
14 CONTINUE
15 CONTINUE
RETURN
END

```

```

SUBROUTINE WF(IDEM,MIN,II,X1,XV,SMLOSS)
C THIS SUBROUTINE CALCULATES THE SMOOTHED LOSS WHEN CASE XV IS
C CLASSIFIED USING THE LOGISTIC REGRESSION FUNCTION CONTAINING THE
C IDEM VARIABLES IDENTIFIED BY THE NONZERO ELEMENTS OF THE VECTOR
C MIN. X1 IS THE DATA MATRIX WITH ROW II DELETED. XV CONTAINS THE
C VALUES OF THE FEATURE VARIABLES FOR THE DELETED CASE.
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=50,MM=50,NNPMM=NN+MM,IPP1=IP+1,NMC=200,
&KCLASS=2,NB=500,KB=200,RHO=0.9D0)
DIMENSION X1(NNPMM-1,IP+1),XX(NNPMM-1,IDEM),XV(IP)
DIMENSION BETA1(IDEM+1,1)
DIMENSION XM1(IDEM),XM2(IDEM),S(IDEM,IDEM),SINV(IDEM,IDEM)
DIMENSION MIN(IP),ICLASS(NNPMM-1)
C
C THE MATRIX XX, CONTAINING ONLY THE IDEM SELECTED VARIABLES, IS FORMED
C
IF (II.LE.NN) THEN
  N1=NN-1
  N2=MM
ENDIF
IF (II.GT.NN) THEN
  N1=NN
  N2=MM-1
ENDIF
DO 10 I=1,N1
ITEL=0
ICLASS(I)=0
DO 5 J=1,IP
IF (MIN(J).EQ.1) THEN
  ITEL=ITEL+1
  XX(I,ITEL)=X1(I,J)
ENDIF
5 CONTINUE
10 CONTINUE
DO 20 I=1,N2

```

```

ITEL=0
ICLASS(N1+I)=1
DO 15 J=1,IP
IF (MIN(J,EQ.1) THEN
  ITEL=ITEL+1
  XX(N1+I,ITEL)=X1(N1+I,J)
ENDIF
15 CONTINUE
20 CONTINUE
NIPN2=N1+N2

IW=0
NITER=100
DSMALL=0.1D0
C
C SUBROUTINE POLY IS USED TO ESTIMATE THE LOGISTIC REGRESSION
C COEFFICIENTS (USING ONLY THE SELECTED VARIABLES AND WITH CASE
C II OMITTED)
C
CALL POLY(IW,ICLASS,NITER,NIPN2,KLASS,ITEL,DSMALL,XX,BETA1)
C
C SUBROUTINE AVGVAR3 IS USED TO CALCULATE THE GROUP MEANS,
C POOLED COVARIANCE MATRIX (AND ITS INVERSE)
C
CALL AVGVAR3(N1,N2,NIPN2,IDEM,XX,XM1,XM2,S,SINV)
C
C THE MAHALANOBIS DISTANCE BETWEEN THE TWO GROUPS (BASED ONLY ON
C THE SELECTED VARIABLES) IS CALCULATED
C
AMAH=0.0D0
DO 50 I=1,IDEM
DO 40 J=1,IDEM
VERS1=XM1(I)-XM2(I)
VERS2=XM1(J)-XM2(J)
AMAH=AMAH+VERS1*SINV(I,J)*VERS2
40 CONTINUE
50 CONTINUE
AMAH=DSQRT(AMAH)
C
C THE CUTOFF POINTS FOR CALCULATION OF THE SMOOTHED LOSS IS DETERMINED
C
CUTOFF1=AMAH/(1.0D0+AMAH)
IF (CUTOFF1.LT.0.5D0) CUTOFF1=0.5D0
CUTOFF2=1.0D0/(1.0D0+AMAH)
IF (CUTOFF2.GT.0.5D0) CUTOFF2=0.5D0

SUM1=BETA1(1,1)
DO 75 J=1,ITEL
SUM1=SUM1+BETA1(J+1,1)*XV(J)
75 CONTINUE
C
C THE POSTERIOR PROBABILITIES FOR CASE XV ARE CALCULATED AND USED TO
C OBTAIN THE SMOOTHED LOSS
C
EE=DEXP(SUM1)

```

```

IF (II.LE.NN) THEN
  PP=1.0D0/(1.0D0+EE)
  SMLOSS=1.0D0-PP
  IF (PP.GT.CUTOFF1) SMLOSS=0.0D0
  IF (PP.LT.CUTOFF2) SMLOSS=1.0D0
ENDIF
IF (II.GT.NN) THEN
  PP=EE/(1.0D0+EE)
  SMLOSS=1.0D0-PP
  IF (PP.GT.CUTOFF1) SMLOSS=0.0D0
  IF (PP.LT.CUTOFF2) SMLOSS=1.0D0
ENDIF
RETURN
END

```

```

SUBROUTINE APPERR(IDEM,RMIN,X1,BETA1,APERR)

```

```

C THIS SUBROUTINE CALCULATES THE APPARENT ERROR RATE OF A LOGISTIC
C DISCRIMINANT RULE BASED ON A SELECTED SUBSET OF VARIABLES.
C INPUT : IDEM=THE NUMBER OF VARIABLES SELECTED
C         RMIN=INDICATOR VECTOR IDENTIFYING THE SELECTED VARIABLES
C         X1=MATRIX CONTAINING THE DATA
C OUTPUT : BETA1=COEFFICIENT OF LOGISTIC CLASSIFICATION FUNCTION
C          BASED ON SELECTED VARIABLES
C          APERR=APPARENT ERROR RATE ASSOCIATED WITH THE LOGISTIC
C          CLASSIFICATION FUNCTION BASED ON THE SELECTED
C          VARIABLES

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)

```

```

PARAMETER (IP=10,NN=50,MM=50,NNPMM=NN+MM,IPP1=IP+1,NMC=200,
&KLASS=2,NB=500,KB=200,RHO=0.9D0)

```

```

DIMENSION X1(NNPMM,IP),XX(NNPMM,IDEM)

```

```

DIMENSION RMIN(IP)

```

```

DIMENSION ICLASS(NNPMM)

```

```

DIMENSION BETA1(IPP1,1)

```

```

N1=NN

```

```

N2=MM

```

```

DO 10 I=1,N1

```

```

  ITEL=0

```

```

  ICLASS(I)=0

```

```

  DO 5 J=1,IP

```

```

    IF (RMIN(J).EQ.1) THEN

```

```

      ITEL=ITEL+1

```

```

      XX(I,ITEL)=X1(I,J)

```

```

    ENDIF

```

```

5  CONTINUE

```

```

10 CONTINUE

```

```

  DO 20 I=1,N2

```

```

    ITEL=0

```

```

    ICLASS(N1+I)=1

```

```

    DO 15 J=1,IP

```

```

      IF (RMIN(J).EQ.1) THEN

```

```

        ITEL=ITEL+1

```

```

        XX(N1+I,ITEL)=X1(N1+I,J)

```

```

      ENDIF

```

```

15 CONTINUE

```

```

20 CONTINUE
  N1PN2=N1+N2
  IW=0
  NITER=100
  DSMALL=0.1D0
  CALL POLY(IW,ICLASS,NITER,N1PN2,KLASS,ITEL,DSMALL,XX,BETA1)
  APERR=0.0D0
  DO 90 I=1,NNPMM
    SUM1=BETA1(1,1)
    DO 75 J=1,ITEL
      SUM1=SUM1+BETA1(J+1,1)*XX(I,J)
75 CONTINUE
  IF ((I.LE.NN).AND.(SUM1.GE.0.0D0)) APERR=APERR+1.0D0
  IF ((I.GT.NN).AND.(SUM1.LT.0.0D0)) APERR=APERR+1.0D0
90 CONTINUE
  APERR=APERR/NNPMM
600 FORMAT(10I5)
610 FORMAT(10(F8.4,2X))
  RETURN
  END

```

SUBROUTINE APPERRV(IDEM,RMIN,X1,BETA1,APERRV)

