# Data Modelling & Bayesian Model Comparison
## *with Spherically Symmetric Priors*

Riyaadh Jamodien

Stellenbosch University

Department of Physics




Thesis presented in partial fulfilment of the requirements for the degree of *Master of Science* at Stellenbosch University.

Supervised by

Prof. H.C. Eggers  Dr M.B. De Kock  Dr J.N. Kriel

DECEMBER 2020

I, Riyaadh Jamodien, hereby declare that the entirety of the work contained herein is my own, original work, that I am the sole author thereof, that reproduction and publication by Stellenbosch University will not infringe any third party rights and that I have not previously, in its entirety or in part, submitted this for obtaining any qualification.

Date: 11/2020

# Abstract


## Data Modelling & Bayesian Model Comparison

***with Spherically Symmetric Priors***

**Riyaadh Jamodien** **MSc (Theoretical Physics)** **November 2020**

The analysis of data is common in many fields of science, and modelling data is one of the standard techniques in such analysis. Models are, of course, not unique and many theoretical models may be constructed to describe the same set of data. When considering many competing models, we naturally ask the question: which model *best describes* the data?

It is well known that the chi-squared criterion, which is commonly cited as a goodness-of-fit between model and data, is inadequate as a measure of model quality. Rather, we employ the Bayesian framework of probability theory in addressing the question of model description of data. Within the Bayesian framework, the evidence (marginal likelihood) is the criterion by which to compare competing theoretical models. The evidence is an integral (over all parameter space) of the likelihood and prior. However, even for the simple case of linear models, there is no consensus or clarity on the choice of the best uninformative prior which enables the unbiased comparison of models with different numbers of parameters. In addressing the concern of the prior, we consider the framework of spherical symmetry, in which the evidence is reduced from an integral over a multi-dimensional space to that of a one-dimensional space, effectively reducing the problem to finding a single, optimal radial prior. We generalise existing results to a family of priors via scale relations in the form of a scaling parameter, of which several scaling relations are tested to find the best scale between models of different dimensions. We also introduce a new hyper-parameter, which had previously been conflated with model dimensions. With these developments we establish a prior that is sensitive to the new hyper-parameter, while insensitive to the model dimension, leading to the establishment of information criteria that are sensitive to these new parameters as well as the model dimension. These criteria are tested and shown to be an improvement over the existing body of work. These information criteria perform on par with widely accepted information criteria in the literature.

# Uittreksel


## Data Modellering & Bayesiaanse Model Vergelyking

***met Sferies-simmetriese Voorafwaarskynlikhede***

**Riyaadh Jamodien** **MSc (Teoretiese Fisika)** **November 2020**

Die analise van data is algemeen in verskeie wetenskaplike velde en een van die standaard tegnieke behels data modellering. Data modelle, is natuurlik nie uniek nie, en verskeie teoretiese modelle kan geskep word om dieselfde data versameling te beskryf. Wanneer verskeie modelle wat meeding in ag geneem word, ontstaan dié vraag: watter model gee die *beste beskrywing* van die data?

Dit is bekend dat die chi-kwadraat kriterion, wat algemeen in die konteks van pasgehalte tussen model en data gebruik word, onvoldoende as 'n maatstaaf van modelgehalte is. Ons maak gebruik van die raamwerk van Bayesiaanse waarskynlikheid om die vraag van modelbeskrywing van data te beantwoord. Met betrekking tot die Bayesiaanse raamwerk is die randaanneemlikheid die maatstaaf waarmee kompeterende modelle vergelyk kan word. Die randaanneemlikheid is 'n integraal (oor die gehele parameter-ruimte) van die aanneemlikheids-waarskynlikheid en die voorafwaarskynlikheid. Selfs vir die eenvoudige geval van lineêre modelle is daar egter geen ooreenstemming of duidelikheid oor die keuse van die beste oningewyde voorafwaarskynlikheid wat onbevooroordeelde vergelykings tussen modelle van verskillende dimensies toelaat nie. Om die saak van die voorafwaarskynlikheid aan te spreek, beskou ons die raamwerk van sferiese simmetrie, waarin die randaanneemlikheid vanaf 'n multi-dimensionele ruimte na 'n eendimensionele ruimte vereenvoudig word. Sodoende vereenvoudig ook die probleem tot die soektog na die enkele, optimale radiale voorafwaarskynlikheid. Ons veralgemeen die bestaande weergawes na dié van 'n familie van voorafwaarskynlikhede met skaalverhoudings in die vorm van 'n skaalparameter. Verskeie skaalverhoudings word getoets om die beste verhouding tussen modelle van verskillende dimensies te bepaal. Ons stel ook 'n nuwe hiperparameter voor wat voorheen met die algemene model dimensie verwar is. Met hierdie veralgemenings skep ons 'n voorafwaarskynlikheid wat die nuwe hiperparameter in ag neem, maar terselfdertyd die model dimensie verontagsaam, waaruit ons inligtingskriteria bepaal wat hierdie nuwe parameters in ag neem asook die model dimensie. Hierdie kriteria word teen die bestaande raamwerk van sferiese simmetrie getoets, en in vergelyking is dit aangetoon dat daar verbeteringe is. Dit is aangedui dat hierdie nuwe inligtingskriteria tred hou met die mees aanvaarde inligtingskriteria in die literatuur.

# 摘要


## 数据建模与贝叶斯模型比较
## 采用球面对称的先验概率

**张瑞雅德** **硕士（理论物理）** **2020 年十一月**

许多科学领域的研究都涉及数据分析，数据建模是数据分析的标准方法之一。数据模型当然不是唯一的，许多理论模型可以被用来描述或分析同样的数据。当考虑几种参考模型时，我们自然而然会问这个问题：哪一种模型最能描述数据？

我们都知道普遍引用的数据和模型之间适合度的卡方准则，不足以用来衡量模型质量。因此，我们采用了贝叶斯框架内的概率理论来研究数据建模的问题。在贝叶斯框架里，证据因子（边际似然）是比较几种互相竞争的数据模型的标准。证据因子是似然函数和先验概率在所有参数空间内的积分，但是，即便是简单的线性模型，都不能达成共识或者可以清楚的选择能够让我们无偏差的比较有着不同数量参数的数据模型的最好的无信息先验概率。考虑到先验概率，我们采用了球面对称的理论框架，通过将证据因子的多维参数空间的积分转化为一维参数空间的积分，问题也简化成了找到一个最佳的径向先验概率。通过测试几种参数的规模关系来寻找有不同维度模型的最佳规模，我们将现有的结论推广到一系列的先验概率。同时，我们也介绍了曾经被合并为模型参数维度的超参数。综合以上，我们建立了一个对新的超参数灵敏，但对模型维度不灵敏的的先验概率，在此引导下建立了既对新参数也对模型维度灵敏的信息准则。通过球面对称测试，这些新准则比现有的工作有进步。此外，这些新信息准则的表现可以媲美参考文献里广泛接受的信息准则。

# Acknowledgements

I extend my lasting gratitude to my teachers (primary school, high school, and beyond), to my lecturers, to the collective staff of the physics department, to the collective staff of the Confucius Institute and the Chinese language department, to those who always believed in me and never gave up on me, to those who lent a helping hand (or a shoulder to cry on), to my friends (*all* of you, no matter where in the world you may be), to my family, and most of all to my mother.

Thank you for being a part of this journey.
*You* empowered me to see further.

Special thanks to *Zero Skateboards*, *Fallen Footwear*, *Chpo*, and *Skateistan* — for all the good that they have done, and continue to do, in this world.

To the *giants*... whose shoulders *I* rest upon.

# Contents

# Preface

The presented content forms part of an ongoing body of work with unknown scope and range. Countless hours have been spent in an effort to answer some questions (whether crucial or trivial) in hopes to provide clarity and foundation to the content. It was always considered important to ask *how* and *why* something is the way it is, rather than to accept it as such, even if it may have been well-known.

Following this line of thought, an emphasis was placed on trying to achieve an intuitive understanding of certain principles that arose during the course of this work, leading it to be seemingly narrow in range. Despite this range (or lack thereof), the aimed emphasis left many known routes available for future endeavours, therefore it should be seen as fortuitous rather than lamentable.

The content is presented in a basic manner, in hopes of producing an accessible text. This resulted in material that may be quite verbose on occasion, but it is in hopes of aiding in conceptual understanding of the content.

Readers who are familiar with the basic principles of calculus and linear algebra should feel comfortable with the presented content, however, the text is designed so that readers without a strong mathematical background may still be able to follow the narrative.

With all things considered, the presented content ultimately serves as an *introductory* text and not as a reference text.

* * *

The journey contained within this text is not one of merely achieving a qualification; it is one in the pursuit of knowledge, the acquisition of experience, the exploration of thought... it is an adventure which details a journey of learning.

# Convention

## Identities

At the start of a chapter or section, a list of identities may be provided if necessary. For example, suppose we are to use Euler's identity, then it would be listed as

**Identities**

$$re^{i\theta} = r(\cos\theta + i\sin\theta)$$

When an identity is used, instructive pink highlights are applied as an indication,

$$\begin{aligned} z &= 1 + i \\ &= \sqrt{2}\left(\cos\tfrac{\pi}{4} + i\sin\tfrac{\pi}{4}\right) \\ &= \sqrt{2}\, e^{i\frac{\pi}{4}} \end{aligned}$$

Of course, a list may often contain more than one identity, and so a specific identity will be referenced when used, and where necessary. This makes the primary content less cumbersome while also providing greater clarity without unnecessary explanations.

## Discrete Mathematics

The content presented mostly follows the convention as set out by Graham et al. (1989). As such, we take note of the following, since it may not be considered a standard convention (in many circles of academia). The choices, however, were not made based on accepted convention, but based on clarity of understanding and ease of reading.

### Rising & Falling Powers

Rising and falling powers (also known as rising/falling factorials) of any number, $z$, are respectively defined as,

$$\begin{aligned} z^{\overline{k}} &:= z(z+1)(z+2)\cdots(z+k-2)(z+k-1) \\ &= \frac{\Gamma(z+k)}{\Gamma(z)}, \\ z^{\underline{k}} &:= z(z-1)(z-2)\cdots(z-k+2)(z-k+1) \\ &= \frac{\Gamma(z+1)}{\Gamma(z-k+1)}, \end{aligned}$$

where $z,\ k \in \mathbb{C}$. Rising powers were historically written as Pochhammer symbols, $z^{\overline{k}} = (z)_k$.

## Hypergeometric Functions

Hypergeometric functions, also known as hypergeometric series, form a large class of functions which include the exponential function, all the Bessel functions, and many others. For historical

reasons, there exists a plethora of different notations and conventions in common use. Consider the following notations which denote hypergeometric functions,

$$\begin{aligned}
{}_0F_1\left(\begin{matrix} \\ a \end{matrix}\,\middle|\, z\right) &= F\left(\begin{matrix} 1 \\ 1,\, a \end{matrix}\,\middle|\, z\right) = {}_0F_1(-; a; z) = {}_0F_1(; a; z) \\
&= \sum_{k \geqslant 0} \frac{1}{a^{\overline{k}}} \frac{1}{k!} z^k, \\
{}_2F_0\left(\begin{matrix} a,\, b \\ \end{matrix}\,\middle|\, z\right) &= F\left(\begin{matrix} a,\, b,\, 1 \\ 1 \end{matrix}\,\middle|\, z\right) = {}_2F_0(a,\, b; -; z) = {}_2F_0(a,\, b; ; z) \\
&= \sum_{k \geqslant 0} a^{\overline{k}} b^{\overline{k}} \frac{1}{k!} z^k.
\end{aligned}$$

Our preferred notation is introduced on the left, equated to that of Graham et al. (1989), and lastly to that of other texts which cover the subject. Unlike Graham et al. (1989), we utilise the traditional subscripts of the hypergeometric functions since these summarise the type of function we are dealing with at a mere glance.

Our notation also establishes a *position sensitive* parameter on the left of the vertical bar. The position implies that parameters appear in the numerator when elevated and in the denominator when lowered (as illustrated).

The notation is read as "zero-$F$-one", "two-$F$-one", or in general "p-$F$-q". Certain prominent hypergeometric functions, which contain paremeters in both the numerator and denominator of the series are less distinct, such as Gauss' hypergeometric function,

$${}_2F_1\left(\begin{matrix} a,\, b \\ c \end{matrix}\,\middle|\, z\right) = \sum_{k \geq 0}^{\infty} \frac{a^{\overline{k}} b^{\overline{k}}}{c^{\overline{k}}} \frac{z^k}{k!} = 1 + \frac{ab}{c} z + \frac{a(a+1)\, b(b+1)}{c(c+1)} \frac{z^2}{2!} + \ldots \quad |z| < 1,$$

or the confluent hypergeometric function,

$${}_1F_1\left(\begin{matrix} a \\ b \end{matrix}\,\middle|\, z\right) = \sum_{k \geq 0} \frac{a^{\overline{k}}}{b^{\overline{k}}} \frac{z^k}{k!}.$$

In general, each term in the series consists of ratios of rising powers of coefficients $a, b, c, \ldots$, while the real argument $z$ is taken to the ordinary power. The subscripts of $F$ indicate the number of coefficients appearing in the numerator and denominator respectively.

We stylise the exponential function as

$${}_1F_1\left(\begin{matrix} 1 \\ 1 \end{matrix}\,\middle|\, z\right) = e^z.$$

In truth, the exponential is a ${}_0F_0$ function, and to maintain the hypergeometric notation, define it with similarity to Graham et al. (1989).

Our notation is further extended when dealing with bivariate hypergeometric functions, as can be seen in Appendix B. When expressing hypergeometric functions inline, we annotate our notation as ${}_0F_1\left(\begin{smallmatrix} \\ a \end{smallmatrix}\,\middle|\, z\right)$, ${}_1F_0\left(\begin{smallmatrix} a \\ \end{smallmatrix}\,\middle|\, z\right)$, or simply ${}_2F_1(a, b; c; z)$ when unambiguous. This maintains succinctness, while also maintaining typographic consistency.

Appendix B, The Theory of Hypergeometric Functions (page 89) provides a brief introduction to readers who may be unfamiliar with this topic.

和乃不同君子之德；
定而后安大学所先。

# Introduction

The progression of this text follows the title: *(i.)* **Data modelling** *(ii.)* **& Bayesian Model Comparison** *(iii.)* ***with Spherically Symmetric Priors***.

**model** /'mɒd(ə)l/

*n.* a simplified (often mathematical) description of a system etc., to assist calculations, predictions, and understanding.

*v.* (modelled, modelling) *tr.* devise a (usually mathematical) model of (a phenomenon, system, etc.).

First we explore concepts regarding data and data modelling. When considering theoretical descriptions of data there is a requirement of *inference*, and our principles of inference are introduced with the Bayesian theory of probability. Inference would naturally depend on the model and what it aims to model from the data.

When modelling data, we postulate theoretical candidate functions (candidate models) which model any given set of data. As such, we always have *many* competing candidate models and we are interested in finding the model which *best describes* the data. In a simple sense, competing models can differ by their number of parameters, the type of functions, etc. Of course, the problem then becomes *how* the best model of many competing models is chosen. Chapters 2 and 3 provide the necessary foundations in inference, data, and data modelling.

**comparison** /kəm'pærɪs(ə)n/

*n.* The act or instance of comparing.

• *bear comparison* (often foll. by *with*) be able to be compared favourably.

Quantities such as the maximum likelihood estimate and the minimum chi-squared criterion are often cited as quantifiers of the best model, however this approach is flawed. First in that it *cannot* compare models of different complexity adequately (e.g. the comparison of models with differing number of parameters, or of different function assignments, etc.). The chi-squared criterion will always be lesser for a model of higher complexity, but this does not necessarily mean that the model provides a better description of the data. Second in that the minimum chi-squared tends to zero as a model's complexity increases.

Therefore, when concerned with model comparison, models which are overly complex are not adequately penalised in this setting.

Developments in information theory led to methods which succeed that of the chi-squared criterion/maximum likelihood in the form of *information criteria*. In general, information criteria incorporate the chi-squared criterion, but also include a *penalty* term, which sufficiently penalises models of high complexity. Akaike (1974) and Schwarz (1978) used Bayesian arguments together with Gaussian likelihoods to argue for the merits of their respective information criteria, namely *Akaike's information criterion* and the *Bayesian information criterion*. Within this context, the *minimum* information criterion refers to the best model.

In Bayesian inference, the *evidence* incorporates not only the likelihood (of the model), but also prior information in the form of a prior probability, and model comparison is effected by

an integral over the likelihood and prior. The evidence for any model has to normalise to unity, therefore a penalty is imposed on models with a greater span in parameter space. This means that the evidence naturally penalises complexity. Within this context the maximum evidence or the minimum negative-log evidence refers to the best model.

Chapter 4 details the pitfalls of the chi-squared criterion, and addresses model comparison in the context of both information criteria and the evidence.

**spherical** /'sferɪk(ə)l/
*adj.* (a) of or relating to the properties of spheres (*spherical geometry*).
(b) formed inside or on the surface of a sphere.
**symmetry** /'sɪmɪtrɪ/
*n.* (*pl.* -ies) a structure that allows an object to be divided into parts of an equal shape and size, and similar position relative to the point or line or plane of division.

With the evidence often being non-trivial to compute (e.g. a high dimensional integral), Zellner (1986) formulated a $g$-prior in an attempt to provide a prior which led to a simple analytical formula for the evidence. Unfortunately, certain choices of the parameter $g$ biases the model comparison towards either complex or simple models, irrespective of the data, thus making it unsuitable for model comparison.

In order to overcome this shortcoming, Liang et al. (2007) proposed a *mixture model* which made the $g$ parameter a variable, thereby avoiding the inconsistencies of Zellner's original $g$-prior. Liang's hyper-$g$ prior gave better results than its predecessor and it has gained traction, however its conceptual basis remains unclear.

De Kock and Eggers (2017a,b) attempted to provide a better conceptual foundation for the $g$-prior, the hyper-$g$ prior, as well as other priors by identifying the underlying concept of *spherical symmetry*. Within the framework of spherical symmetry, all of these priors were shown to be special cases that are part of the general idea of $r$-priors.

Like its predecessors, $r$-priors maintained analytic results for the evidence, and further demonstrated that spherical symmetry enables the evidence to be reduced from a high dimensional integral in parameter space, to a one dimensional integral in spherically symmetric space (along the radius). Chapter 5 details $r$-priors and the framework of spherical symmetry.

**Aim**

Within the framework of spherical symmetry there are conceptual inconsistencies that are yet to be resolved. These inconsistencies include the radius introduced with the $r$-prior and it's relation between models, the possibility of a scale between this radius and that of models in spherical space (where models have their own respective radii), the nature of how a radius could span the entire model space (from the minimal to the maximal model), and its effect on the evidence, potential for scales between fixed model parametrisations, and the like.

With these inconsistencies being identified, this text's primary investigation aims to resolve them and extend (or ammend) the existing body of spherical symmetry. These advances are identified in chapter 5, then tested and concluded upon in chapter 6.

CHAPTER 1

# Mathematical Overview

This chapter serves as an extension to the introduction, providing a brief scope of the preliminaries that are within this text. Everything that is set out here shall be considered in detail later.

## 1.1 Data with Experimental Uncertainties

We begin with an explanation of what we mean by the word data within our context. The basic situation is shown by example in figure 1.1. A scientist in some field, such as physics, psychology, economics or one of many others, has made $N$ sets of repeated measurements as a function of an independent variable $x$.

The $N = 6$ measurement points, $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$, are fixed and known exactly; they could be real numbers or bin midpoints. There is no need for equal spacing or even ordering in the sense that $x_3$ has to lie to the right of $x_2$ and to the left of $x_4$, although that is often the case. Associated with each measurement point $x_n$, there is a *data point* $y_n$, represented as a black dot in figure 1.1. Each $y_n$ is itself the result of an earlier process of deliberation and measurement. Typically, $y_n$ is an average over a number of underlying measurements, all taken at $x_n$, while the so-called *statistical error*, $s_n$, is calculated as the square root of the sum of squared variations of these underlying measurements. In data language, $y_n$ would be the sample mean and $s_n^2$ would be the sample variance of the underlying measurements. The details of calculating $y_n$ and $s_n$ do not matter for the time being; what matters is that both $y_n$ and $s_n$ are fully determined by the measurements performed by the scientist — without theoretical modelling or interpretation.

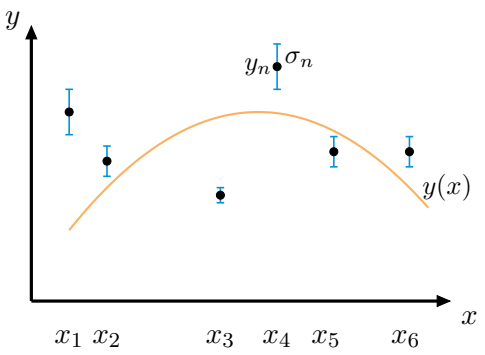



Figure 1.1: Data

To the statistical error $s_n$, the scientist will often add a *systematic error* reflecting his or her best judgement of possible underlying measurement bias or other experimental sources of error or uncertainty. The resulting *experimental uncertainty*, $\sigma_n$, is typically the square root of a sum of statistical and systematic errors,

$$\sigma_n^2 = s_n^2 + (\text{systematic error})^2$$

or some other combination of $s_n$ and the systematic error. These experimental uncertainties $\sigma_n$ are shown in figure 1.1 as vertical bars above and below each data point $y_n$, so that the total bar gives a visual indication of the uncertainty around $y_n$.

At the point where our analysis starts, both the set of measurements $\boldsymbol{y}$ consisting of all the $y_n$ as well as the set of experimental uncertainties $\boldsymbol{\sigma}$ consisting of all the $\sigma_n$ are considered fixed and given.

## 1.2 Descriptions of Data in Brief

Given data, $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{\sigma}) = \{y_n, \sigma_n\}_{n=1}^{N}$ and the known measurement points $\boldsymbol{x}$, the goal is to describe this data as accurately, but also as economically as possible. In this context, to *describe* means to find some curve which we call the *trial function* or *candidate function/model*, such that difference between the data $y_n$ and the candidate function at that point is minimal. Minimal in the sense that the sum of the squared differences between all data points and the curve at the same $x_n$, should be as small as possible.

$$Q\big(y(x), \boldsymbol{y}, \boldsymbol{\sigma}\big) = \frac{1}{N} \sum_{n=1}^{N} \frac{\big(y_n - y(x_n)\big)^2}{\sigma_n^2}, \tag{1.1}$$

We represent the candidate function with the notation $y(x)$, noting that it is a function of $x$. The $\sigma_n^2$ in the denominator of equation (1.1) ensure that those data points $y_n$ which have a small experimental uncertainty are weighted more heavily than those which do not. We include a prefactor $\frac{1}{N}$ in the definition of the squared difference $Q$ so that it converges to a constant for large $N$.

The most basic of tasks, known as *least squares* in the literature, is often applied to find some $y(x)$ such that $Q\big(y(x), \boldsymbol{y}, \boldsymbol{\sigma}\big)$ is minimised. This is, of course, as simple as much as we could hope for: simply try out all sorts of $y(x)$ and choose that one which minimises $Q$ best, then call it the *best-fit solution*. Unfortunately, that is easier said than done.

Firstly, the candidate functions could have any number and combination of functional forms such as polynomials, exponentials, trigonometric and all sorts of special mathematical functions, so $Q$ and its partial derivatives could easily become quite complicated. The choice of form is entirely up to the scientist. Minimisation in this primitive form would involve infinitely many choices of $y(x)$ with blind calculation of infinitely many values of $Q$.

Secondly, even for a given form of the candidate function, there is great variability due to different values of the underlying *parameters* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ entering into that function, which from now on we denote as $y(x \,|\, \boldsymbol{\alpha})$. By convention, all quantities taken as known appear to the right of the solid line $|$, the "solidus", while those left of the solidus are the variables. A slightly more tractable version would hence be to find the global minimum for

$$Q\big(y(x \,|\, \boldsymbol{\alpha}), \boldsymbol{y}, \boldsymbol{\sigma}\big) = \frac{1}{N} \sum_{n=1}^{N} \frac{\big(y_n - y(x_n \,|\, \boldsymbol{\alpha})\big)^2}{\sigma_n^2}, \tag{1.2}$$

with respect to the parameters by finding the stationary point $\widehat{\boldsymbol{\alpha}}$ at which

$$\left.\frac{\partial Q}{\partial \alpha_k}\right|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} = 0 \qquad \forall\, k = 1, 2, \ldots, K. \tag{1.3}$$

Of course, this may turn out to be of exceptional technical difficulty, since the dependence of $Q$ on each parameter can be highly non-linear. It is also common to find strong correlations between two or more parameters in the sense that they may vary together over a large range with only a minimal change in the value of $Q$ itself.

A technically simpler situation arises if the candidate functions, $y(x \,|\, \boldsymbol{\alpha})$, are restricted to be linear in the parameters*. The scientist in this case typically uses a set of "basis functions" $f_k(x)$

* This does not mean that they are linear in $x$.

which are added linearly to give trial functions of the form,

$$y(x \,|\, \boldsymbol{\alpha}) = \sum_{k=1}^{K} f_k(x_n)\, \alpha_k. \tag{1.4}$$

$$\mathbb{X}_{nk} = \frac{f_k(x_n)}{\sigma_n} \qquad n = 1, \ldots, N \qquad k = 1, \ldots, K, \tag{1.5}$$

which, together with the vector of standardised data

$$\boldsymbol{z} = \{z_n\}_{n=1}^{N} = \left\{ \frac{y_n}{\sigma_n} \right\}_{n=1}^{N} \tag{1.6}$$

and on substituting equation (1.4) into equation (1.2) gives us a vector-matrix form as the inner product of a row vector $(\boldsymbol{z} - \mathbb{X}\boldsymbol{\alpha})^{\mathsf{T}}$ and a corresponding column vector,

$$Q(\boldsymbol{\alpha}, \mathbb{X}, K, \boldsymbol{z}) = \tfrac{1}{N}(\boldsymbol{z} - \mathbb{X}\boldsymbol{\alpha})^{\mathsf{T}}(\boldsymbol{z} - \mathbb{X}\boldsymbol{\alpha}). \tag{1.7}$$

Minimisation by means of equation (1.3) leads to the minimum-$Q$ solution in terms of $\mathbb{X}$, its transpose $\mathbb{X}^{\mathsf{T}}$ and the data $\boldsymbol{z}$, the *Moore-Penrose pseudo-inverse*,

$$\hat{\boldsymbol{\alpha}} = (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\boldsymbol{z}. \tag{1.8}$$

## 1.3 From Least Squares to Probability Theory

The most serious problem with all the above methods is not technical but conceptual in nature. On its own, least squares is an ad hoc principle, not a "law of nature" or any fundamental concept, and it seems odd to base important judgements on the quality of physical interpretations $\big($as provided by $y(x \,|\, \boldsymbol{\alpha})\big)$ on an unjustified principle.

This is not an isolated case; we have seen that in spite of the wide use and large need of best-fit solutions, the motivation for the methods used is often not well understood, and so different proposals co-exist with different methods and different answers. For most scientists, such ambiguity is undesirable.

A first step towards a solid foundation is to understand $Q$ as being directly proportional to the *logarithm of a Gaussian probability*. The joint probability for all $N$ data points for $\boldsymbol{y}$ and $\boldsymbol{z}$ respectively, when written as Gaussians automatically contain $\frac{1}{2}NQ$ in the exponents,

$$p(\boldsymbol{y} \,|\, \boldsymbol{\alpha}) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\tfrac{1}{2\sigma_n^2}\big(y_n - y(x_n \,|\, \boldsymbol{\alpha})\big)^2\big) \tag{1.9}$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}\left(\frac{y_n}{\sigma_n} - \frac{y(x_n \,|\, \boldsymbol{\alpha})}{\sigma_n}\right)^2\right) \tag{1.10}$$

$$= (2\pi\sigma^2)^{-N/2} \exp\big(-\tfrac{1}{2}NQ\big),$$

$$p(\boldsymbol{z} \,|\, \boldsymbol{\alpha}) = (2\pi)^{-N/2} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}\left(z_n - \frac{y(x_n \,|\, \boldsymbol{\alpha})}{\sigma_n}\right)^2\right) \tag{1.11}$$

$$= (2\pi)^{-N/2} \exp\big(-\tfrac{1}{2}NQ\big).$$

The proportionality $p \propto \exp(-\frac{1}{2}NQ)$ means that

$$Q \propto -2 \log p, \tag{1.12}$$

therefore casting $Q$ as a log-probability implies that minimisation of $Q$ is equivalent to maximising $p(\boldsymbol{z} \mid \boldsymbol{\alpha})$. Conceptually, least squares therefore amounts to the *maximum likelihood*.

## 1.4 Concepts in Model Comparison

### 1.4.1 Occam's Razor & Information Criteria

A simple least squares approach has at least one other serious deficiency; minimisation of $Q$ for an unlimited choice of functions $y(x)$ and arbitrarily many parameters $\boldsymbol{\alpha}$ will almost always yield a perfect solution (i.e. $Q = 0$), if $y(x)$ is made complicated enough. For the linear candidate functions of equation (1.4), a perfect solution is almost guaranteed if a sufficient number of terms with free parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are admitted in equation (1.4).

Put differently, an almost arbitrary collection of $N$ data points $y_n$ can be reproduced exactly by using $K = N$ free parameters $\alpha_k$, while for $K > N$ there will be infinitely many perfect solutions — a highly undesirable situation.

Clearly, the formulation is missing a key ingredient, namely *economy of description*, colloquially called *Occam's razor*: the "simplicity" of a candidate function $y(x)$ should count in its favour compared to a more complicated one, even if the latter results in a smaller least squares solution of $Q$ (McElreath 2015, p. 165).

For linear cases, Occam's razor demands that a trial function with a smaller number of parameters $K$ should be able to compete with one with larger $K$, even if it does not fit as well. To compensate for the decrease of the least squares sum $Q(\boldsymbol{\alpha}, \mathbb{X}, K, \boldsymbol{z})$ with increasing $K$, we need an additional "penalty function" $F$ which increases with $K$, so that the competition between $Q$ and $F$ can lead to an overall minimum. This approach is commonly called the *information criterion* approach.