```

C THIS SUBROUTINE CALCULATES THE ERROR RATE WHEN CLASSIFYING THE
C DATA IN X1 USING THE LOGISTIC REGRESSION FUNCTION WITH COEFFICIENTS
C IN BETA1 (WHICH IS INPUT)
C INPUT : IDEM=THE NUMBER OF VARIABLES SELECTED
C         RMIN=INDICATOR VECTOR IDENTIFYING THE SELECTED VARIABLES
C         X1=MATRIX CONTAINING THE DATA
C         BETA1=COEFFICIENT OF LOGISTIC CLASSIFICATION FUNCTION
C         CALCULATED ON ANOTHER DATA SET.
C OUTPUT : APERRV=ERROR RATE OBTAINED WHEN CLASSIFYING THE DATA IN X1
C          USING THE LOGISTIC CLASSIFICATION FUNCTION WITH COEFFICIENTS
C          IN BETA1
C

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)

PARAMETER (IP=10,NN=50,MM=50,NNPMM=NN+MM,IPP1=IP+1,NMC=200,
&KLASS=2,NB=500,KB=200,RHO=0.9D0)

DIMENSION X1(NNPMM,IP+1),XX(NNPMM,IDEM)

DIMENSION RMIN(IP)

DIMENSION BETA1(IPP1,1)

N1=NN

N2=MM

DO 10 I=1,N1

ITEL=0

DO 5 J=1,IP

IF (RMIN(J).EQ.1) THEN

ITEL=ITEL+1

XX(I,ITEL)=X1(I,J)

ENDIF

5 CONTINUE

10 CONTINUE

DO 20 I=1,N2

ITEL=0

DO 15 J=1,IP

```

IF (RMIN(J).EQ.1) THEN
  ITEL=ITEL+1
  XX(N1+I,ITEL)=X1(N1+I,J)
ENDIF
15 CONTINUE
20 CONTINUE
  N1PN2=N1+N2
  APERRV=0.0D0
  DO 90 I=1,NNPMM
    SUM1=BETA1(1,1)
    DO 75 J=1,ITEL
      SUM1=SUM1+BETA1(J+1,1)*XX(I,J)
75 CONTINUE
  IF ((I.LE.NN).AND.(SUM1.GE.0.0D0)) APERRV=APERRV+1.0D0
  IF ((I.GT.NN).AND.(SUM1.LT.0.0D0)) APERRV=APERRV+1.0D0
90 CONTINUE
  APERRV=APERRV/NNPMM
  RETURN
END

```

SUBROUTINE ERROR(AMU,RSIG,XX,IMIN,HIST,ACTERR,IW)

```

C
C THIS SUBROUTINE USES SIMULATION TO ESTIMATE THE POST-SELECTION
C ACTUAL ERROR RATE
C

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=50,MM=50,NNPMM=NN+MM,IPP1=IP+1,NMC=200,
&KCLASS=2,NB=500,KB=200,RHO=0.9D0)
DIMENSION XX(NNPMM,IP+1),XKIES(NNPMM,IMIN)
DIMENSION RNX1(NB,IP),RNX2(NB,IP)
DIMENSION RNX1K(NB,IP),RNX2K(NB,IP)
DIMENSION AMU(2,IP),RSIG(IP,IP)
DIMENSION XB(IP),HIST(IP)
DIMENSION ICLASSF(NNPMM)
DIMENSION BETA1F(IPP1,1)
ITEL=0
DO 1 I=1,NN
  ICLASSF(I)=0
1 CONTINUE
DO 2 I=1,MM
  ICLASSF(NN+I)=1
2 CONTINUE
DO 5 L=1,IP
  IF (HIST(L).EQ.1) THEN
    ITEL=ITEL+1
    DO 3 I=1,NNPMM
      XKIES(I,ITEL)=XX(I,L)
3 CONTINUE
  ENDIF
5 CONTINUE
  IWF=0
  NITERF=100
  DSMALLF=0.1D0
  CALL POLY(IWF,ICLASSF,NITERF,NNPMM,KCLASS,IMIN,DSMALLF,

```



```

&XKIES,BETA1F)
IF (IWF.NE.0) THEN
  IW=1
  RETURN
ENDIF
CALL DRNMVN(NB,IP,RSIG,IP,RNX1,NB)
CALL DRNMVN(NB,IP,RSIG,IP,RNX2,NB)
ACT1=0.0D0
ACT2=0.0D0
ITEL=0
DO 40 L=1,IP
IF (HIST(L).EQ.1) THEN
  ITEL=ITEL+1
  DO 31 I=1,NB
    RNX1K(I,ITEL)=RNX1(I,L)+AMU(1,L)
31  CONTINUE
  DO 32 I=1,NB
    RNX2K(I,ITEL)=RNX2(I,L)+AMU(2,L)
32  CONTINUE
  ENDF
40  CONTINUE
  DO 99 II=1,NB
  DO 60 JJ=1,IMIN
  XB(JJ)=RNX1K(II,JJ)
60  CONTINUE
  SUM1=BETA1F(1,1)
  DO 75 J=1,IMIN
  SUM1=SUM1+BETA1F(J+1,1)*XB(J)
75  CONTINUE
  IF (SUM1.GE.0.0D0) ACT1=ACT1+1.0D0
99  CONTINUE
  DO 199 II=1,NB
  DO 160 JJ=1,IMIN
  XB(JJ)=RNX2K(II,JJ)
160 CONTINUE
  SUM1=BETA1F(1,1)
  DO 175 J=1,IMIN
  SUM1=SUM1+BETA1F(J+1,1)*XB(J)
175 CONTINUE
  IF (SUM1.LT.0.0D0) ACT2=ACT2+1.0D0
199 CONTINUE
  ACTERR=(ACT1+ACT2)/(2.0D0*NB)
  RETURN
END

```

SUBROUTINE AVGVAR3(N,M,NPM,IT,XX,S,SINV,XM1,XM2)

```

C
C THIS SUBROUTINE CALCULATES THE MEAN VECTORS OF THE TWO GROUPS (XM1
C AND XM2) AS WELL AS THE POOLED COVARIANCE MATRIX (S) AND ITS
C INVERSE (SINV). THIS ROUTINE IS USED FOR THE MATRIX CONTAINING
C ONLY A SUBSET OF THE ORIGINAL NUMBER OF ROWS.
C INPUT: THE MATRIX XX(NPM,IPP1) - THE FIRST N ROWS OF XX CONTAIN THE
C OBSERVATIONS FROM GROUP1 AND THE NEXT M ROWS CONTAIN THE
C OBSERVATIONS FROM GROUP2. ONLY THE FIRST IT COLUMNS ARE