There are many information criteria in the literature, they all follow the same scheme whereby the criterion is a sum of the original $Q$, which decreases with $K$, and some specific penalty function $F$ which increases with $K$.

Two common information criteria, which this thesis will use as reference points, the first proposed by Akaike (1974), known as *Akaike's information criterion* (AIC)

$$\mathrm{AIC} = NQ + 2K \tag{1.13}$$

and the second proposed by Schwarz (1978), known as the *Schwarz information criterion*, more commonly known as the *Bayesian information criterion* (BIC),

$$\mathrm{BIC} = NQ + K \log N. \tag{1.14}$$

### 1.4.2 The Bayesian Evidence for Model Comparison

The main insight is that, given the data at hand $\boldsymbol{z}$, *model comparison* of two models $\mathcal{M}_1, \mathcal{M}_2$ can be done by means of the probabilities for each model, $p(\mathcal{M}_1 \mid \boldsymbol{z}), p(\mathcal{M}_2 \mid \boldsymbol{z})$. By first considering Bayes' rule in terms of any model $\mathcal{M}_i$, with $i = 1, 2, \ldots$,

$$p(\mathcal{M}_i \mid \boldsymbol{z}) = \frac{p(\boldsymbol{z} \mid \mathcal{M}_i)\, p(\mathcal{M}_i)}{p(\boldsymbol{z})} \tag{1.15}$$

We then consider the comparison of two models by taking the ratio of the two model probabilities as given by Bayes' rule for each respective model, i.e. $\mathcal{M}_1$ and $\mathcal{M}_2$, which turns out to be the ratio of the two *evidences*, $p(\boldsymbol{z} \mid \mathcal{M}_1)$ and $p(\boldsymbol{z} \mid \mathcal{M}_2)$,

$$\frac{p(\mathcal{M}_1 \mid \boldsymbol{z})}{p(\mathcal{M}_2 \mid \boldsymbol{z})} = \frac{p(\boldsymbol{z} \mid K_1, \mathcal{M}_1)}{p(\boldsymbol{z} \mid K_2, \mathcal{M}_2)} \tag{1.16}$$

where the evidence for model $\mathcal{M}_1$ can be found from an integral over the $K_1$ parameters $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^{K_1}$ entering into the likelihood $p(\boldsymbol{z} \mid \boldsymbol{\alpha}, K_1, \mathcal{M}_1)$ and a "prior probability" on the parameters $p(\boldsymbol{\alpha} \mid K_1, \mathcal{M}_1)$,

$$\begin{aligned} p(\boldsymbol{z} \mid K_1, \mathcal{M}_1) &= \int \mathrm{d}\boldsymbol{\alpha}\, p(\boldsymbol{z} \mid \boldsymbol{\alpha}, K_1, \mathcal{M}_1)\, p(\boldsymbol{\alpha} \mid K_1, \mathcal{M}_1) \\ \text{(evidence)} &= \int \mathrm{d}\boldsymbol{\alpha}\, \text{(likelihood) (prior)}. \end{aligned} \tag{1.17}$$

and likewise for the evidence $p(\boldsymbol{z} \mid \mathcal{M}_2)$ which is a $K_2$-fold integral over the $K_2$ parameters of $\mathcal{M}_2$. Maximising the evidence includes minimisation of $Q$ but also penalises models with larger $K$, thereby providing a natural inclusion of Occam's razor. characteristics of the evidence and its associated quantities are detailed in chapter 4.

Both the AIC and BIC emerge as special cases of evidence calculations, as do maximum likelihood approaches, and of course all lower-order criteria such as least squares.

The evidence and its calculation is therefore crucial; if the underlying likelihood is non-linear in the parameters $\boldsymbol{\alpha}$, the $K$-fold integral must be calculated numerically, which for large $K$ may be quite challenging.

For linear models as in equation (1.4) entering the likelihood and some choices of prior, however, these integrals can be calculated analytically.

The problems identified so far have been solved conceptually on the most fundamental level by the Bayesian viewpoint, including calculations thereof. However, the introduction of priors of the type $p(\boldsymbol{\alpha} \mid \mathcal{M})$ results in an additional concern. The main challenge is to construct a prior, in mathematical form, which reflects our state of (available) knowledge on the one hand, but does not overspecify in the process. Within our context, we address this concern in section 4.5.

CHAPTER 2

# The Bayesian Theory of Probability

## 2.1 Introduction

Two friends, RiRi and JamJam, are visiting a nearby farm to collect fresh fruit from the trees on a cool, cloudy morning. Upon blueberry and strawberry picking, they notice that the area around the trees is fairly moist. RiRi looks down to notice sprinklers and concludes that the moisture is a result of the sprinkers. JamJam pauses for a moment, then states that things aren't that simple, since there could be many possible reasons. She hypothesises that there may be a possibility of it being due to morning dew, a morning drizzle, the sprinklers, or a number of any other possibilities. RiRi asserts his argument as a certainty, while JamJam asserts that his argument is only the most probable via her reasoning.

At first glance, RiRi may seem to have deduced the correct answer, while JamJam provided reasonable counter arguments based on plausible possiblities — if, at any moment, it starts to drizzle, then JamJam discounts RiRi's argument even though he stated it as a certainty. The primary difference between their means of reasoning is that RiRi reaches his conclusions based upon statements and premises, while JamJam, on the contrary, reaches her conclusions based upon weighing considerations of probable possibilities based on observations.

Notice that RiRi's reasoning takes the cause (sprinkler) and reasons a consequence (moisture), we may refer to this type of reasoning as *deductive reasoning*. JamJam's reasoning, on the other hand, takes a hypothesis or probable cause (dew, drizzle, sprinkler, etc.) following an observation (moisture), which we refer to as *inductive reasoning*.

(i) Deductive reasoning (RiRi): black points represent a cause and coloured points represent its consequences.

(ii) Inductive reasoning (JamJam): coloured points represent (observed) consequences while black points represent their probable causes.

Figure 2.1: Visual comparison of deductive and inductive reasoning.

If we are to categorise the manner in which RiRi and JamJam reason, we are able to present two models of reasoning — one involving certainties (RiRi: deductive) and another involving inferences or uncertainties (JamJam: inductive).

Figure 2.1(ii) illustrates that inductive reasoning may be more involved than that of deductive reasoning, since any observation may have many probable causes, all of which need to be weighed or considered against each other. Furthermore, several observations may have the same probable causes and it is necessary to infer which probable cause provides the most reasonable explanation of any given observation. While, on the other hand, the consequences of deductive reasoning are (often) independent results stemming from an initial cause and are then reached having followed a successive logical steps (no inferences or uncertainties).

Deductive reasoning is the type of reasoning used by mathematicians, where a mathematical

proof is based upon a premise and proven either true or false. Inductive reasoning is the type of reasoning used by scientists, who are often faced with some natural phenomena (potentially represented by data) and aim to describe the underlying principle thereof (this is often referred to as inverse or plausible reasoning).

Logic is an algebra founded on modelling systems of *reasoning*. In other words, a construction which attempts to explain *how we reason*. As such, we may attempt to explain (or model) the presented systems of reasoning, as was used by RiRi and JamJam, through a framework of logic.

## 2.2 Foundations of Bayesian Probability Theory

### 2.2.1 Logic of Propositions

Propositional logic is an algebra that studies logical expressions involving propositions and their truths or falsities. For example, the state of any proposition $x$ being true means that the proposition *not* $x$ is false. This form of propositional logic is known as classical propositional logic, because it models propositions based on Boolean values: $1 \equiv$ true and $0 \equiv$ false.

Propositional logic introduces an algebra for *logical expressions* with Boolean algebraic operands as well as logical operators, such as `AND`, `OR`, and `NOT`. The operators operate on propositions (the operands) with Boolean values (Aho and Ullman 1994, p. 642). As such, (classical) propositional logic forms a Boolean algebra.

**Example 2.1** **A Computer Science Student's Struggle**

A student trying to understand the usage of conditional statements programmed the following in the Python programming language,

$$\texttt{if } \big(a > b \texttt{ or } (a <= b \texttt{ and } c < d)\big)\texttt{:}$$

This conditional was programmed to include three expressions which needed to be validated, *i.* $(a > b)$, *ii.* $(a <= b)$ as well as *iii.* $(c < d)$, with expressions *ii* and *iii* being validated together.

They later realised that the above conditional may be replaced by the following one which contains only two expressions, yet produces the same result,

$$\texttt{if } (a > b \texttt{ or } c < d)\texttt{:}$$

They reasoned that the first conditional succeeded if, in one case, the expression $(a > b)$ is true, and in the other case when $(a <= b)$ and $(c < d)$ are both true given that the first expression is false. They also reasoned that $(a <= b)$ is the negation (i.e. falsehood) of the first expression (meaning that $a \leq b$ in this case), therefore if the statement of the first expression fails, it implies that it is false, thus only the third expression is necessary, yielding the second conditional which contains only two expressions (*i* and *iii*).

The expressions, which can be referred to as *logical statements*, presented within the conditionals of example 2.1 are either true or false, provided their conditions are fulfilled (or not). Therefore we can, for example, represent $(a > b)$ by $x$, $(a <= b)$ by `NOT` $x$ and $(c < d)$ by $y$ — these symbols (or combinations thereof) are generally known as propositional variables since they can

represent any proposition (i.e. logical statement) which can be either true or false. These are known as atomic operands of propositional logic.

As stated, propositional logic forms an algebra, and under the given operators contained within this logic, may, for any propositional variables $x$ and $y$, be defined as

- Negation: (NOT $x$) is true when $x$ is false, and it is false otherwise.
- Conjunction: ($x$ AND $y$) is true if both $x$ and $y$ are true, it is false otherwise.
- Disjunction: ($x$ OR $y$) is true if either $x$ or $y$ or both are true, it is false otherwise.

Here we note that propositional variables, along with the Boolean values, represent what is known as *logical expressions*. Logical expressions may be constructed through the binary operators AND and OR as well as the unary operator NOT (Aho and Ullman 1994, p. 645).

In addition to these core operators, there are four additional logical operators,

- Implication: ($x \implies y$) means "if $x$ then $y$" and is only false when $x$ is false while $y$ is true.
- Equivalence: ($x \equiv y$) is true if and only if $x$ and $y$ are either both true or both false, it is false otherwise.
- Negated Conjunction: ($x$ NAND $y$) applies the AND operator followed by the NOT operator to its operands.
- Negated Disjunction: ($x$ NOR $y$) applies the OR operator followed by the NOT operator to its operands.

The operators have precedence in the following descending order: NOT, NAND, NOR, AND, OR, $\implies$, and $\equiv$. Precedence implies that operators have a logical order, for instance, consider that NOT $x$ $\equiv$ NOT $y$ AND $z$ would be grouped as (NOT $x$) $\equiv$ $\big($(NOT $y$) AND $z$$\big)$ (Aho and Ullman 1994, p. 652). Truth tables may, naturally, be defined for all of these operators and while the study of truth tables may be common in classical propositional logic, it provides little gain within the current setting.

These operators (omitting joint operators) have symbolic representations defined as,

$$\texttt{NOT} \quad \overline{x} \tag{2.1}$$

$$\texttt{AND} \quad (x, y) \tag{2.2}$$

$$\texttt{OR} \quad (x + y) \tag{2.3}$$

The logical AND may often also be represented by $xy$, much akin to multiplication. The operator NAND is symbolically known as $\Uparrow$, while NOR is symbolically known as $\Downarrow$.

Logical expressions involving propositions and operators have various possible results based on the assignments of the Boolean values and for any expression, $\mathcal{F}$, this collection of results (stemming from an expression or expressions) is known as a Boolean function*. Any Boolean function can be represented by the core operators given by equations (2.1) to (2.3) (Aho and Ullman 1994, p. 658).

* This is similar to the standard definition (in analysis) of a function, $f$, which maps $x$ to $y$.

The three core operators form a complete set, meaning that every Boolean function can be represented through these three operators. Additionally, this means that the logical operators which are not part of the core three (such as implication) may all be represented by the core three operators. We may see this through the following representations of implication and equivalance. Let the statement "if and only if" (symbolised by $\iff$) represent equality between two logical expressions. Meaning one expression is true *if and only if (!)* the other is true.

$$\begin{aligned} &\text{Implication} && (\overline{x} + y) \iff (x \implies y) \\ &\text{Equivalence} && \big((x \implies y), (y \implies x)\big) \iff (x \equiv y) \end{aligned}$$

Here we first defined the implication, afterwhich we used it to define equivalence. From this, it is clear that all logical principles may follow merely from the core operators.

The NAND and NOR operators are both, individually, complete, because they can independently represent the three core operators (Aho and Ullman 1994, p. 659). For example, for NAND, we have that,

$$\begin{aligned} (\texttt{NOT}\ x) &\equiv (x\ \texttt{NAND}\ 1) & \overline{x} &\equiv (x \Uparrow 1) \\ (x\ \texttt{AND}\ y) &\equiv \big((x\ \texttt{NAND}\ y)\ \texttt{NAND}\ 1\big) & (x, y) &\equiv \big((x \Uparrow y) \Uparrow 1\big) \\ (x\ \texttt{OR}\ y) &\equiv \big((x\ \texttt{NAND}\ 1)\ \texttt{NAND}\ (y\ \texttt{NAND}\ 1)\big) & (x + y) &\equiv \big((x \Uparrow 1) \Uparrow (y \Uparrow 1)\big) \end{aligned}$$

#### 2.2.1.1 Algebraic Rules for Propositional Logic

Within the Boolean algebraic structure, the operators obey the laws of the algebra and we may expand our knowledge of logical expressions to encompass these laws.

Since the $\equiv$ operator is the operator with the lowest precedence, it is also the one fulfilling the most algebraic properties. As such, we first consider the range of this operator before the others.

$$\begin{aligned} &\text{Law of Reflexivity} && x \equiv x \\ &\text{Law of Transitivity} && \big((x \equiv y)\ \texttt{AND}\ (y \equiv z)\big) \implies (x \equiv z) \\ &\text{Law of Commutativity} && (x \equiv y) \equiv (y \equiv x) \\ &\text{Equivalence of Negations} && (x \equiv y) \equiv (\overline{x} \equiv \overline{y}) \end{aligned}$$

The three core operators are analogous to the standard operators of arithmetic $(-, +, \times)$ (Aho and Ullman 1994, p. 675),

$$\text{Law of Associativity} \qquad \begin{aligned} \big(x, (y, z)\big) &\equiv \big((x, y), z\big) \\ \big(x + (y + z)\big) &\equiv \big((x + y) + z\big) \end{aligned} \tag{2.4}$$

$$\text{Law of Commutativity} \qquad \begin{aligned} (x, y) &\equiv (y, x) \\ (x + y) &\equiv (y + x) \end{aligned} \tag{2.5}$$

$$\text{Law of Distributivity} \qquad \begin{aligned} \big(x, (y + z)\big) &\equiv \big((x, y) + (x, z)\big) \\ \big(x + (y, z)\big) &\equiv \big((x + y), (x + z)\big) \end{aligned} \tag{2.6}$$

The commutative law asserts that the order of operation refers only to the order in which they are stated and not the order in which events occur* Cox (1961, p. 5).

* In spoken language, on the contrary, stating $x$ first and then $y$ may have a different meaning than stating $y$ then $x$.