```

```
C      TAKEN INTO ACCOUNT
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
      DIMENSION XX(NNPMM,IP+1),XX1(N,IP),XX2(M,IP)
      DIMENSION XM1(IP),XM2(IP)
      DIMENSION S(IP,IP),SINV(IP,IP),S1(IP,IP),S2(IP,IP)
      EXTERNAL DCORVC,DLINDS
      DO 10 I=1,N
      DO 5 J=1,IT
      XX1(I,J)=XX(I,J)
5      CONTINUE
10     CONTINUE
      DO 20 I=1,M
      DO 15 J=1,IT
      XX2(I,J)=XX(N+I,J)
15     CONTINUE
20     CONTINUE
      IDO=0
      NVAR=IT
      IFRQ=0
      IWT=0
      MOPT=0
      ICOPT=0
      LDCOV=IP
      LDINCD=1
      NROW=N
      LDX=N
      CALL DCORVC(IDO,NROW,NVAR,XX1,LDX,IFRQ,IWT,MOPT,
&      ICOPT,XM1,S1,LDCOV,INCD,LDINCD,NOBS,
&      NMISS,SUMWT)
      NROW=M
      LDX=M
      CALL DCORVC(IDO,NROW,NVAR,XX2,LDX,IFRQ,IWT,MOPT,
&      ICOPT,XM2,S2,LDCOV,INCD,LDINCD,NOBS,
&      NMISS,SUMWT)
      NPM2=NPM-2
      DO 30 I=1,IT
      DO 25 J=1,IT
      S(I,J)=((N-1)*S1(I,J)+(M-1)*S2(I,J))/NPM2
25     CONTINUE
30     CONTINUE
      CALL DLINDS(IT,S,IP,SINV,IP)
      RETURN
      END
```

PROGRAM 4

```

C IN THIS PROGRAM, THE PTq (PRE-TEST q) METHOD IS USED TO SELECT
C VARIABLES. THE POST-SELECTION ERROR RATE IS ESTIMATED BY MEANS
C OF A LEAVE-ONE-OUT STRATEGY, WHERE THE LEAVE-ONE-OUT PROCESS
C PRECEDES THE SELECTION PROCESS.
C THE PROPERTIES OF THIS PROCEDURE ARE INVESTIGATED BY MEANS OF
C SIMULATION. IT IS ASSUMED THAT THE FEATURE VARIABLES ARE
C UNCORRELATED AND NORMALLY DISTRIBUTED.
C
C PARAMETERS :
C IP=THE TOTAL NUMBER OF AVAILABLE FEATURE VARIABLES
C NN=THE TRAINING SAMPLE SIZE FROM GROUP 1
C MM=THE TRAINING SAMPLE SIZE FROM GROUP 2
C NMC=THE NUMBER OF MONTE CARLO REPETITIONS
C
C THE FOLLOWING IMSL-SUBROUTINES ARE USED IN THE MAIN PROGRAM:
C 1. DLINDS: FINDS THE INVERSE OF A GIVEN COVARIANCE MATRIX
C 2. DCHFAC: FINDS THE CHOLESKY DECOMPOSITION OF A GIVEN MATRIX
C 3. DRNMVN: GENERATES VALUES FROM A MULTIVARIATE NORMAL DISTRIBUTION
C 4. DSVRGP: SORTS A REAL ARRAY BY ALGEBRAICALLY INCREASING VALUE
C 5. DCORVC: COMPUTES A COVARIANCE OR CORRELATION MATRIX
C
C IMPLICIT DOUBLE PRECISION (A-H,O-Z)
C PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1,NMC=5000)
C DIMENSION AMU(2,IP),SIGMAM(IP,IP),RSIG(IP,IP)
C DIMENSION RNX1(NN,IP),RNX2(MM,IP),RESP(NNPMM)
C DIMENSION XX(NNPMM,IPP1),XX1P(NNPMM-1,IPP1),XX2P(NNPMM,IPP1)
C DIMENSION THSEL(2,IPP1),SSEL(IP,IP),SINV(IP,IP)
C DIMENSION PSELVARP(IP),PSELNUMP(IP),ERRP(NNPMM)
C DIMENSION XV(IP),TV(IP),ATV(IP),Z(IP),AVG(2,IP),CRIT(0:IP)
C DIMENSION IPERM(IP),INDH(IP)
C
C CHARACTER*70 FILEOUT
C FILEOUT='/ptq.d'
C
C NONZERO IS THE NUMBER OF NONZERO ELEMENTS OF THE MEAN VECTOR OF THE
C SECOND GROUP - ALL THE ELEMENTS OF THE MEAN VECTOR OF THE FIRST
C GROUP ARE TAKEN EQUAL TO ZERO
C
C NONZERO=1
C DO 2 I=1,IP
C   SIGMAM(I,I)=1.0D0
C DO 1 J=1,IP
C   SIGMAM(I,J)=0.0D0
1 CONTINUE
2 CONTINUE
C DO 11 I=1,IP
C   AMU(1,I)=0.0D0
11 CONTINUE
C
C THE LOOP UP TO 500 SYSTEMATICALLY INCREASES THE DISTANCE BETWEEN
C THE TWO GROUPS.

```

C THE FOLLOWING SIMULATION COUNTERS ARE INITIALISED:
 C 1. PSELVARP(J): THE ESTIMATED PROBABILITY OF SELECTING VARIABLE J
 C 2. PSELNUMP(J): THE ESTIMATED PROBABILITY OF SELECTING J VARIABLES
 C 3. EREP: THE AVERAGE ERROR RATE ESTIMATOR OF THE PTQ METHOD
 C 4. AMSEOPP: THE MEAN SQUARED ERROR OF ESTIMATING THE OPTIMAL ERROR RATE
 C 5. AUMSEP: THE MEAN SQUARED ERROR OF ESTIMATING THE ACTUAL ERROR RATE
 C 6. ERACTP: THE AVERAGE POST-SELECTION ACTUAL ERROR RATE
 C 7. EROPTP: THE AVERAGE POST-SELECTION OPTIMAL ERROR RATE
 C

```
DO 500 IS=0,6
IF (IS.LE.4) D2=1.0D0*IS
IF (IS.EQ.5) D2=6.0D0
IF (IS.EQ.6) D2=9.0D0
DO 12 J=1, NONZERO
AMU(2,J)=DSQRT(D2/(1.0D0*NONZERO))
PSELVARP(J)=0.0D0
PSELNUMP(J)=0.0D0
```

```
12 CONTINUE
IF (NONZERO.LT.IP) THEN
DO 13 J=NONZERO+1,IP
AMU(2,J)=0.0D0
PSELVARP(J)=0.0D0
PSELNUMP(J)=0.0D0
```

```
13 CONTINUE
```

```
ENDIF
EREP=0.0D0
AMSEOPP=0.0D0
AUMSEP=0.0D0
ERACTP=0.0D0
EROPTP=0.0D0
CPCSP=0.0D0
SELOVERP=0.0D0
SELUNDERP=0.0D0
SELMIXP=0.0D0
```

```
TOL=1.0D2*DMACH(4)
CALL DCHFAC(IP, SIGMAM, IP, TOL, IRANK, RSIG, IP)
```

C
 C THE SIMULATION LOOP BEGINS, AND THE FIRST STEP IS TO GENERATE THE
 C REQUIRED TRAINING DATA SETS FROM THE RELEVANT MULTIVARIATE NORMAL
 C DISTRIBUTIONS - NOTE THAT THE MEAN VALUES ARE ADDED SEPARATELY
 C

```
MC=0
```

```
14 CALL DRNMVN(NN, IP, RSIG, IP, RNX1, NN)
CALL DRNMVN(MM, IP, RSIG, IP, RNX2, MM)
```

```
DO 16 I=1, NN
DO 15 J=1, IP
RNX1(I,J)=RNX1(I,J)+AMU(1,J)
```

```
15 CONTINUE
```

```
16 CONTINUE
```

```
DO 20 I=1, MM
DO 19 J=1, IP
RNX2(I,J)=RNX2(I,J)+AMU(2,J)
```

```
19 CONTINUE
```

```
20 CONTINUE
```

```

DO 25 I=1,NN
RESP(I)=1.0D0
25 CONTINUE
DO 30 I=NN+1,NNPMM
RESP(I)=2.0D0
30 CONTINUE
C
C A SINGLE DATA MATRIX XX (NNPMM x IP+1) IS FORMED. THE FIRST IP COLUMNS
C CONTAIN THE FEATURE VARIABLES, WHILE COLUMN (IP+1) CONTAINS THE
C RESPONSE VARIABLE VALUES INDICATING GROUP MEMBERSHIP.
C
DO 45 J=1,IP
DO 35 I=1,NN
XX(I,J)=RNX1(I,J)
35 CONTINUE
DO 40 I=1,MM
XX(NN+I,J)=RNX2(I,J)
40 CONTINUE
45 CONTINUE
DO 50 I=1,NN
XX(I,IP+1)=RESP(I)
50 CONTINUE
DO 55 I=1,MM
XX(NN+I,IP+1)=RESP(NN+I)
55 CONTINUE
C
C THIS IS THE BEGINNING OF THE LOOP WHERE THE ROWS OF THE ORIGINAL
C DATA MATRIX ARE OMITTED ONE BY ONE.
C SELECTION BY MEANS OF THE PTq-METHOD IS THEN DONE ON THE REMAINING
C DATA, AND THE OMITTED CASE IS THEN CLASSIFIED USING THE LINEAR
C DISCRIMINANT FUNCTION BASED ON THE SELECTED VARIABLES.
C
DO 120 II=1,NNPMM
IF (II.LE.NN) THEN
NNEW=NN-1
MNEW=MM
ENDIF
IF (II.GT.NN) THEN
NNEW=NN
MNEW=MM-1
ENDIF
CALL LOO(II,XX,XX1P)
C
C THE PTq METHOD STARTS HERE - IT IS APPLIED TO THE DATA MATRIX
C WITH ROW NUMBER II OMITTED (XX1P)
C
VERM=DSQRT(1.0D0*NNEW*MNEW/(NNEW+MNEW))
DO 70 J=1,IP
AVG(1,J)=0.0D0
DO 65 I=1,NNEW
AVG(1,J)=AVG(1,J)+XX1P(I,J)
65 CONTINUE
AVG(2,J)=0.0D0
DO 66 I=1,MNEW
AVG(2,J)=AVG(2,J)+XX1P(NNEW+I,J)

```

```

66 CONTINUE
   AVG(1,J)=AVG(1,J)/NNEW
   AVG(2,J)=AVG(2,J)/MNEW
   TV(J)=VERM*(AVG(1,J)+AVG(2,J))
   ATV(J)=DABS(TV(J))
   IPERM(J)=J
70 CONTINUE
   SUM=0.0D0
   DO 80 J=1,IP
   DO 75 I=1,NNEW
   SUM=SUM+(XX1P(I,J)-AVG(1,J))**2.0D0
75 CONTINUE
   DO 76 I=1,MNEW
   SUM=SUM+(XX1P(NNEW+I,J)-AVG(2,J))**2.0D0
76 CONTINUE
80 CONTINUE
   SHAT2=SUM/(IP*(NNEW+MNEW-2.0D0))
   SHAT=DSQRT(SHAT2)
   CALL DSVRGP(IP,ATV,Z,IPERM)
   CRIT(0)=IP*SHAT2
   DO 100 IQ=1,IP-1
   CRIT(IQ)=IP*SHAT2
   SUM=0.0D0
   DO 85 I=1,IQ
   SUM=SUM+Z(I)*Z(I)-2.0D0*SHAT2+2.0D0*SHAT*Z(IQ+1)*
& (PHI((Z(IQ+1)-Z(I))/SHAT)+PHI((Z(IQ+1)+Z(I))/SHAT))
85 CONTINUE
   CRIT(IQ)=CRIT(IQ)+SUM
   SUM=0.0D0
   DO 90 I=IQ+1,IP
   SUM=SUM+2.0D0*SHAT*Z(IQ)*
& (PHI((Z(IQ)-Z(I))/SHAT)+PHI((Z(IQ)+Z(I))/SHAT))
90 CONTINUE
   CRIT(IQ)=CRIT(IQ)+SUM
   IF (CRIT(IQ).LT.0.0D0) CRIT(IQ)=0.0D0
100 CONTINUE

   AMIN=CRIT(0)
   IQHAT=0
   DO 110 J=1,IP-1
   IF (CRIT(J).LT.AMIN) THEN
     AMIN=CRIT(J)
     IQHAT=J
   ENDIF
110 CONTINUE

   NROW=NNPMM-1
   NVAR=IP-IQHAT
   DO 111 J=1,IP
   INDH(J)=0
111 CONTINUE
   DO 115 J=1,NVAR
   INDH(IPERM(IQHAT+J))=1
115 CONTINUE

```

```
IC=0
DO 116 J=1,IP
IF (INDH(J).EQ.1) THEN
  IC=IC+1
  XV(IC)=XX(IL,J)
ENDIF
116 CONTINUE
C
C SUBROUTINE WF IS USED TO CLASSIFY THE OMITTED CASE USING ONLY THE
C VARIABLES SELECTED BY THE PTq METHOD
C
CALL WF(INDH,NNEW,MNEW,XX1P,XV,WW)
IF (WW.GT.0) GRP=1
IF (WW.LE.0) GRP=2
C
C THE 0-1 LOSS ASSOCIATED WITH CLASSIFICATION OF THE OMITTED
C CASE IS RECORDED
C
IF(IL.LE.NN) ERRP(IL)=DABS(1-GRP)
IF(IL.GT.NN) ERRP(IL)=DABS(2-GRP)

120 CONTINUE
C
C THIS IS THE END OF THE LOOP WHERE THE ROWS ARE OMITTED ONE BY ONE.
C
C THE ERROR RATE ESTIMATE IS NOW CALCULATED.
C
ERRORP=0.0D0
DO 125 I=1,NNPMM
ERRORP=ERRORP+ERRP(I)
125 CONTINUE
ERRORP=ERRORP/NNPMM
C
C ERRORP IS THE ERROR RATE ESTIMATE FOR THE PTq METHOD.
C
C THE PTq SELECTION METHOD IS NOW APPLIED TO THE FULL DATA SET (XX)
C TO SELECT THE FINAL MODEL
C
VERM=DSQRT(1.0D0*NN*MM/(NN+MM))
DO 170 J=1,IP
AVG(1,J)=0.0D0
DO 165 I=1,NN
AVG(1,J)=AVG(1,J)+XX(I,J)
165 CONTINUE
AVG(2,J)=0.0D0
DO 166 I=1,MM
AVG(2,J)=AVG(2,J)+XX(NN+I,J)
166 CONTINUE
AVG(1,J)=AVG(1,J)/NN
AVG(2,J)=AVG(2,J)/MM
TV(J)=VERM*(AVG(1,J)+AVG(2,J))
ATV(J)=DABS(TV(J))
IPERM(J)=J
170 CONTINUE
SUM=0.0D0
```

```

DO 180 J=1,IP
DO 175 I=1,NN
SUM=SUM+(XX(I,J)-AVG(1,J))**2.0D0
175 CONTINUE
DO 176 I=1,MM
SUM=SUM+(XX(NN+I,J)-AVG(2,J))**2.0D0
176 CONTINUE
180 CONTINUE
SHAT2=SUM/(IP*(NN+MM-2.0D0))
SHAT=DSQRT(SHAT2)
CALL DSVRGP(IP,ATV,Z,IPERM)
CRIT(0)=IP*SHAT2
DO 200 IQ=1,IP-1
CRIT(IQ)=IP*SHAT2
SUM=0.0D0
DO 185 I=1,IQ
SUM=SUM+Z(I)*Z(I)-2.0D0*SHAT2+2.0D0*SHAT*Z(IQ+1)*
& (PHI((Z(IQ+1)-Z(I))/SHAT)+PHI((Z(IQ+1)+Z(I))/SHAT))
185 CONTINUE
CRIT(IQ)=CRIT(IQ)+SUM
SUM=0.0D0
DO 190 I=IQ+1,IP
SUM=SUM+2.0D0*SHAT*Z(IQ)*
& (PHI((Z(IQ)-Z(I))/SHAT)+PHI((Z(IQ)+Z(I))/SHAT))
190 CONTINUE
CRIT(IQ)=CRIT(IQ)+SUM
IF (CRIT(IQ).LT.0.0D0) CRIT(IQ)=0.0D0
200 CONTINUE
AMIN=CRIT(0)
IQHAT=0
DO 210 J=1,IP-1
IF (CRIT(J).LT.AMIN) THEN
AMIN=CRIT(J)
IQHAT=J
ENDIF
210 CONTINUE

IMIN=IP-IQHAT
DO 211 J=1,IP
INDH(J)=0
211 CONTINUE
DO 215 J=1,IMIN
INDH(IPERM(IQHAT+J))=1
215 CONTINUE
DO 245 J=1,IMIN
DO 241 I=1,NNPMM
XX2P(I,J)=XX(I,IPERM(IQHAT+J))
241 CONTINUE
THSEL(1,J)=AMU(1,IPERM(IQHAT+J))
THSEL(2,J)=AMU(2,IPERM(IQHAT+J))
DO 242 I=1,IMIN
SSEL(I,J)=SIGMAM(IPERM(IQHAT+I),IPERM(IQHAT+J))
242 CONTINUE
245 CONTINUE
CALL DLINDS(IMIN,SSEL,IP,SINV,IP)