The distributive law of AND distributing over a proposition of OR states that either ($x$ AND $y$) is true or that ($x$ AND $z$) is true. The distributive law of OR distributing over a proposition of AND states that both ($x$ OR $y$) and ($x$ OR $z$) are true.

Note that the distributive law of AND distributing over OR is analogous to the distributive law in arithmetic, while the distributive law of OR distributing over AND is *neither* analogous *nor* equivalent to the distributive law of arithmetic.

Given that AND as well as OR are both associative, it implies successive propositions may be grouped without internal ordering, we may therefore consider the natural extension of these operators to $K$ propositions. Thus, for any collection of $K$ propositions $x_1, x_2, \ldots, x_K$ we have a representation* for the disjunctions and conjunctions for collections of propositions,

$$\begin{aligned}\left(x_1 \texttt{ AND } x_2 \texttt{ AND } \ldots \texttt{ AND } x_K\right) \equiv \left(\texttt{AND } x_1, x_2, \ldots, x_K\right) &\equiv \left(x_1, x_2, \ldots, x_K\right) \\ &\equiv \left(\bigotimes_{k=1}^{K} x_k\right)\end{aligned} \tag{2.7}$$

$$\begin{aligned}\left(x_1 \texttt{ OR } x_2 \texttt{ OR } \ldots \texttt{ OR } x_K\right) \equiv \left(\texttt{OR } x_1, x_2, \ldots, x_K\right) &\equiv \left(x_1 + x_2 + \cdots + x_K\right) \\ &\equiv \left(\bigoplus_{k=1}^{K} x_k\right)\end{aligned} \tag{2.8}$$

The centre expressions apply the operators to the collection of $K$ propositions. In the case of the AND expression between $K$ propositions, the value of this expression will be true when all propositions are true. In the case of the OR expression, the value will be false when all propositions are false and true otherwise (Aho and Ullman 1994, p. 651). This property is known as *consistency* under extension to $K$ propositions.

If in a collection of propositions *at least one* of many propositions is true, then the disjuntion between them is true and spans the entire collection of propositions, then such a collection is known as *mutually exhaustive*. If in a collection of propositions the conjunction between a combination of any and all propositions is false, then such a collection is known as *mutually exclusive*.

Here we introduced the symbolic shorthand notation for AND as well as OR for a collection of propositions. Each are analogous to indexed notations from arithmetic, where the indexed AND notation, $\bigotimes_k$, is analogous to the notation of products, $\prod_k$, and the indexed OR notation, $\bigoplus_k$, is analogous to that of summations, $\sum_k$.

Furthermore, there are additional rules of note for both the operators AND as well as OR,

$$\text{Identity} \qquad \begin{aligned}(x, 1) &\equiv 1 \\ (0 + x) &\equiv x\end{aligned} \tag{2.9}$$

$$\text{Annihilation} \qquad \begin{aligned}(x, 0) &\equiv 0 \\ (1 + x) &\equiv 1\end{aligned} \tag{2.10}$$

$$\text{Idempotence} \qquad \begin{aligned}(x, x) &\equiv x \\ (x + x) &\equiv x\end{aligned} \tag{2.11}$$

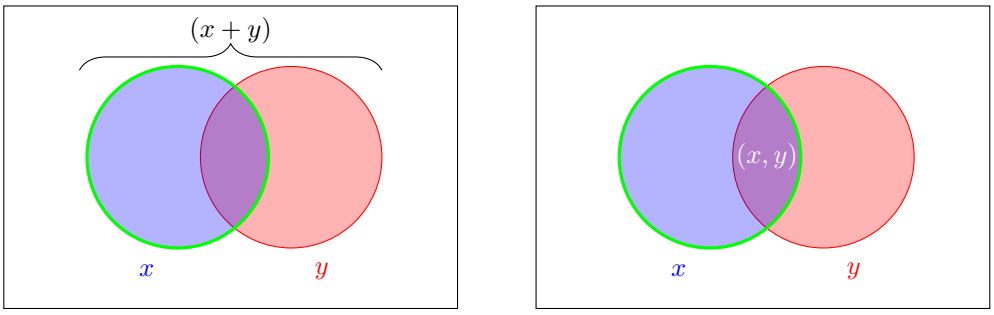
* To avoid ambiguity between a collection of propositions and a conjunction of a collection of propositions, note that conjunctions are always contained in parentheses.

Subsumption

$$\big(x,(x+y)\big) \equiv x \tag{2.12}$$

$$\big(x+(x,y)\big) \equiv x \tag{2.13}$$

Firstly, note that idempotence asserts (in both cases) that a proposition has been *stated twice* and *not* that an event has occurred twice Cox (1961, p. 6). Secondly, note that subsumption may (in both cases) be evaluated through the distributive law.




(i) Subsumption over conjunction (AND) (ii) Subsumption over disjunction (OR)

Figure 2.2: Venn diagrams illustrating subsumption

In figure 2.2(i) we have the conjunction between $x$ and $(x+y)$, which we refer to as subsumption over AND $\big($equation (2.12)$\big)$. In figure 2.2(ii) we have the disjunction between $x$ and $(x,y)$, which we refer to as subsumption over OR $\big($equation (2.13)$\big)$. In either case (over AND as well as over OR) the illustrations establish the propositions and expressions, then perform the respective operations *between* them, with the result receiving a green highlight ($x$ in both cases).

Conceptually, we may think of subsumption as follows, suppose $x \equiv$ "blueberries in a basket" and $y \equiv$ "strawberries in a basket", the disjunction (OR) between these two propositions would be to have two baskets with each, while the conjunction (AND) between these two would be to have a small bowl and adding a little of both berries, but the bowl is only given upon request.

Now, subsumption as illustrated in figure 2.2(i) $\big($over AND, given by equation (2.12)$\big)$ requests "only blueberries" between all of the options (two baskets in this case), thus the result is the basket containing the blueberries ($x$). Similarly, subsumption as illustrated in figure 2.2(ii) $\big($over OR, given by equation (2.13)$\big)$ requests "all blueberries", but between the small bowl (which was first requested) and the basket containing blueberries, but the blueberries are removed from the bowl and added to the basket, hence the result is the original basket of blueberries ($x$).

There are two additional laws, within the structure of Boolean algebra, that allow for the expressions of conjunctions and disjunctions to be expressed through the use of negations (Aho and Ullman 1994, p. 677). These laws are known as De Morgan's laws and relates an expression of a negated conjunction to an expression of a disjunction and the expression of a negated disjunction to an expression of a conjunction,

De Morgan's First Law $$\overline{(x,y)} \equiv (\overline{x}+\overline{y}) \tag{2.14}$$

De Morgan's Second Law $$\overline{(x+y)} \equiv (\overline{x},\overline{y}) \tag{2.15}$$

De Morgan's first law states that both $x$ and $y$ are false when at least one of them is false, and the second law states that neither $x$ nor $y$ is true if, and only if, they are both false. Both laws may be extended to a collection of $K$ propositions without loss of generality.

When observing De Morgan's laws, note that if we exchange `AND` in the first law with `OR`, we immediately recover the second law (and vice versa). Furthermore, observing the other laws introduced, equivalences seem to appear in pairs, in which `AND` and `OR` are exchanged. Such exchanges may be extended to propositions, such as between *true* and *false*, any proposition becoming `NOT` of itself, etc. — but we see that the algebra *remains logically consistent*. This phenomenon is known as *the principle of duality*, which in essence states that exchanging every atomic operand by its dual (the counterpart within a pair), yields the dual result.

Lastly, we introduce the following,

$$(x + \overline{x}) \equiv 1 \tag{2.16}$$

This is known as *the law of excluded middle*, and it states that either a proposition is true or its negation is true — *nothing* inbetween*. And similarly,

$$(x, \overline{x}) \equiv 0 \tag{2.17}$$

Is known as *the law of non-contradiction* and it states that two contradictory propositions cannot be true at the same time. We can refer to the statement made by equation (2.16) as a *truism*; it states a definite truth, and the statement made by equation (2.17) as a *falsity*; it states a definite fallacy.

As a consequence of these two laws, we have the principle of elimination:

$$\begin{aligned} \big(x + (\overline{x}, y)\big) &\equiv (x + y) \\ \big(x, (\overline{x} + y)\big) &\equiv (x, y) \end{aligned} \tag{2.18}$$

Which results from the laws of excluded middle and non-contradiction, and it is essentially subsumption over negation.

It is of interest to note that set theory also forms a Boolean algebra, therefore the principles of propositional logic are analogous to that of set theory. When considering this analogy, propositions are equivalent to sets and the core operators of propositional logic can be mapped to those of set theory as (`NOT`, `AND`, `OR`) $\mapsto$ (`complement`, `intersection`, `union`). In addition to this, for any two propositions $x$ and $y$, and any two sets $X$ and $Y$, we may represent this mapping symbolically as $\big(\overline{x},\ (x, y),\ (x + y)\big) \mapsto (X^c,\ X \cap Y,\ X \cup Y)$.

### 2.2.2 Logic of Uncertainties

Propositional logic provides a framework of reasoning involving strict truths and falsities — reasoning involving certainties. In other words, propositional logic characterises *deductive* reasoning. The natural extension to this framework would be to produce a framework of reasoning involving *uncertainties* — one which characterises *inductive* reasoning. We consider propositional (i.e. classical) logic as the foundation to the development of a non-classical theory of logic that is capable of reasoning within uncertainties.

* Many non-classical frameworks of logic are founded based on the *exclusion* of this law, for example, intuitionist logic.

We are first required to motivate the necessities of reasoning under uncertainties; even though the foundation of this reasoning consists of propositional logic, succeeding strict certainties naturally implies that outcomes can no longer be restricted to that of only true and false, meaning that Boolean values are no longer the only possibilities. Within propositional logic we were interested in the evaluation of the state of a proposition $x$ (as well as logical expressions thereof) being either true or false.

Therefore, it is necessary to introduce the *state of uncertainty* in the proposition $x$, namely $\psi(x)$. The state of uncertainty in a proposition evaluates that proposition and provides an assertion about how convincing (or plausible) it may be*. The level of how convincing a state of uncertainty may be is referred to as our *conviction* in the state of uncertainty in a proposition. We shall refer to this construction of logic as *uncertainty logic*.

As such, we adapt the following desiderata†, provided by Jaynes (2003, p. 19), as the premise of what is necessary to produce rational reasoning within uncertainties.

**Definition 2.1** **Jaynes' Desiderata**

I. States of uncertainty are represented by real numbers.

II. Qualitative correspondence with common sense.
   i. If knowledge about the truth in a proposition increases, then the conviction about its state of uncertainty increases.
   ii. Small changes in knowledge of a proposition yields small changes in the conviction about its state of uncertainty.

III. Consistency within reasoning of propositional logic.
   i. All possible ways to deduce a particular result should lead to the same result.
   ii. All known propositions that can be taken into account should be taken into account.
   iii. Equivalent knowledge in states of propositions are represented by equivalent convictions in states of uncertainties.

Having described the states of uncertainty in any propositions $x$ and $y$ as $\psi(x)$ and $\psi(y)$, we now introduce the state of uncertainty in $x$ provided knowledge on the proposition $y$: $\psi(x \mid y)$. This notation is read as "the state of uncertainty in $x$ *given* $y$". Conversely, if knowledge on the state of uncertainty in proposition $x$ is known, then the state of uncertainty in $y$ *given* $x$ is $\psi(y \mid x)$.

Regarding any *knowledge* on propositions, as described in the desiderata, we note that it may interchangeably be referred to as information and should always to be considered when evaluating a state of uncertainty. We denote this as $\psi(x \mid y, \kappa)$, where knowledge is represented by $\kappa$. We have that all states of uncertainty are sensitive to one's knowledge, and therefore we accept knowledge to be *implicit*, $\psi(x \mid \kappa) \equiv \psi(x)$, for any state of uncertainty.

This notation is known as the *conditional*‡, since the state of uncertainty in one proposition is

* Exactly how JamJam's reasoning meant that the sprinkler may have been most reasonable, but given an instance of rain, the rain became the more convincing reason for the presence of moisture. † Desideratum (*pl.* desiderata) refers to something that is lacking but needed or desired. ‡ In spoken and computational languages the conditional refers to *if-then* statements. These two conditionals should not be confused.

conditioned on the knowledge of another proposition. This notation provides us with a measure of assessing our conviction in a state of uncertainty in one proposition provided knowledge of another (known) proposition.

### 2.2.3 Axioms of Uncertainty Logic

An in depth, modern formalism for uncertainty logic is developed in Appendix A Foundations of Bayesian Probability Theory (page 70), which establishes consistency between uncertainty logic and probability theory.

Readers who are more interested in the *usage* of uncertainty logic, may proceed directly to the following subsection with knowing that we may henceforth express our *conviction* in a state of uncertainty in a proposition as the probability of that proposition: $\psi(x) \equiv p(x)$, in other words probability measures (or evaluates) our conviction.

### 2.2.4 Rules of Probability Theory

The consistency between uncertainty logic and probability theory means that we are now only to review the rules of probability theory from the established framework of uncertainty logic. We reiterate that all probabilities are always conditioned on one's knowledge, $\kappa$, even if not explicitly expressed.

The truism and falsity are defined as the definite probabilities,

$$p(x \,|\, x) = 1 \tag{2.19}$$

$$p(x \,|\, \overline{x}) = 0 \tag{2.20}$$

Equation (2.19) establishes the *scale* (of magnitude) of normalised probabilities*, with the truism of propositional logic corresponding to the truism in probability, and equation (2.20) corresponds to the falsity in propositional logic. Equation (2.20) is also the probabilistic equivalent to the law of non-contradiction — there can be no assured probability of $x$ when $\overline{x}$ is known to be true.

Following this reasoning, we express the total probability of this system as

$$1 = p(x \,|\, \kappa) + p(\overline{x} \,|\, \kappa) \tag{2.21}$$

Which is known as *the sum rule*. This rule is much alike to the disjuntion between a proposition and its contradiction: $x + \overline{x} \equiv 1$.

Upon further exploration we now consider compound propositions with the operators `AND` as well as `OR` (having just established `NOT`). The probability of a logical conjunction (`AND`) may be represented as a *joint* probability between propositions,

$$p\big((x, y) \,|\, \kappa\big) = p(x \,|\, y, \kappa)\, p(y \,|\, \kappa) = p(y \,|\, x, \kappa)\, p(x \,|\, \kappa) \tag{2.22}$$

This decomposition of the joint probability is known as *the product rule*. This rule originates from our desiderata, which asserts that compound statements should decompose into simpler

* The preceding section affirm this scale to be arbitrary

statements (in agreement with propositional logic). Note that there is no ordering within the product rule, which is inherited from the commutativity of conjunctions of logical propositions.

If it so happens that knowledge on the state of proposition $y$ has no influence on the state of knowledge in proposition $x$ (or vice versa), then the conditional in the probability $p(x \mid y, \kappa)$ or $p(y \mid x, \kappa)$ loses meaning of that which it is conditioned on, meaning that the probability of $x$ is unaffected by any knowledge gained from knowing $y$ (or vice versa). This is known as *conditional independence:*

$$p(x \mid y, \kappa) \overset{\mathsf{CI}}{=} p(x \mid \kappa) \tag{2.23}$$

As a result of conditional independence, the product rule reduces into a product of the individual probabilities of the conjoined propositions. In other words, joint probabilities factorise into a product of the probabilites of the propositions in the conjunction,

$$p\big((x, y) \mid \kappa\big) \overset{\mathsf{LI}}{=} p(x \mid \kappa)\, p(y \mid \kappa) \tag{2.24}$$

This is known as *logical independence*. It is worth noting that logical independence is *not* a property of an event or an object — the assertion of logical independence is based on our state of knowledge.