```



```

DELTA2=0.0D0
DO 248 I1=1,IMIN
DO 247 I2=1,IMIN
V1=THSEL(1,I1)-THSEL(2,I1)
V2=THSEL(1,I2)-THSEL(2,I2)
DELTA2=DELTA2+V1*SINV(I1,I2)*V2
247 CONTINUE
248 CONTINUE
C
C THE POST-SELECTION OPTIMAL ERROR RATE IS CALCULATED
C
OPT=DNORDF(-0.5D0*DSQRT(DELTA2))
EROPTP=EROPTP+OPT
C
C SUBROUTINE ERACTP IS USED TO CALCULATE THE POST-SELECTION ACTUAL
C ERROR RATE
C
CALL ERACTP(IMIN,THSEL,SSEL,XX2P,ACT)
ERACTP=ERACTP+ACT
EREP=EREP+ERRORP
AMSEOPP=AMSEOPP+((ERRORP-OPT)**2.0D0)
AUMSEP=AUMSEP+((ERRORP-ACT)**2.0D0)
DO 250 J=1,IMIN
JJ=IPERM(IQHAT+J)
PSELVARP(JJ)=PSELVARP(JJ)+1.0D0
250 CONTINUE
PSELNUMP(IMIN)=PSELNUMP(IMIN)+1.0D0

IF (IMIN.EQ.NONZERO) THEN
ISELR=1
DO 251 J=1,NONZERO
IF (INDH(J).LT.0.1D0) ISELR=0
251 CONTINUE
CPCSP=CPCSP+ISELR
ENDIF
IF (IMIN.GT.NONZERO) THEN
ISELR=1
DO 252 J=1,NONZERO
IF (INDH(J).LT.0.1D0) ISELR=0
252 CONTINUE
IF (ISELR.EQ.1) SELOVERP=SELOVERP+1.0D0
ENDIF
IF (IMIN.LT.NONZERO) THEN
ISELW=0
DO 253 J=NONZERO+1,IP
IF (INDH(J).GT.0.1D0) ISELW=1
253 CONTINUE
IF (ISELW.EQ.0) SELUNDERP=SELUNDERP+1.0D0
ENDIF

ISELM=0
DO 254 J=NONZERO+1,IP
IF (INDH(J).GT.0.1D0) ISELM=1
254 CONTINUE
IF (ISELM.EQ.1) THEN

```

```

    NREG=0
    DO 255 J=1, NONZERO
    IF (INDH(J).GT.0.1D0) NREG=NREG+1
255  CONTINUE
    IF ((NREG.GT.0).AND.(NREG.LT.NONZERO)) SELMIXP=
&    SELMIXP+1.0D0
    ENDF
C
C  THE PTq PROCEDURE ENDS HERE
C
    MC=MC+1
    IF (MC.LT.NMC) GOTO 14
C
C  THE MONTE CARLO LOOP STOPS HERE. THE SIMULATION COUNTERS ARE NOW
C  DIVIDED BY THE NUMBER OF MC REPETITIONS.
C
400  IF (PSELNUMP(NONZERO).LT.0.5D0) PSELNUMP(NONZERO)=-1.0D0
    EREP=EREP/NMC
    ERACTP=ERACTP/NMC
    EROPTP=EROPTP/NMC
    BIASP1=EREP-EROPTP
    BIASP2=EREP-ERACTP
    AMSEP1=AMSEOPP/NMC
    AMSEP2=AUMSEP/NMC
    CPCSP=CPCSP/PSELNUMP(NONZERO)
    PCSP=(CPCSP*PSELNUMP(NONZERO))/NMC
    SELOVERP=SELOVERP/NMC
    SELUNDERP=SELUNDERP/NMC
    SELMIXP=SELMIXP/NMC
    DO 410 J=1, IP
    PSELNUMP(J)=PSELNUMP(J)/NMC
    PSELVARP(J)=PSELVARP(J)/NMC
410  CONTINUE
C
C  RESULTS FOR THIS SEPARATION BETWEEN THE TWO GROUPS ARE WRITTEN TO FILE
C
    OPEN(1, FILE=FILEOUT, ACCESS='APPEND')
    WRITE(1,600) IS,(AMU(2,J),J=1,IP)
    WRITE(1,600)
    WRITE(1,610) EROPTP,ERACTP
    WRITE(1,610) BIASP1,AMSEP1
    WRITE(1,610) BIASP2,AMSEP2
    WRITE(1,620) (PSELVARP(J),J=1,IP)
    WRITE(1,620) (PSELNUMP(J),J=1,IP)
    WRITE(1,620) CPCSP,PCSP,SELOVERP,SELUNDERP,SELMIXP
    WRITE(1,600)
    WRITE(1,*)
    CLOSE(1)
500  CONTINUE
C
C  GO BACK AND REPEAT FOR ANOTHER VALUE OF THE MAHALANOBIS DISTANCE
C  BETWEEN THE TWO GROUPS
C
600  FORMAT(I4,2X,5(F10.5,2X))
610  FORMAT(F12.6,2X,F12.6,2X,F12.6)

```

620 FORMAT(10(F10.5,2X))

1000 STOP
END

SUBROUTINE ERACTP(IT,AMU,SIGMAM,XX,ACT)

C
C THIS SUBROUTINE CALCULATES THE ACTUAL ERROR RATE OF THE LDF
C BASED ON A SELECTED SUBSET OF VARIABLES

C
C INPUT : IT=THE NUMBER OF COLUMNS TO BE TAKEN INTO ACCOUNT.
C AMU=THE MATRIX CONTAINING THE MEANS.
C SIGMAM=THE COVARIANCE MATRIX.
C XX=THE DATA MATRIX.
C OUTPUT : ACT=THE ACTUAL ERROR RATE.

C
C IMPLICIT DOUBLE PRECISION (A-H,O-Z)
C PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
C DIMENSION XX(NNPMM,IP+1),S(IP,IP),SINV(IP,IP),XM1(IP),XM2(IP)
C DIMENSION AMU(2,IP),SIGMAM(IP,IP),SXM12(IP)

C
C SUBROUTINE AVGVARV IS USED TO CALCULATE THE GROUP MEANS (XM1 AND XM2
C AS WELL AS THE POOLED COVARIANCE MATRIX (AND ITS INVERSE).

C
CALL AVGVARV(IT,XX,S,SINV,XM1,XM2)

SUM1=0.0D0

SUM2=0.0D0

DO 10 I1=1,IT

SXM12(I1)=0.0D0

DO 5 I2=1,IT

V1=AMU(1,I1)-(XM1(I1)+XM2(I1))/2.0D0

V2=XM1(I2)-XM2(I2)

SUM1=SUM1+V1*SINV(I1,I2)*V2

SXM12(I1)=SXM12(I1)+SINV(I1,I2)*V2

V1=AMU(2,I1)-(XM1(I1)+XM2(I1))/2.0D0

SUM2=SUM2+V1*SINV(I1,I2)*V2

5 CONTINUE

10 CONTINUE

DSMU1=SUM1

DSMU2=SUM2

V=0.0D0

DO 20 I1=1,IT

DO 15 I2=1,IT

V=V+SXM12(I1)*SIGMAM(I1,I2)*SXM12(I2)

15 CONTINUE

20 CONTINUE

P1=DNORDF(-DSMU1/DSQRT(V))

P2=DNORDF(DSMU2/DSQRT(V))

ACT=0.5D0*(P1+P2)

RETURN

END

SUBROUTINE AVGVARV(IC,XX,S,SINV,XM1,XM2)

```

C
C THIS SUBROUTINE CALCULATES THE GROUP MEANS (XM1 AND XM2) AND THE
C POOLED COVARIANCE MATRIX (S) AS WELL AS ITS INVERSE (SINV).
C THIS ROUTINE IS FOR THE MATRIX CONTAINING ALL THE ROWS.
C INPUT : XX(NNPMM,IP) = THE FIRST NN ROWS OF XX CONTAIN THE OBSERVATIONS
C FOR GROUP 1 AND THE NEXT MM ROWS CONTAIN THE
C OBSERVATIONS FOR GROUP 2. ONLY THE FIRST IC
C COLUMNS ARE TAKEN INTO ACCOUNT.
C IC=THE NUMBER OF COLUMNS TO BE TAKEN INTO ACCOUNT.
C OUTPUT : XM1=MEAN OF GROUP 1.
C XM2=MEAN OF GROUP 2.
C S=POOLED COVARIANCE MATRIX.
C SINV=INVERSE OF POOLED COVARIANCE MATRIX.
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
DIMENSION XX(NNPMM,IP+1),XX1(NN,IP),XX2(MM,IP)
DIMENSION XM1(IP),XM2(IP)
DIMENSION S(IP,IP),SINV(IP,IP),S1(IP,IP),S2(IP,IP)
EXTERNAL DCORVC,DLINDS
DO 10 I=1,NN
DO 5 J=1,IC
XX1(I,J)=XX(I,J)
5 CONTINUE
10 CONTINUE
DO 20 I=1,MM
DO 15 J=1,IC
XX2(I,J)=XX(NN+I,J)
15 CONTINUE
20 CONTINUE
IDO=0
NVAR=IC
IFRQ=0
IWT=0
MOPT=0
ICOPT=0
LDCOV=IP
LDINCD=1
NROW=NN
LDX=NN
CALL DCORVC(IDO,NROW,NVAR,XX1,LDX,IFRQ,IWT,MOPT,
& ICOPT,XM1,S1,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)
NROW=MM
LDX=MM
CALL DCORVC(IDO,NROW,NVAR,XX2,LDX,IFRQ,IWT,MOPT,
& ICOPT,XM2,S2,LDCOV,INCD,LDINCD,NOBS,
& NMISS,SUMWT)
NNPMMM2=NNPMM-2
DO 30 I=1,IC
DO 25 J=1,IC
S(I,J)=((NN-1)*S1(I,J)+(MM-1)*S2(I,J))/NNPMMM2
25 CONTINUE
30 CONTINUE
CALL DLINDS(IC,S,IP,SINV,IP)
RETURN

```

END

SUBROUTINE LOO(II,X,X1)

```

C
C THIS SUBROUTINE OMITTS ONE ROW FROM THE DATA MATRIX.
C INPUT : II=THE NUMBER OF THE ROW TO BE OMITTED.
C X=THE MATRIX CONTAINING ALL THE ROWS.
C OUTPUT : X1=THE MATRIX WITH ROW II OMITTED.
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM,IPP1=IP+1)
DIMENSION X(NNPMM,IPP1),X1(NNPMM-1,IPP1)
N=NNPMM
IF (II.EQ.1) THEN
  DO 5 I=1,N-1
    DO 1 J=1,IPP1
      X1(I,J)=X(I+1,J)
1 CONTINUE
5 CONTINUE
ENDIF
IF ((II.GT.1).AND.(II.LT.N)) THEN
  DO 15 I=1,II-1
    DO 10 J=1,IPP1
      X1(I,J)=X(I,J)
10 CONTINUE
15 CONTINUE
  DO 25 I=II,N-1
    DO 20 J=1,IPP1
      X1(I,J)=X(I+1,J)
20 CONTINUE
25 CONTINUE
ENDIF
IF (II.EQ.N) THEN
  DO 35 I=1,N-1
    DO 30 J=1,IPP1
      X1(I,J)=X(I,J)
30 CONTINUE
35 CONTINUE
ENDIF
RETURN
END

```

SUBROUTINE WF(MIN,N1,N2,X1,XV,WW)

```

C THIS SUBROUTINE CALCULATES THE ANDERSON CLASSIFICATION STATISTIC, WW
C (BASED ON THE SELECTED VARIABLES) TO CLASSIFY THE OMITTED CASE, XV.
C INPUT : MIN=INDICATOR VECTOR USED TO IDENTIFY SELECTED VARIABLES.
C N1=NUMBER OF OBSERVATIONS FROM GROUP 1 IN X1.
C N2=NUMBER OF OBSERVATIONS FROM GROUP 2 IN X1.
C X1=THE DATA MATRIX WITH ONE ROW OMITTED.
C XV=THE OMITTED CASE (ROW).
C OUTPUT : WW=THE ANDERSON CLASSIFICATION STATISTIC FOR CASE XV.
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10,NN=25,MM=25,NNPMM=NN+MM)
DIMENSION X1(NNPMM-1,IP+1),XX(NNPMM-1,IP+1),XV(IP)