The probability of the logical disjunction (`OR`) can be evaluated via De Morgan's second law (equation (2.15)) as follows,

$$\begin{aligned} p\big((x + y) \mid \kappa\big) &= 1 - p(\overline{x} \mid \overline{y}, \kappa)\, p(\overline{y} \mid \kappa) \\ &= p(y \mid \kappa) - p\big((x, \overline{y}) \mid \kappa\big) \\ &= p(y \mid \kappa) - p(x \mid \kappa)\, p(\overline{y} \mid x, \kappa) \\ &= p(y \mid \kappa) - p(x \mid \kappa)\big(1 - p(y \mid x, \kappa)\big) \end{aligned}$$

From which *the generalised sum rule* follows,

$$p\big((x + y) \mid \kappa\big) = p(x \mid \kappa) + p(y \mid \kappa) - p\big((x, y) \mid \kappa\big) \tag{2.25}$$

This rule is also the probabilistic equivalent of *the inclusion-exclusion principle*.

If a collection of propositions is mutually exclusive (the conjunction between propositions is false), then the generalised sum rule reduces to

$$p\big((x + y) \mid \kappa\big) = p(x \mid \kappa) + p(y \mid \kappa) \tag{2.26}$$

Extending the product and generalised sum rules to a collection of $K$ propositions, $x_1, x_2, \ldots, x_K$, the product rule yields

$$p\Big(\bigotimes_{k=1}^{K} x_k \;\Big|\; \kappa\Big) = p(x_1 \mid x_2, \ldots, x_K)\, p(x_2 \mid x_3 \ldots, x_K) \cdots p(x_{K-1} \mid x_K)\, p(x_K \mid \kappa) \tag{2.27}$$

Similarly, the generalised sum rule (or the generalised inclusion-exclusion principle) becomes

$$p\Big(\bigoplus_{k=1}^{K} x_k \;\Big|\; \kappa\Big) = \sum_{j=1}^{K} (-1)^{j+1} \sum_{k=1}^{\binom{K}{j}} p\Big(\bigotimes_{\ell = k \in \sigma_j(K)} x_\ell \;\Big|\; \kappa\Big) \tag{2.28}$$

Here $\sigma_j(\kappa)$ represents all unique permutations of cardinality $j$ from the set $\{1, 2, \ldots, K\}$. An example hereof could be seen with $\sigma_3(4) = \big\{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}\big\}$, and these (set) elements are the values of the index $\ell$ for respective $j$ and $k$ (of which there are $\binom{K}{j}$ terms), so if $k = 3$, then $\ell$ performs the conjunction $(x_1, x_3, x_4)$.

For any collection of mutually exhaustive and exclusive propositions, $y_1, y_2, \ldots, y_K$, we have

$$p\big((y_1 + \cdots + y_K) \,|\, \kappa\big) = 1 \qquad\qquad p\big((y_i, y_j) \,|\, \kappa\big) = 0 \quad , \quad i \neq j$$

We may then partition the probability of proposition into this collection as a joint probability,

$$\begin{aligned} p(x \,|\, \kappa) &= p\big((x, y_1) + \cdots + (x, y_K) \,|\, \kappa\big) \\ &= \sum_{k=1}^{K} p\big((x, y_k) \,|\, \kappa\big). \end{aligned} \tag{2.29}$$

Summing or integrating over a parameter is known as *marginalisation* and the partitioning of a probability of a proposition into the collection as a marginalised joint probability is known as the *marginal*, and the marginalised quantity is known as the *evidence*.

We may thus state the idea of logical independence as a joint probability which factorises into a product of its marginals (given a state of knowledge).

From the symmetry of the product rule $\big($equation (2.22)$\big)$, we may express the conditional probability of a proposition as

$$p(x \,|\, y, \kappa) = \frac{p(y \,|\, x, \kappa)\, p(x \,|\, \kappa)}{p(y \,|\, \kappa)}. \tag{2.30}$$

Which is a generic form of *Bayes' rule*. From here onward, we relax our notation for the probabilities of compound propositions,

$$p\big((x, y) \,|\, \kappa\big) \equiv p(x, y \,|\, \kappa) \qquad\qquad p\big((x + y) \,|\, \kappa\big) \equiv p(x + y \,|\, \kappa)$$

CHAPTER 3

# Data & Data Modelling

## 3.1 The Relationship Between Data, Description & Simulation

It is very important to understand that there are three *different* aspects pertaining to data. While the mathematics pertaining to each aspect may appear to be similar or even the same, the concepts in each aspect are very distinct.

I. **Physical Data and How it Originated**

The concept of physical data was briefly introduced in section 1.1; we treat the topic more thoroughly in section 3.2 below.

II. **The Theoretical Description of Existing data**

Once data is available and it is believed to be true and interesting, science has a need to describe it as well as possible. The description comes on four levels:

i. On the lowest level, the existing data is cast into a different form or summarised in terms of purely data-derived quantities. No theoretical input into these quantities is allowed or desired, although of course judgement and theory would motivate their definition and use. The most well-known examples are the sample mean and sample variance

$$\begin{aligned}\langle y\rangle &:= \frac{1}{N}\sum_{n=1}^{N} y_n \\ k_2 = \mathrm{var}(y) &:= \langle y^2\rangle - \langle y\rangle^2 \\ &= \left[\frac{1}{N}\sum_{n=1}^{N} y_n^2\right] - \left[\frac{1}{N}\sum_{n=1}^{N} y_n\right]^2\end{aligned}$$

which can be computed directly from the data without any theory.

ii. On the next level, the scientist uses their insight and inspiration to propose a *model*, a theoretical description of the data. There are many different types of models. In this text, we concentrate only on *linear regression* models for the likelihood (treated at length in section 3.3) coupled to various priors which form the subject of later sections and chapters, as well as the focus of this text.

Very importantly, *Bayesian* models consist not only of a choice of the usual probability $p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \mathcal{M}_K)$ for the data given the parameters i.e. the likelihood, but also of the *prior* for those parameters, $p(\boldsymbol{\alpha} \mid \mathcal{M}_K)$.

iii. On the third level, there are two or more competing models proposed for the same data. The question is then which of these competing models provides the best description thereof (as detailed in chapter 4).

iv. On the fourth and highest level, the goal is to find and implement a systematic set of criteria and methods to design and compare different model comparison methods. In other words, there is a need to not only mechanically apply a particular method, but also to see why and how this method works relative to another method, and then

to use this insight to construct a *best method* which can be expected to work for as many data sets as possible.

The background for best-method tests is not treated at length within this, since we apply this evaluation through the mean squared error (see chapter 6).

III. **The Numerical Simulation of Artificial Data**

The third and final aspect regarding data is to test the different theoretical descriptions of not only for one specific experiment's data, but for many different data sets. While these may in some cases be available from experiments, it is often necessary to generate such data sets by means of computer based simulations. Simulations can be used as inputs into testing the above third and fourth level of theoretical description.

Our particular implementation is covered in chapter 6.

## 3.2 Physical Data

### 3.2.1 Data Within Context

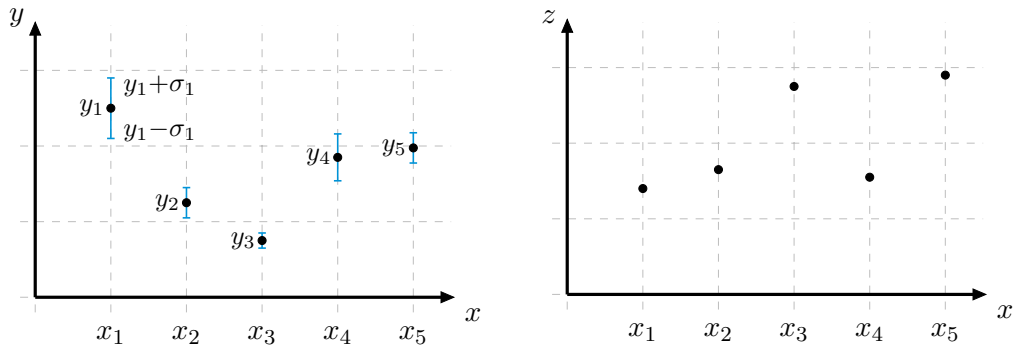
There is a huge variety of physical situations, an equally large variety of experiments and measuring techniques, and a large variety in time and money available for gathering data. Section 3.1 covers them all, but only generically*.

We specialise *only* to such data as previously illustrated, namely,

$$\mathcal{D} := (\boldsymbol{y}, \boldsymbol{\sigma}) = \{y_n, \sigma_n\}_{n=1}^{N}, \tag{3.1}$$

with data points $\boldsymbol{y}$ and experimental uncertainties $\boldsymbol{\sigma}$. Also we *only* look at models using Gaussian likelihoods, and we *only* consider candidate functions which are *linear* in the parameters.

We also consider *only* the case when the $\boldsymbol{\sigma}$ are fixed and known, rather than unknown.




(i) The basic data: $x_n, y_n, \sigma_n$

(ii) Uncertainty-rescaled data $z_n = y_n/\sigma_n$

Figure 3.1: Unrescaled and rescaled data

Figure 3.1 qualitatively illustrates some basic features of the type of data which we consider. On the horizontal axis, we have $N$ points at which measurements or observations were made, $\boldsymbol{x} = \{x_n\}_{n=1}^{N}$. In this example, the five measurement points are equally spaced and consecutive, but that is not a general requirement. These measurement points are taken as fixed and known

* Of course, this text covers only a very small part of this vast field.

throughout the analysis. They are therefore not explicitly included in notation, but considered as part of our state of knowledge*.

The five data points, $y_1, y_2, \ldots, y_5$, are shown as black dots and depending on the circumstances, they can be measured directly or they are the result of processing some "raw data". For example, there may have been 25 measurements $\eta_1, \ldots, \eta_{25}$ made at the point $x_2$, and the final data point is the sample mean $y_2 = \frac{1}{25}\sum_{i=1}^{25} \eta_i$.

The experimental uncertainties, $\sigma_1, \ldots, \sigma_5$ are shown in as blue bars above and below the respective data points $y_n$. We show for the measurement point $x_1$ the values of the bottom and top as $y_1 - \sigma_1$ as well as $y_1 + \sigma_1$.

The intention of the experimental uncertainties is to give the reader an indication of the degree of confidence which the experimentalist has in the value of the measured $y_n$: small $\sigma_n$ indicates an accurate and precise $y_n$, while a large $\sigma_n$ indicates that the value of $y_n$ is not well determined.

The determination of experimental uncertainties is a difficult and complicated topic, because it is supposed to give information not only on statistical fluctuations, but also on systematic uncertainties or systematic errors. Systematic errors or uncertainties arise from many causes, including apparatus bias, calibration, human error and the like. Statistical fluctuations arise because some underlying variable is not controlled or there is a random physical influence such as thermal noise.

We assume that both statistical and systematic errors have been taken into account in compiling the experimental uncertainties†.

Figure 3.1(ii) illustrates what happens when the data $y_n$ is converted into "uncertainty-rescaled data" by means of $z_n = \frac{y_n}{\sigma_n}$: those points $y_n$ with small $\sigma_n$ become larger and vice versa. It is, in fact, possible to do the entire analysis in terms of $\boldsymbol{z}$ instead of $(\boldsymbol{y}, \boldsymbol{\sigma})$ (as seen in section 3.4).

* * *

**Example 3.1** **Systematic and Statistical Errors: The Ten Metre Measure**

In the hard labour industry, when labourers need to make a quick measurement while they do not have their equipment on hand, they make use of clever innovations. For example, they measure 10 metres by walking 13 paces. Given this practice, suppose we know that the following rule or "law" applies:

*10 metres have the same length as 13 steps.*

In practice, we ask people at random to take 13 steps in a straight line from a defined starting point. We measure the length of the 13 steps for each person and compare these measurements to 10 metres. Not surprisingly, different people produce varying lengths with 13 steps.

In order to account for this variation, we are required to pinpoint some source thereof. It could be the result of several variables, but we may be able to eliminate the evidently false ones, like the brand of shoes having any remarkable influence. Alternatively, we

* The choice of $x_n$ is of course important, since this is the subject matter of the field of experimental design which tries to optimise the design matrix introduced in section 3.3. † The experimentalist ensures that these numbers properly reflect the state of uncertainty.

may consider a more relevant influence, such as the height of the individual: taller people take longer steps while shorter people take shorter ones.

With this in mind, we see the systematic source of the variation, i.e. the systematic error, primarily as the result of an individual's natural stride length.

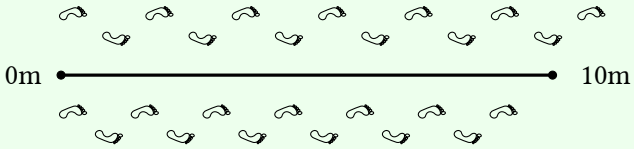
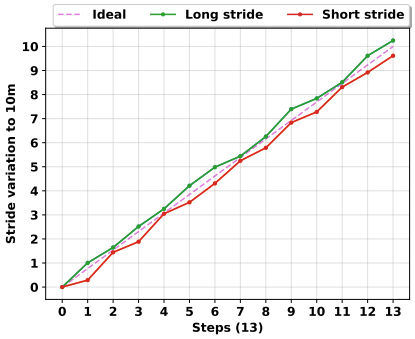
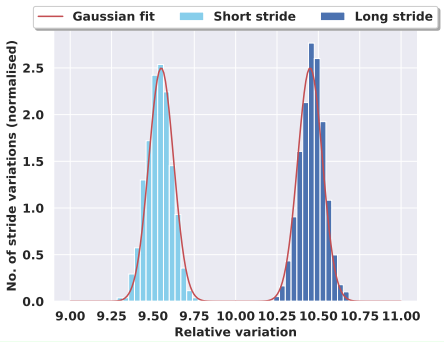



Figure 3.2: Footstep variations indicating how 13 steps may either overestimate or underestimate 10m, depending on the length of steps.

Naturally, we expect varying strides to produce varying lengths; longer strides are expected to exceed the 10/13 rule while shorter strides are expected to fall systematically short of it. Figure 3.3(i) illustrates results obtained from the first two participants. Evidently, the person with the longer stride overshot the 10/13 rule, shown as the straight line, while the person with the shorter stride "fell short".




(i) Systematic error: tall and short individuals' stride lengths.




(ii) Statistical error on top of systematic error.

Figure 3.3: Systematic and statistical error present within the experiment.

In addition to the non-random "stride length" represented by systematic error, every individual person's steps would vary if the experiment was to be repeated. These variations are however random and will tend to be symmetric about some mean.