```

```

DIMENSION S(IP,IP),SINV(IP,IP),XM1(IP),XM2(IP)
DIMENSION MIN(IP)
C
C THE INDICATOR VECTOR MIN IS USED TO FORM THE MATRIX XX CONTAINING
C ONLY THE SELECTED VARIABLES.
C
DO 10 I=1,N1
  IC=0
  DO 5 J=1,IP
    IF (MIN(J).GT.0) THEN
      IC=IC+1
      XX(I,IC)=X1(I,J)
    ENDIF
  5 CONTINUE
10 CONTINUE
  DO 20 I=1,N2
    IC=0
    DO 15 J=1,IP
      IF (MIN(J).GT.0) THEN
        IC=IC+1
        XX(N1+I,IC)=X1(N1+I,J)
      ENDIF
    15 CONTINUE
  20 CONTINUE
  N1PN2=N1+N2
C
C THE SUBROUTINE AVGVARD IS USED TO CALCULATE THE GROUP MEANS,
C POOLED COVARIANCE MATRIX AND ITS INVERSE. ONLY THE IC SELECTED
C VARIABLES CONTAINED IN XX, ARE TAKEN INTO ACCOUNT
C
CALL AVGVARD(N1,N2,N1PN2,IC,XX,S,SINV,XM1,XM2)
SUM1=0.0D0
DO 95 I1=1,IC
  DO 90 I2=1,IC
    V1=XV(I1)-(XM1(I1)+XM2(I1))/2.0D0
    V2=XM1(I2)-XM2(I2)
    SUM1=SUM1+V1*SINV(I1,I2)*V2
  90 CONTINUE
95 CONTINUE
C WW IS THE ANDERSON CLASSIFICATION STATISTIC THAT IS USED TO CLASSIFY
C THE OMITTED CASE
  WW=SUM1
  RETURN
  END

```

```

SUBROUTINE AVGVARD(N,M,NPM,IC,XX,S,SINV,XM1,XM2)
C
C THIS SUBROUTINE CALCULATES THE GROUP MEANS (XM1 AND XM2) AND THE
C POOLED COVARIANCE MATRIX (S) AS WELL AS ITS INVERSE (SINV).
C THIS ROUTINE IS FOR THE MATRIX CONTAINING ONLY A SUBSET OF THE ROWS.
C INPUT : XX(NPM,IP) = THE FIRST N ROWS OF XX CONTAIN THE OBSERVATIONS
C FOR GROUP 1 AND THE NEXT M ROWS CONTAIN THE
C OBSERVATIONS FOR GROUP 2. ONLY THE FIRST IC
C COLUMNS ARE TAKEN INTO ACCOUNT.

```

```

C      IC=THE NUMBER OF COLUMNS TO BE TAKEN INTO ACCOUNT.
C      OUTPUT : XM1=MEAN OF GROUP 1.
C      XM2=MEAN OF GROUP 2.
C      S=POOLED COVARIANCE MATRIX.
C      SINV=INVERSE OF POOLED COVARIANCE MATRIX.
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
PARAMETER (IP=10)
DIMENSION XX(NPM,IP+1),XX1(N,IP),XX2(M,IP)
DIMENSION XM1(IP),XM2(IP)
DIMENSION S(IP,IP),SINV(IP,IP),S1(IP,IP),S2(IP,IP)
EXTERNAL DCORVC,DLINDS
DO 10 I=1,N
DO 5 J=1,IC
XX1(I,J)=XX(I,J)
5  CONTINUE
10 CONTINUE
DO 20 I=1,M
DO 15 J=1,IC
XX2(I,J)=XX(N+I,J)
15 CONTINUE
20 CONTINUE
IDO=0
NVAR=IC
IFRQ=0
IWT=0
MOPT=0
ICOPT=0
LDCOV=IP
LDINCD=1
NROW=N
LDX=N
CALL DCORVC(IDO,NROW,NVAR,XX1,LDX,IFRQ,IWT,MOPT,
&      ICOPT,XM1,S1,LDCOV,INCD,LDINCD,NOBS,
&      NMISS,SUMWT)
NROW=M
LDX=M
CALL DCORVC(IDO,NROW,NVAR,XX2,LDX,IFRQ,IWT,MOPT,
&      ICOPT,XM2,S2,LDCOV,INCD,LDINCD,NOBS,
&      NMISS,SUMWT)
NPMM2=NPM-2
DO 30 I=1,IC
DO 25 J=1,IC
S(I,J)=((N-1)*S1(I,J)+(M-1)*S2(I,J))/NPMM2
25 CONTINUE
30 CONTINUE
CALL DLINDS(IC,S,IP,SINV,IP)
RETURN
END

FUNCTION PHI(Z)
C
C      CALCULATES THE DENSITY FUNCTION OF THE STANDARD NORMAL DISTRIBUTION
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)

```

**PHI=0.3989422D0*DEXP(-0.5*Z*Z)
RETURN
END**

REFERENCES

- Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.
- Anderson, T. W. (1951). Classification by multivariate analysis. *Psychometrika* **16**, 31-50.
- Begg, C. B. and Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualised regressions. *Biometrika* **71**, 11-18.
- Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and control, 2nd edition*. San Francisco: Holden-Day.
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression : X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738-754.
- Breiman, L., Friedman, J. H., Ohlsen, R. A. and Stone, C.J. (1984). *Classification and Regression Trees*. New York: Chapman and Hall.
- Bull, S.B. and Donner, A. (1987). The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *Journal of the American Statistical Association* **82**, 1118-1122.
- Byth, K. and McLachlan, G. J. (1980). Logistic regression compared to normal discrimination for non-normal populations. *Australian Journal of Statistics* **22**, 188-196.
- Campbell, M. K., Donner, A.P. and Webster, K.M. (1991). Are ordinal models useful for classification? *Statistics in Medicine* **10**, 383-394.
- Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1**, 245-276.
- Chatterjee, S. and Chatterjee, S. (1983). Estimation of misclassification probabilities by bootstrap methods. *Communications in Statistics - Computation and Simulation* **12**, 645-656.
- Cheng, B. and Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science* **9**, 2-30.
- Chernick, M. R., Murthy, V. K. and Nealy, C. D. (1985). Application of bootstrap and other resampling techniques: evaluation of classifier performance. *Pattern Recognition Letters* **3**, 167-178.

- Chernick, M. R., Murthy, V. K. and Nealy, C. D. (1986a). Correction note to Application of bootstrap and other resampling techniques: evaluation of classifier performance. *Pattern Recognition Letters* **3**, 167-178.
- Chernick, M. R., Murthy, V. K. and Nealy, C. D. (1986b). Estimation of error rate for linear discriminant functions by resampling: non-Gaussian populations. *Computers and Mathematics with Applications* **15**, 29-37
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70**, 892-898.
- Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: SIAM.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316-331.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461-470.
- Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, jackknife, and cross-validation. *The American Statistician* **37**, 36-48.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179-188.
- Flury, B. W. (1989). Understanding partial statistics and redundancy of variables in regression and discriminant analysis. *The American Statistician* **43**, 27-31.
- Flury, B.W. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.
- Gabriel, K. R. (1969). Simultaneous test procedures - some theory of multiple comparisons. *Annals of Mathematical Statistics* **40**, 224-250.
- Ganeshanandam, S. and Krzanowski, W. J. (1989). On selecting variables and assessing their performance in linear discriminant analysis. *Australian Journal of Statistics* **31**, 433-447.

- Ganeshanandam, S. and Krzanowski, W. J. (1990). Error-rate estimation in two-group discriminant analysis using the linear discriminant function. *Journal of Statistical Computation and Simulation* **36**, 157-175.
- Geisser, S. (1964). Posterior odds for multivariate normal classifications. *Journal of the Royal Statistical Society B* **26**, 69-76.
- Geisser, S. (1966). Predictive discrimination. In *Multivariate Analysis*, P. R. Krishnaiah (Ed.). New York: Academic Press, pp.149-163.
- Geisser, S. (1982). Bayesian discrimination. In *Handbook of Statistics* (Vol. 2), P. R. Krishnaiah and L. N. Kanal (Eds.). Amsterdam: North-Holland, pp. 101-120.
- Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition* **10**, 211-222.
- Gnanadesikan, R. et al. (1989). Discriminant analysis and clustering: Panel on discriminant analysis, classification, and clustering. *Statistical Science* **4**, 34-69.
- Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. *Journal of the American Statistical Association* **81**, 108-113.
- Grizzle, J., Starmer, F. and Koch, G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489-504.
- Habbema, J. D. F. and Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics* **19**, 487-493.
- Hald, A. (1952). *Statistical theory with engineering applications*. New York: Wiley.
- Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* **89**, 1255-1270.
- Hawkins, D. M. (1976). The subset problem in multivariate analysis of variance. *Journal of the Royal Statistical Society B* **38**, 132-139.
- Hjorth, U. (1994). *Computer Intensive Statistical Methods*. London: Chapman and Hall.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

- Johnson, M. E. (1987). *Multivariate Statistical Simulation*. New York: Wiley.
- Konishi, S. and Honda, M. (1990). Comparison of procedures for the estimation of error rates in discriminant analysis under nonnormal populations. *Journal of Statistical Computation and Simulation* **36**, 105-115.
- Kshirsagar, A. M. (1972). *Multivariate Analysis*. New York: Marcel Dekker.
- Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics* **23**, 639-645.
- Lachenbruch, P. A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics* **24**, 823-834.
- Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1-11.
- Lesaffre, E. and Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society B* **51**, 109-116.
- Linhart, H. and Zucchini, W. (1986). *Model Selection*. New York: Wiley.
- Mallows, C. L. (1973). Some comments on C_p . *Technometrics* **15**, 661-675.
- Mardia, K.V., Kent, J.T. and Bibby, J. M. (1988). *Multivariate Analysis*. London: Academic Press.
- McKay, R. J. (1976). Simultaneous procedures in discriminant analysis involving two groups. *Technometrics* **18**, 47-53.
- McKay, R. J. (1977). Simultaneous procedures for variable selection in multiple discriminant analysis. *Biometrika* **64**, 283-290.
- McKay, R. J. and Campbell, N. A. (1982a). Variable selection techniques in discriminant analysis I. Description. *British Journal of Mathematical and Statistical Psychology* **35**, 1-29.
- McKay, R. J. and Campbell, N. A. (1982b). Variable selection techniques in discriminant analysis II. Allocation. *British Journal of Mathematical and Statistical Psychology* **35**, 30-41.

- McLachlan, G. J. (1973). An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. *Australian Journal of Statistics* **15**, 210-214.
- McLachlan, G. J. (1974). An asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics* **30**, 239-249.
- McLachlan, G. J. (1975). Confidence intervals for the conditional probability of misallocation in discriminant analysis. *Biometrics* **32**, 161-167.
- McLachlan, G. J. (1976a). A criterion for selecting variables for the linear discriminant function. *Biometrics* **32**, 529-534.
- McLachlan, G. J. (1976b). The bias of the apparent error rate in discriminant analysis. *Biometrika* **63**, 239-244.
- McLachlan, G. J. (1980a). On the relationship between the F-test and the overall error rate for variable selection in two-group discriminant analysis. *Biometrics* **36**, 501-510.
- McLachlan, G. J. (1980b). The efficiency of Efron's "bootstrap" approach applied to error rate estimation in discriminant analysis. *Journal of Statistical Computation and Simulation* **11**, 273-279.
- McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.
- McLachlan, G. J. and Byth, K. (1979). Expected error rates for logistic regression versus normal discriminant analysis. *Biometrics Journal* **21**, 47-56.
- Miller, A. J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.
- Murray, G. D. (1977). A cautionary note on selection of variables in discriminant analysis. *Applied Statistics* **26**, 246-250.
- O'Gorman, T. W. O. and Woolson, F. (1991). Variable selection to discriminate between two groups: stepwise logistic regression or stepwise discriminant analysis? *The American Statistician* **45**, 187-193.
- Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Annals of Mathematical Statistics* **34**, 1286-1301.
- Olivier, P. (1990). *Mislukkingsvoorspellings vir handels- en vervaardigings-ondernemings, veral met inagneming van verskillende tydskedimensies*. Unpublished Ph.D. Thesis, University of Stellenbosch.

- Page, J. T. (1985). Error-rate estimation in discriminant analysis. *Technometrics* 27, 189-198.
- Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics* 9, 705-724.
- Press, S. J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association* 73, 699-705.
- Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.
- Rencher, A.C. (1992). Bias in apparent classification rates in stepwise discriminant analysis. *Communications in Statistics - Computation and Simulation* 21, 373-389.
- Rencher, A.C. and Larson, S. F. (1980). Bias in Wilks' Λ in stepwise discriminant analysis. *Technometrics* 22, 349-356.
- Rudolpher, S. M. , Watson, P. C. and Lesaffre, E. (1995). Are ordinal models useful for classification? A revised analysis. *Journal of Statistical Computation and Simulation* 52, 105-132.
- Ruiz-Velasco, S. (1991). Asymptotic efficiency of logistic regression relative to linear discriminant analysis. *Biometrika* 78, 235-243.
- Rutter, C., Flack, V. and Lachenbruch, P. (1991). Bias in error rate estimates in discriminant analysis when stepwise variable selection is employed. *Communications in Statistics - Computation and Simulation* 20(1), 1-22.
- Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics* 24, 220-238.
- Sánchez, J. M. P. and Cepeda, X. L. O. (1989). The use of smooth bootstrap techniques for estimating the error rate of a prediction rule. *Communications in Statistics - Computation and Simulation* 18, 1169-1186.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Smith, C. A. B. (1947). Some examples of discrimination. *Annals of Eugenics* 18, 272-282.

- Snapinn, S. M. and Knoke, J. D. (1984). Classification error rate estimators evaluated by unconditional mean squared error. *Technometrics* **26**, 371-378.
- Snapinn, S. M. and Knoke, J. D. (1985). An evaluation of smoothed classification error-rate estimators. *Technometrics* **27**, 199-206.
- Snapinn, S. M. and Knoke, J. D. (1988). Bootstrapped and smoothed classification error rate estimators. *Communications in Statistics - Computation and Simulation* **17**, 1135-1153.
- Snapinn, S. M. and Knoke, J. D. (1989). Estimation of error-rates in discriminant analysis with selection of variables. *Biometrics* **45**, 289-299.
- S-PLUS Reference Manual. (1991). Statistical Sciences, Inc., Seattle.
- Toussaint, G. T. (1974). Bibliography on estimation of classification. *IEEE Transactions on Information Theory* **20**, 472-479.
- Van Ness, J.W. and Simpson, C. (1976). On the effects of dimension in discriminant analysis. *Technometrics* **18**, 175-187.
- Venter, J.H. and Steel, S.J. (1993). Simultaneous selection and estimation for the some zeros family of normal models. *Journal of Statistical Computation and Simulation* **45**, 129-146.
- Venter, J.H. and Steel, S.J. (1994). Pre-test type estimators for selection of simple normal models. *Journal of Statistical Computation and Simulation* **51**, 31-48.