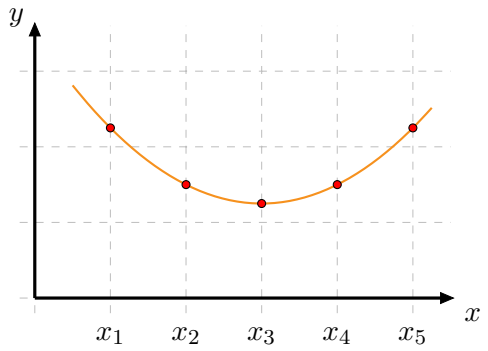
This is illustrated in figure 3.3(ii), where both the long strided and the short strided participants took 13 steps many times and the final distance for each trial over all repetitions was recorded and then histogrammed. This illustrates the statistical error. Figure 3.3(ii) also reflects the fact that such histograms often take the shape of Gaussian (normal) distributions, which may be fitted with suitable Gaussians functions.

### 3.2.2 Data From Additive Systematic Error & Random Variables

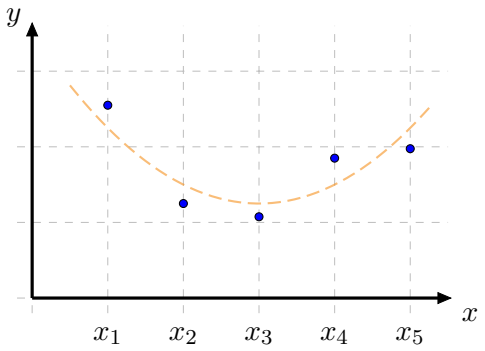
As previously stated, quantifying the experimental uncertainties, $\boldsymbol{\sigma}$, of the data is not a primary concern; these are taken as given. Nevertheless, it is tempting to make a priori physical models which "explain" how the data which we finally observe came about.

At the beginning of this toy-model explanation, there is a rule or "law" for which the phe-

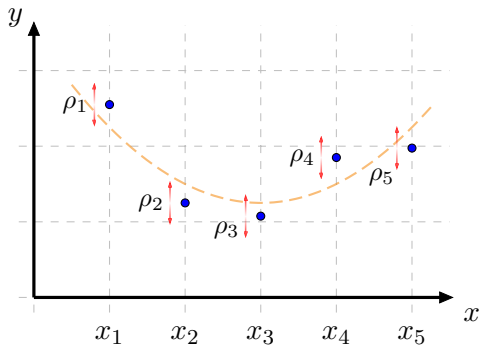
nomenon being investigated follows and it is this rule that we consider as underlying the data.




(i) The underlying law or rule




(ii) Systematic distortion




(iii) Random component

Figure 3.4: Additive generation of data

Figure 3.4(i), this law is represented by the orange curve which we could call the "true" $y(x)$ — but remembering that it is not possible to call any explanation "true". In the trivial case where the law translates directly into data, we would have obtained the five readings shown as red circles on top of the curve, i.e. $y(x_n), n = 1, \ldots, 5$.

In this "explanation", these "true" readings are systematically distorted by numbers $a_n$, one at each $x_n$: as illustrated in figure 3.4(ii), with the (now translated) blue circles. The changes $a_n$ from red to blue circles would always be *the same* if we were to repeat the experiment under the same conditions.

On top of these systematic changes, there is an additional random variation $\rho_n$ which would be *different* for every repetition. This random physical variation often occurs in experimental situations; one example is "thermal noise". The $\rho_n$ would therefore change for every repetition of the experiment; this is indicated schematically by the red arrows in figure 3.4(iii).

Within the context of this additive physical "explanation", the final data is the result of adding the effects of the law, the systematic distortion and the random variation,

$$y_n = y(x_n) + a_n + \rho_n \tag{3.2}$$

We note again that this "explanation" may be plausible and convenient, but in real experimental situations may not be true. Nevertheless, this physical picture of a *true* $y(x)$ changed by *systematic* distortions and *random* additional component, provides a convenient starting point for the *numerical simulation of data*.

## 3.3 Linear Regression

Having dealt with data and its origins, we now turn to the theoretical description, and with regards to our current state of knowledge, we have previously encountered the four levels of the theoretical description of data.

In this section, we make a start in constructing a model, or rather a class of models often called linear regression.

The formulae derived below are standard textbook cases, with the important difference that they form only a part of the picture in a Bayesian context. At present, we are dealing with only the likelihood. A complete model specification can only follow much later once the various priors have been introduced.

The specific form of the likelihood used for linear regression is that of a multivariate Gaussian. The Gaussian comes about as follows, given data in the form $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{\sigma})$, the aim is to use candidate functions (also called parametrisations), $y(x)$, which are constructed from a suitable set of functions $f_k(x)$, with $k = 1, \ldots, K$,

$$y(x \,|\, \boldsymbol{\alpha}) = \sum_{k=1}^{K} f_k(x)\, \alpha_k, \tag{3.3}$$

Note that these are linear in the $K$ free parameters $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^{K}$, not in $x$. Moreover, we refer to the number of free parameters as the *model order*. Figure 3.5 illustrates two of the infinitely many possible candidate functions which can be formed by varying the values of the parameters

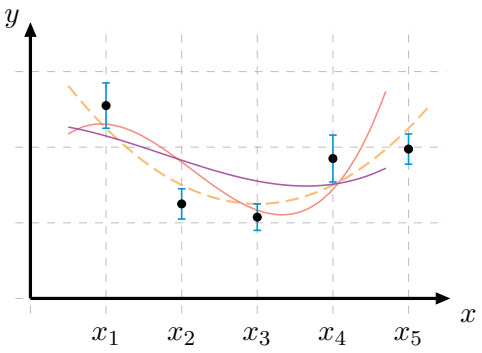



Figure 3.5: Candidate functions attempting to model the data. The true function, represented by the dashed line, is unknown.

The underlying law or true function shown in dashed lines is, of course, unknown within the analysis of data. Ideally, it would be the best case scenario to find the candidate function $y(x)$ which approaches that unknown function as closely as possible, but that aim cannot be achieved directly. All that we have is the data, which as discussed deviates from the underlying law in a way which we cannot reconstruct. The true law cannot be known, even in principle, because the given data could have originated from more than one possible true law.

Note also that each data point $y_n$ contains information which pertains only to the candidate and true functions at the point $x_n$; it tells us nothing about the true law $y(x)$ and note that the candidate functions $y(x \,|\, \boldsymbol{\alpha})$ of equation (3.3) at any other $x \neq x_n$. For any inference, we must therefore refrain from using equation (3.3), and rather only the $N$ samples of the candidate functions given by

$$y(x_n \,|\, \boldsymbol{\alpha}) = \sum_{k=1}^{K} f_k(x_n)\, \alpha_k \qquad \forall\, x_n \in \boldsymbol{x}. \tag{3.4}$$

This agrees with the least squares prescription, namely to minimise the differences between the given data $y_n$ and the candidate functions only at the points $x_n$,

Given particular values of $\boldsymbol{r}$, there are therefore exactly $N$ differences between data and candidate function,

$$\varepsilon_n := y_n - y(x_n \,|\, \boldsymbol{\alpha}) \tag{3.5}$$

which are collectively minimised by

$$Q(\boldsymbol{\varepsilon}) = \frac{1}{N}\sum_{n=1}^{N}\frac{\varepsilon_n^2}{\sigma_n^2} = \frac{1}{N}\sum_{n=1}^{N}\frac{\big(y_n - y(x_n \,|\, \boldsymbol{\alpha})\big)^2}{\sigma_n^2}, \tag{3.6}$$

where $\boldsymbol{\varepsilon} = \{\varepsilon_n\}_{n=1}^N$. This quantity is often called $\chi^2$ or chi-squared criterion in the literature.

As explained before, the minimisation of $Q$ can be understood as the maximisation of the likelihood if we argue that the distribution of differences between data and the trial function at point $x_n$ should be Gaussian,

$$p(\varepsilon_n) = \frac{\exp\big(-\frac{\varepsilon_n^2}{2\sigma_n^2}\big)}{\sqrt{2\pi\sigma_n^2}} \qquad n = 1, \ldots, N, \tag{3.7}$$

and that, for independent $\varepsilon_n$, the likelihood for all the differences $\varepsilon_1, \ldots, \varepsilon_N$ is therefore the product*

$$p(\boldsymbol{\varepsilon}) = p(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N) \overset{\text{ui}}{=} \prod_{n=1}^{N} p(\varepsilon_n) = \frac{\exp\big(-\frac{1}{2}NQ(\boldsymbol{\varepsilon})\big)}{(2\pi)^{N/2}\prod_n \sigma_n} \tag{3.8}$$

With the help of equation (3.5) this is easily transformed into a likelihood for the data $\boldsymbol{y}$

$$p(\boldsymbol{y} \,|\, \boldsymbol{\alpha}) = p(\boldsymbol{\varepsilon}) \overset{\text{ui}}{=} \frac{1}{(2\pi)^{N/2}\prod_n \sigma_n} \exp\left(-\frac{1}{2N}\sum_{n=1}^{N}\frac{\big(y_n - y(x_n)\big)^2}{\sigma_n^2}\right) \tag{3.9}$$

This can be cast in terms of vectors and matrices. Let $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{y}$, and define the $N \times K$ design matrix $\mathbb{V}$ with matrix elements $\mathbb{V}_{nk} := f_k(x_n)$, so that equation (3.3) becomes $\boldsymbol{y} = \mathbb{V}\boldsymbol{\alpha}$. Also let $\boldsymbol{\Sigma}$ be the $N \times N$ diagonal matrix with matrix elements $\boldsymbol{\Sigma}_{nm} = \sigma_n\sigma_m$, i.e. $\sigma_n^2$ on the diagonal†. then the chi-squared criterion can be written as

$$Q(\boldsymbol{y}) = \tfrac{1}{N}(\boldsymbol{y} - \mathbb{V}\boldsymbol{\alpha})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mathbb{V}\boldsymbol{\alpha}) \tag{3.10}$$

and the likelihood becomes

$$p(\boldsymbol{y} \,|\, \boldsymbol{\alpha}) = \frac{\exp\big(-\frac{1}{2}(\boldsymbol{y} - \mathbb{V}\boldsymbol{\alpha})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mathbb{V}\boldsymbol{\alpha})\big)}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} = \frac{\exp\big(-\frac{1}{2}NQ(\boldsymbol{y})\big)}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \tag{3.11}$$

There are now certain aspects which are worthy of note,

1. The Gaussian in Equation (3.7) and the resulting Gaussian forms for the likelihood are a *model choice* and not at all compulsory; one could choose many other functions for the likelihood. The choice of Gaussian has, of course, substantial advantages: firstly it results in the convenient and beautiful linear algebra theory set out in this chapter and thousands of other books and papers. More fundamentally, the central limit theorem implies that most non-Gaussian probabilities eventually converge to a Gaussian under convolution. There are, however, data cases and types which are best described in terms of other functional forms.

* We could write this probability as $p(\boldsymbol{\varepsilon} \,|\, \boldsymbol{\sigma})$, but since $\boldsymbol{\sigma}$ here and its generalisation $\boldsymbol{\Sigma}$ remains constant throughout, we omit it from the notation. † This general form may be reduced to the previous form by considering the matrix elements to be such that $\boldsymbol{\Sigma}_{nm} = \sigma_n\sigma_m\delta_{nm}$.

A deeper explanation of this is that knowledge of some data only in the form of the data mean and data variance, without additional information, leads to a Gaussian likelihood by maximisation of entropy, without any claim or need that the underlying data must follow a Gaussian (Jaynes 2003, pp. 208–210).

2. The definition of $\varepsilon_n$ as the difference between the data $y_n$ and the candidate function at that point $y(x_n \,|\, \boldsymbol{\alpha})$ resembles the variable $\rho_n = y_n - y(x_n) - a_n$ of equation (3.2), and consequently the sum $y(x_n) + a_n$ (true function plus systematic changes $a_n$) is also often confused with the candidate model $y(x_n \,|\, \boldsymbol{\alpha})$. However, $\varepsilon_n$ and $\rho_n$ are conceptually far apart: $\rho_n$ is a random variable designed to mimick the randomness of measurement in experimental situations, while $\varepsilon_n$ is a variable within model building.
3. As already stated in section 3.1, a third kind of Gaussian with a third set of random variables is introduced in the course of data simulation. While one of these may be used to motivate another, there is no principle or theorem which requires the physical data, the model building and the data simulation to have the same mathematical structure.
4. It is of interest to note that the candidate model takes on the role of the mean in the Gaussian, thus, when finding the optimum candidate model, we are optimising with respect to the mean.

## 3.4 Minimum Chi-squared & Maximum Likelihood

As already outlined in the introduction, the derivation so far can be understood entirely within the traditional least squares theory or alternatively that of the maximum likelihood, because minimisation of $Q$ is equivalent to maximisation of the likelihood through equation (3.11).

Given the analytical results, minimising the chi-squared criterion with respect to the $K$ parameters of $\boldsymbol{\alpha}$ is a simple matter. We first consider the algebraic first and second derivatives, and then quote without derivation the vector-matrix form of the same derivation.

To find the values of the parameters $\widehat{\boldsymbol{\alpha}}$ at which $Q$ is an extremum, the derivative with respect to each parameter $\alpha_i$ must be zero,

$$0 \stackrel{!}{=} \frac{\partial}{\partial \alpha_i} N Q(\boldsymbol{\alpha}) \bigg|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} = 2 \sum_{n,m=1}^{N} \left( y_n - \sum_{k=1}^{K} \mathbb{V}_{nk}\, \alpha_k \right) (\boldsymbol{\Sigma}^{-1})_{nm} (-\mathbb{V}_{mi}) \tag{3.12}$$

which results in the vector-matrix equation

$$0 \stackrel{!}{=} 2(\mathbb{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbb{V} \widehat{\boldsymbol{\alpha}} - \mathbb{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}) \tag{3.13}$$

so that the minimum-$Q$ (maximum likelihood) parameter values are given by

$$\widehat{\boldsymbol{\alpha}} = (\mathbb{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbb{V})^{-1}\, \mathbb{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y} \tag{3.14}$$

and the second derivative is

$$\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} N Q(\boldsymbol{\alpha}) \bigg|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} = 2\, (\mathbb{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbb{V})_{ij} \tag{3.15}$$

If consider the case in which no covariances are present, such as with Identically and Independently Distributed (IID) data, and the covariance matrix reduces to the diagonal matrix of $\sigma_n^2$, then we may consider this simplification as leading to the use of the rescaled data,

$$z_n = \left\{ \frac{y_n}{\sigma_n} \right\}_{n=1}^{N} \tag{3.16}$$

and on rescaling our basis functions also,

$$g_k(x_n) = \frac{f_k(x_n)}{\sigma_n} \tag{3.17}$$

we thus obtain a rescaled design matrix, which we shall denote by $\mathbb{X}$. The chi-squared criterion then simplifies to

$$\begin{aligned} NQ(\boldsymbol{\alpha}) &= \boldsymbol{\chi}^2 \\ &= (\boldsymbol{z} - \mathbb{X}\boldsymbol{\alpha})^{\mathsf{T}} (\boldsymbol{z} - \mathbb{X}\boldsymbol{\alpha}) \\ &= (\boldsymbol{z} - \boldsymbol{\zeta})^{\mathsf{T}} (\boldsymbol{z} - \boldsymbol{\zeta}). \end{aligned} \tag{3.18}$$

Here we define $\boldsymbol{\zeta} := \mathbb{X}\boldsymbol{\alpha}$ as the vector representing the model, and the maximum-likelihood parameter vector of equation (3.14) becomes the well-known Moore-Penrose pseudo-inverse,

$$\widehat{\boldsymbol{\alpha}} = (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\,\mathbb{X}^{\mathsf{T}}\boldsymbol{z}. \tag{3.19}$$

In the general case, this result is valid also for non-IID covariance matrices $\boldsymbol{\Sigma}$ since it is positive semi-definite and therefore its inverse has a Cholesky decomposition*: $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^{\mathsf{T}}\boldsymbol{L}$ and the quantities derived in equations (3.14) and (3.15) may be defined in a fashion near identical to equation (3.16) to yield

$$(\mathbb{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbb{V})^{-1}\,\mathbb{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \big((\boldsymbol{L}\mathbb{V})^{\mathsf{T}}(\boldsymbol{L}\mathbb{V})\big)^{-1}(\boldsymbol{L}\mathbb{V})^{\mathsf{T}}(\boldsymbol{L}\boldsymbol{y}) \tag{3.20}$$

which is identical to equation (3.19) on setting $\boldsymbol{L}\mathbb{V} = \mathbb{X}$ and $\boldsymbol{L}\boldsymbol{y} = \boldsymbol{z}$. Given our assumptions and from this pseudo-inverse we learn that we need only the data and the design matrix to determine the most suitable candidate model. The $K \times K$ matrix of the second derivatives in equation (3.15) is known as the Hessian. We prefer to define our Hessian matrix in a slightly modified way,

$$\mathbb{H} := \frac{\mathbb{X}^{\mathsf{T}}\mathbb{X}}{N} \qquad\qquad \mathbb{H}_{ij} = \frac{1}{2N}\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}Q(\boldsymbol{\alpha})\bigg|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}}. \tag{3.21}$$

Noting that higher-order derivatives with respect to $\boldsymbol{\alpha}$ are zero, and expanding $Q$ in a Taylor series about $\widehat{\boldsymbol{\alpha}}$ provides us with the following, quadratic form

$$Q(\boldsymbol{\alpha}) = Q(\widehat{\boldsymbol{\alpha}}) + (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})^{\mathsf{T}}\,\mathbb{H}\,(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}). \tag{3.22}$$

The first term in the expansion is the minimum $\chi^2$ criterion, while the second term offers us grounds for exploration.

By rescaling the data, the likelihood of equation (3.11) can be written in terms of the chi-squared criterion as expressed in equation (3.22), which results in

$$p(\boldsymbol{z}\,|\,\boldsymbol{\alpha}) = \frac{\exp\left(-\frac{1}{2}NQ(\boldsymbol{\alpha})\right)}{(2\pi)^{N/2}}. \tag{3.23}$$

* The Cholesky decomposition is the decomposition of a Hermitian positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose.

## 3.5 Geometric & Linear Algebraic Perspective

We shall now consider an alternative geometric approach to our current analysis, while still bearing everything up to this point in mind. Suppose we represent the data as a vector, $\boldsymbol{z}$ in an $N$-dimensional "data space", and we desire to approximate this vector by defining any vector, $\boldsymbol{w}$, from the vector space $W$, the "model space", which contains all possible candidate function vectors, $\boldsymbol{w} \in \big(\zeta(x_1|\boldsymbol{\alpha}), \zeta(x_2|\boldsymbol{\alpha}), \ldots, \zeta(x_N|\boldsymbol{\alpha})\big)$. The model is an approximation, and so the model will be a subspace of the data space, and any approximation will result in an error vector of the form $\boldsymbol{\varepsilon} = \boldsymbol{z} - \boldsymbol{w}$. Any possible candidate function vector in $W$ will result in a *non-zero* error vector, unless $\boldsymbol{z}$ is itself contained in the vector space — which implies that if the data contains unknowns, then the error cannot be zero. It is then of interest to know *which* $\boldsymbol{w}$ best approximates the response vector $\boldsymbol{z}$. Naturally, the best approximation would be the one which has a *minimum length* in $\boldsymbol{\varepsilon}$ (i.e. the shortest norm $\|\boldsymbol{\varepsilon}\|$), meaning that the lesser the distance, the better the approximation.

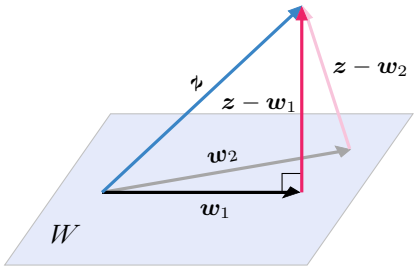



Figure 3.6: Construction of vectors in $W$ which approximate the vector $\boldsymbol{z}$.

In addition, this also means that the distance between $\boldsymbol{z}$ and the vector space $W$ is being minimised — we want $\boldsymbol{z}$ to be as close to $W$ as possible (Anton and Rorres 2005, p. 332). Among all vectors in $W$, suppose there exists at least one vector, $\widehat{\boldsymbol{w}}$ such that

$$\|\boldsymbol{z} - \widehat{\boldsymbol{w}}\| < \|\boldsymbol{z} - \boldsymbol{w}\|$$

for all other vectors $\boldsymbol{w}$ in $W$. Then, we may explore this using the Pythagorean theorem and linearity in the inner product,

$$\|\boldsymbol{z}\|^2 = \|\boldsymbol{w}\|^2 + \|\boldsymbol{\varepsilon}\|^2 + 2\langle \boldsymbol{w}, \boldsymbol{\varepsilon} \rangle \tag{3.24}$$

From this we learn that the vector $\widehat{\boldsymbol{w}}$ minimising this error vector is the one in which the inner product between the cross terms is smallest, i.e. when these vectors are orthogonal. Therefore, the vector in $W$ which best approximates the response vector, is the *orthogonal projection* of $\boldsymbol{z}$ onto $W$,

$$\widehat{\boldsymbol{w}} = \text{proj}_W \, \boldsymbol{z} \tag{3.25}$$

i.e. the best approximation to $\boldsymbol{z}$ is given by $\boldsymbol{z} - \text{proj}_W \, \boldsymbol{z}$. To find an expression for the orthogonal projection for our purposes, we take a brief excursion into the equivalent *additive noise model*. In general, when linear systems of equations of the form $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$, are indeterminate*, then there exists no exact solution to the system and a similar principle applies. In this system, $\mathbb{A}$ is an $N \times K$ dimensional matrix, $\boldsymbol{x}$ is a $K$-dimensional vector, and $\boldsymbol{b}$ is an $N$-dimensional vector.

With indeterminate systems, the quantity $\|\boldsymbol{b} - \mathbb{A}\boldsymbol{x}\|$ may (once again) be viewed as the error, meaning that $\boldsymbol{b}$ is approximated, with $\boldsymbol{x}$ serving as the approximate solution. The vector space $W$

* An indeterminate system is one for which more than one solution exists.

may be considered to be the column space of $\mathbb{A}$, for which the product $\mathbb{A}\boldsymbol{x}$ is a linear combination of the column vectors of $\mathbb{A}$, for each vector $\boldsymbol{x}$. This means that $\mathbb{A}\boldsymbol{x}$ varies over the entire column space $W$ (Anton and Rorres 2005, p. 333).

Minimising the error means that $\boldsymbol{b}$ is projected onto the column space, and the orthogonal projection would yield the best approximation: $\widehat{\boldsymbol{x}} = \text{proj}_W \boldsymbol{b}$, where $\|\mathbb{A}\widehat{\boldsymbol{x}} - \boldsymbol{b}\|$ is a minimum. Since the minimum error vector is perpendicular to the column space of $\mathbb{A}$, it implies that the quantity $\boldsymbol{b} - \mathbb{A}\widehat{\boldsymbol{x}}$ lies in the nullspace of $\mathbb{A}^\mathsf{T}$ (Anton and Rorres 2005, p. 334), and as a result of othogonality it then follows,

$$\begin{aligned} \mathbb{A}^\mathsf{T}(\boldsymbol{b} - \mathbb{A}\widehat{\boldsymbol{x}}) &= 0 \\ \widehat{\boldsymbol{x}} &= (\mathbb{A}^\mathsf{T}\mathbb{A})^{-1}\mathbb{A}^\mathsf{T}\boldsymbol{b} \end{aligned} \tag{3.26}$$

The equations associated with this solution are known as *normal equations* and this solution is known as the least squares solution. Moreover, we note that this is the same solution that we obtained in equation (3.19). If the columns of $\mathbb{A}$ are linearly independent, then the linear system has a *unique* least squares solution (Anton and Rorres 2005, p. 335). In this setting, the orthogonal projection of $\boldsymbol{b}$ on $W$ is given by,

$$\text{proj}_W \boldsymbol{b} = \mathbb{A}\widehat{\boldsymbol{x}} = \mathbb{A}(\mathbb{A}^\mathsf{T}\mathbb{A})^{-1}\mathbb{A}^\mathsf{T}\boldsymbol{b} \tag{3.27}$$

We may refer to the matrix $\mathbb{P} = \mathbb{A}(\mathbb{A}^\mathsf{T}\mathbb{A})^{-1}\mathbb{A}^\mathsf{T}$ as the *projection matrix* or simply the *projector*. Translating this into our geometric analysis, we identify $\boldsymbol{b} = \boldsymbol{z}$, $\mathbb{A} = \mathbb{X}$ and the projector as

$$\mathbb{P} := \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T} \tag{3.28}$$

in terms of which the minimum chi-squared criterion simplifies to

$$NQ(\widehat{\boldsymbol{\alpha}}) = \boldsymbol{z}^\mathsf{T}\boldsymbol{z} - \boldsymbol{z}^\mathsf{T}\mathbb{P}\boldsymbol{z} \tag{3.29}$$

Now we may collate our knowledge from the formalisms of linear regression and least squares, and refine our view on modelling data. Firstly, we note that the $N \times K$ design matrix $\mathbb{X}$ must always contain more observations than parameters. This upper limit exists, because if it so happens that $K > N$, then the system has infinitely many solutions for the linear system. Hence we must have $K \leq N$ and there is an upper limit on the number of parameters. In all cases in which $K < N$, the system is indeterminate and the maximum likelihood estimate (MLE), which is also the least squares solution, becomes the best approximation.

Secondly, the data ($\boldsymbol{z}$), the noise ($\boldsymbol{\varepsilon}$), and the model vector ($\boldsymbol{\zeta}$ constructed from the candidate functions) are all defined from their respective vector spaces, which we refer to as *data space*, *noise space*, and *model space*. The space of $\boldsymbol{\alpha}$ we refer to as *parameter space*.

We note that minimising the projection error and minimising $\boldsymbol{\chi}^2$ (or $NQ$) is the same procedure and yields the same results. We conclude that the error may be expressed simply as $\boldsymbol{\chi}$ and the optimisation procedure of minimising $\boldsymbol{\chi}^2$ is equivalent to minimising the least squares solution. Geometrically, the minimum $\boldsymbol{\chi}^2$ solution is the orthogonal projection of the data vector onto parameter space.

## 3.6 Transformation to a Symmetric Likelihood

The Hessian, as in equation (3.21), is positive semi-definite and symmetric, therefore it may be diagonalised via eigen decomposition,

$$\mathbb{H}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j \qquad j = 1, \ldots, K. \tag{3.30}$$

Since the Hessian is symmetric, it's eigenvectors are orthogonal and we may always construct an orthonormal basis therefrom, therefore we have that $\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_j = \delta_{ij}$, which implies that meaning that $\mathbb{S}^{-1} = \mathbb{S}^{\mathsf{T}}$ and $\det \mathbb{S} = 1$. Therefore $\mathbb{H}$ can be diagonalised by means of

$$\mathbb{H} = \mathbb{S}^{-1}\boldsymbol{\Lambda}\,\mathbb{S} \tag{3.31}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbb{H}$. The parameters in quadratic form thus undergo a transformation,

$$Q(\boldsymbol{\alpha}) = Q(\widehat{\boldsymbol{\alpha}}) + \left(\boldsymbol{\Lambda}^{1/2}\,\mathbb{S}\,(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})\right)^{\mathsf{T}}\left(\boldsymbol{\Lambda}^{1/2}\,\mathbb{S}\,(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})\right). \tag{3.32}$$

Defining transformed parameters

$$\boldsymbol{\beta} = \boldsymbol{\Lambda}^{1/2}\,\mathbb{S}\boldsymbol{\alpha} \tag{3.33}$$

Thus the likelihood of equation (3.23) becomes,

$$p(\boldsymbol{z}\,|\,\boldsymbol{\beta}) = \frac{\exp\left(-\frac{1}{2}NQ(\boldsymbol{\beta})\right)}{(2\pi)^{N/2}} \tag{3.34}$$

where

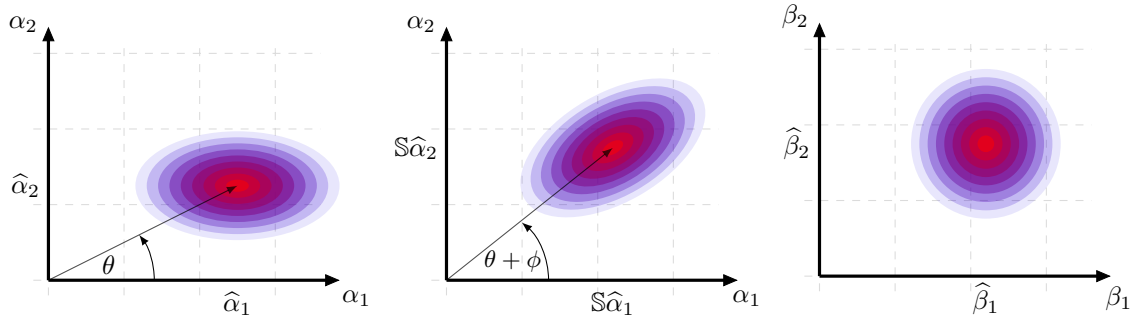
$$Q(\boldsymbol{\beta}) = Q(\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\mathsf{T}}\,(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \tag{3.35}$$

$$\widehat{\boldsymbol{\chi}}^2 = NQ(\widehat{\boldsymbol{\beta}}) = \langle \boldsymbol{z}^2 \rangle - \widehat{\boldsymbol{\beta}}^2 \tag{3.36}$$

with $\widehat{\boldsymbol{\beta}} = \boldsymbol{\Lambda}^{1/2}\,\mathbb{S}\widehat{\boldsymbol{\alpha}}$ and where we introduce shorthand notations

$$\langle \boldsymbol{z}^2 \rangle = \tfrac{1}{N}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{z} = \frac{1}{N}\sum_{n=1}^{N} z_n^2 \qquad \text{and} \qquad \widehat{\boldsymbol{\beta}}^2 = \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\widehat{\boldsymbol{\beta}} = \sum_{k=1}^{K} \widehat{\beta}_k^2.$$

There is a geometric meaning behind the transformation from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ and the effect may be visualised through the likelihood. In Figure 3.7(iii), we show an example of the transformation of a likelihood with two parameters, $(\alpha_1, \alpha_2) \to (\beta_1, \beta_2)$. Firstly, the orthonormal eigenvector matrix, $\mathbb{S}$, acts as a rotation matrix on the parameters, i.e. the parameters undergo a rotation.




(i) Initial likelihood (ii) Eigenvector matrix rotation (iii) Symmetrised likelihood

Figure 3.7: Rotated and rescaled contour plots of the likelihood for $K$=2

Secondly the square root of the eigenvalues rescale the parameters, resulting in a spherically symmetric form for $Q$, with the sphere centered at $(\widehat{\beta}_1, \widehat{\beta}_2)$, the MLEs of the transformed parameters. The example is trivially extended to $K$ parameters: the transformed likelihood will have contours which are spherically symmetric on the $K$-sphere (hypersphere in $K$ dimensions) with the centre of the $K$-sphere located at $\widehat{\boldsymbol{\beta}}$. This symmetry reflects the rotation and rescaling of the parameter space or, statistically speaking, the removal of covariances and rescaling of all variances to one. A spherically symmetric Gaussian distribution has spherical contours of equal probability centred on the point of maximum likelihood.

* * *

**Example 3.2** **Sally's Seashells**

Sally said she sells seashells by the seashore. She usually collects the most beautiful seashells after high tide. In order to assure herself about the ideal times to collect, she conducted a small experiment: everyday, at the start of each hour, between 07:00 and 17:00, she measured the water level at the pier to determine how the tide varies.

After one month, she took the mean value of her readings for each time measured, afterwhich she managed to produce the following data,

| 07:00 | 08:00 | 09:00 | 10:00 | 11:00 | 12:00 | 13:00 | 14:00 | 15:00 | 16:00 | 17:00 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.121 | 0.885 | 1.456 | 1.220 | 0.375 | 0.087 | -0.726 | -0.107 | 0.268 | 0.794 | 1.489 |

Given that there is no reference to a mid-point between high tide and low tide, she considered her *first* reading to be a point of origin. Note that repeated measurements and the subsequent averaging process accounted for a band of variation. Her goal is to now model this data, and under observation she notices that there is a visible *oscillatory trend*, as can be seen in figure 3.8 — meaning that the "true law" may be oscillatory. She decides to consider two possible models: *i.* a polynomial model and *ii.* a sinusoidal model. With either of these models, the primary assumption is that the data is modelled as a *linear system*, therefore the principles of linear regression/least squares are applicable. Provided with this specification, Sally denotes the polynomial model as $v$, while she denotes the sinusoidal model as $w$, thus the candidate models are described by

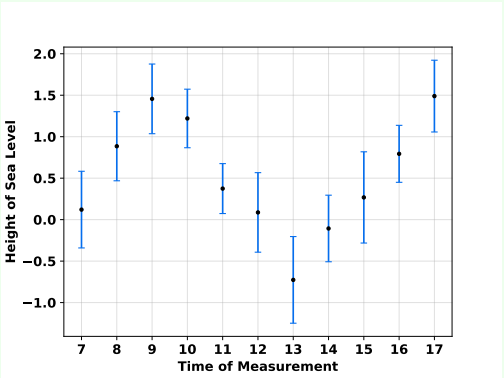
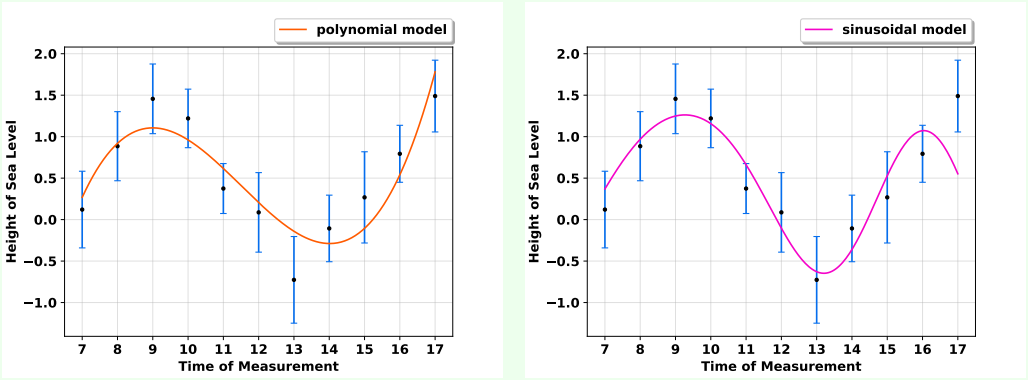



Figure 3.8: Sally's data after one month.

$$
\begin{aligned}
v_n(x_n \,|\, \boldsymbol{\alpha}) &= \sum_{k=1}^{K} \alpha_k \, x_n^{k-1} \\
w_n(x_n \,|\, \boldsymbol{\alpha}) &= \frac{1}{\sqrt{2}} \sum_{\ell=1}^{L} \alpha_\ell \, \sin\left(\frac{\pi(2n+1)(\ell+1)}{2N}\right)
\end{aligned}
\tag{3.37}
$$

Observe that the order of each model need not be the same; it is not necessary for $K = L$, even though the same data is being modelled. Moreover, the sinusoidal model is defined to be orthogonal, since this offers greater stability (among other things compared to a standard sine representation). Naturally, Sally opts for a fourth order (i.e. cubic) polynomial, while tuning the sine model (by eye).

The best parameters are determined via the pseudo-inverse given by equations (3.19) and (3.26), and the candidate models consequently model the data as




(i) Polynomial candidate model of order $K = 4$. (ii) Sinusoidal candidate model of order $L = 5$.

Figure 3.9: Sally's candidate models

Evidently, both candidate models capture the oscillatory behaviour of the data and provide a seemingly suitable description thereof, even though each model is of a different order. Having modelled the data, Sally is now faced with a new problem: which candidate model is the best model?

Note that the fundamental assumptions are about the data (Gaussian noise, linear system, etc.) and any knowledge about the underlying "true" law is drawn therefrom. While candidate models, such as those in figure 3.9, may provide adequate descriptions and predictions of data, there is *no* fundamental knowledge gained in knowing anything about this law.

Furthermore, various different models may provide equally accurate descriptions of data, but we have two new concerns, firstly the model order — since higher model order may provide better predictions*, but this introduces the misconception that we are more interested in the data than we are in the underlying phenomenon.

The second concern is model quality; we may provide many unique model descriptions, but we are currently at a loss as to which model models the underlying phenomenon or "true" law best.

* This may be clear when thinking of the polynomial model: linear versus quadratic versus cubic.

CHAPTER 4

# Model Comparison

## 4.1 Model Complexity

Section 1.4.1 briefly described the necessity for economy when modelling data, since the least squares method would always prefer the model with more parameters. Now we shall provide insights into this issue on a more fundamental level, but first consider it informally.

The situation is that we have one particular data set $\mathcal{D}$, with specific numbers, but that a repetition of the same experiment under the same conditions would yield slightly different data on every repetition. The best explanation even for our present data, $\mathcal{D}$, must therefore be probabilistic. The question is then: *which probability density best explains the data?*

There will almost always be more than one possible model which may describe the data adequately, however, the question becomes how does one strike a balance between choosing a simple model over a more complex model, one with more and one with fewer parameters? Furthermore, how helpful will a simple and a more complex model be in predicting future data?

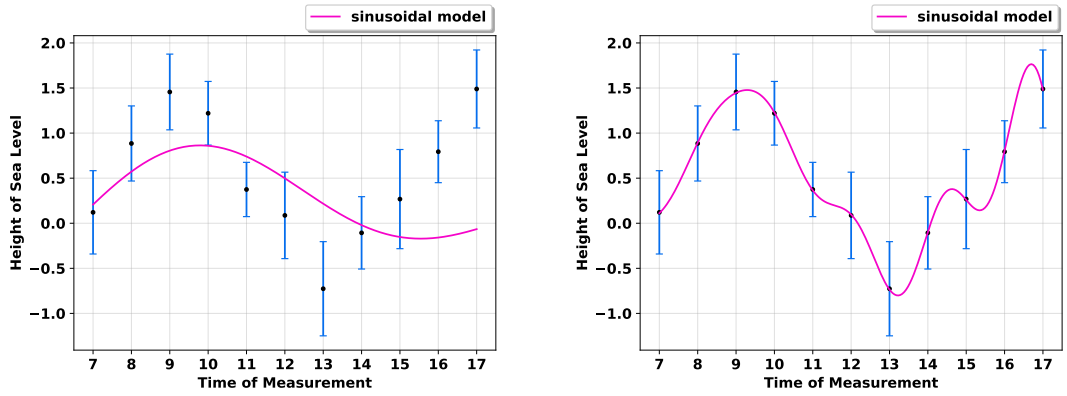
We now face the the first of the two raised concerns, in which the predictive power of a model may increase with complexity. When do we prefer a model of order $K$ or another of $K+1$ which provide similar predictions of data?

One answer to this question is, of course, tied to our state of knowledge. Preference of a more complex model may imply knowledge of the context of the data beyond what is merely observed. Therefore a preference of simpler models (lower order) over complex models (higher order) which make predictions of similar accuracy is rational to our reasoning.

As a matter of fact, *Occam's razor* is often considered a guiding principle of this choice, stating that *complexity should not come without necessity* (McElreath 2015, p. 165). This is commonly rephrased as *a simpler hypothesis may often be the more suitable hypothesis*. Here complexity does not imply anything more than what the simplest case provides, on the contrary, it seems as though it is apparent that one's state of knowledge would imply that one knows more than what is merely observed here in order to justify cases such as those.

We may now approach the second of the two raised concerns, which regards predictive quality, but recall that this is with relation to the unknown "true" law and not merely to the data. This draws a fine line between what is known as *overfitting* and *underfitting* data. Overfitting is an occurrence when the candidate model can reproduce the data with minimal or no discrepancy, which may lead to poor prediction since the model acquired too much from the data, thus loses predictive quality in relation to the law or underlying phenomenon which we aim to model. Underfitting is an occurrence when the candidate model models the data with too little accuracy, thus also losing predictive quality in relation to the the underlying phenomenon.

Figure 4.1 revisits Sally's data (example 3.2) with an illustration of what it means to overfit or underfit the data with respect to model order. The overfitted model introduces a model with too many parameters (high order) while the underfitted model introduces a model with too few parameters (low order).




(i) Underfitting data ($L = 2$) (ii) Overfitting data ($L = 10$)

Figure 4.1: Sally's data revisited.

The question remains as to which model models the data best. Overfitting acquires too much information specific to the current sample being modelled; simply stated, the model becomes sensitive to the purely random fluctuations in the data over and above the desirable sensitivity to the underlying law.

Put differently, A model which overfits a sample is sensitive to irrelevant components in the data and may lead to worse predictions for any future data samples (McElreath 2015, p. 168). When underfitting, on the other hand, the model acquires too little information from the data at hand, is thereby insensitive to the details conveyed by that data, and will therefore also make poor predictions (McElreath 2015, p. 173).

## 4.2 Bayesian Model Comparison in General

At this point, we are aware that the goal is to compare both general hypotheses as well as more specific models which have different degrees of complexity. But how, exactly, does one find the model which best represents the data?

The Bayesian answer to this question is simple but far-reaching. Recall that in equation (2.30), the general form of Bayes' rule, we had already seen that, for any logical propositions $x, y$ and for any background knowledge $\kappa$ whatsoever, one can always express the probability for $x$ given $y$ in terms of the inverse probability of $y$ given $x$,

$$p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)}. \tag{4.1}$$

Noting that probabilities are always conditioned on $\kappa$ (even if not explicit). Given some data $\mathcal{D}$, assume that there are exactly $M$ hypotheses or models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M$ which we consider as possible explanations. For any one model $\mathcal{M}_m$ let $x$ be the proposition "$\mathcal{M}_m$ is true" and $y$ stand for "the data is $\mathcal{D}$". Then equation (4.1) provides an expression for the probability of $\mathcal{M}_m$ being true given $\mathcal{D}$,

$$p(\mathcal{M}_m \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M}_m)\, p(\mathcal{M}_m)}{p(\mathcal{D})}, \tag{4.2}$$

Firstly, the *model prior* is the probability that $\mathcal{M}_m$ is true, given *only* our state of knowledge $\kappa$, but no data. Like all priors, the model prior must be assigned by us based on our knowledge, information and judgement. By our own judgement, one of them must be true so $\sum_{m=1}^{M} p(\mathcal{M}_m) = 1$. If we have no a priori reason to prefer one model over another, we can set all the model priors equal, $p(\mathcal{M}_m) = \frac{1}{M}$ for all $m$. If we do have pertinent reason to consider a particular $\mathcal{M}_a$ more credible than others, we could assign $p(\mathcal{M}_a) > \frac{1}{M}$, so that freedom remains available if we need it.

Secondly, the *model likelihood* is considered to be *our* model of the data and provides a representation of the data *if* the data were to be modelled according to $\mathcal{M}_m$.

Thirdly, the *evidence* can be expressed in terms of the probabilities appearing in the numerator,

$$p(\mathcal{D}) = \sum_{\mu=1}^{M} p(\mathcal{D} \,|\, \mathcal{M}_\mu)\, p(\mathcal{M}_\mu) \tag{4.3}$$

so that the posterior becomes

$$p(\mathcal{M}_m \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \mathcal{M}_m)\, p(\mathcal{M}_m)}{\sum_{\mu=1}^{M} p(\mathcal{D} \,|\, \mathcal{M}_\mu)\, p(\mathcal{M}_\mu)} \tag{4.4}$$

in other words, the probability of a model being correct can be calculated completely once the evidence for all the models has been calculated.

## 4.3 Model Odds & Reference Models

While Equation (4.2) is correct and complete, there is a more convenient formulation in the form of odds. Of the $M$ evidences for models $\mathcal{M}_m$, with $m = 1, \ldots, M$, which have been calculated, we choose one as the reference or base model. Consider this model to be $\mathcal{M}_b$, then we may determine the odds of all $m = 1, \ldots, M$ model posteriors by using equation (4.2) with respect to this base model,

$$\frac{p(\mathcal{M}_m \,|\, \mathcal{D})}{p(\mathcal{M}_b \,|\, \mathcal{D})} = \frac{p(\mathcal{D} \,|\, \mathcal{M}_m)\, p(\mathcal{M}_m)}{p(\mathcal{D})} \cdot \frac{p(\mathcal{D})}{p(\mathcal{D} \,|\, \mathcal{M}_b)\, p(\mathcal{M}_b)} \tag{4.5}$$

$$= \frac{p(\mathcal{D} \,|\, \mathcal{M}_m)}{p(\mathcal{D} \,|\, \mathcal{M}_b)} \cdot \frac{p(\mathcal{M}_m)}{p(\mathcal{M}_b)} \tag{4.6}$$

The so-called *Bayes factor* is defined as the ratio of evidences,

$$\mathcal{B}_{m,b} = \frac{p(\mathcal{D} \,|\, \mathcal{M}_m)}{p(\mathcal{D} \,|\, \mathcal{M}_b)} \tag{4.7}$$

and so we can write the model posterior odds $\frac{p(\mathcal{M}_m \,|\, \mathcal{D})}{p(\mathcal{M}_b \,|\, \mathcal{D})}$ in terms of the Bayes factor and the prior odds $\frac{p(\mathcal{M}_m)}{p(\mathcal{M}_b)}$. Since we have no reason to prefer any of our models, the prior odds equals unity, hence the model posterior odds is equal to the Bayes factor (evidence odds).

In our case, much of the work in calculating evidences can be done analytically. Taking ratios of evidences cancels all common prefactors, which makes the resulting formulas compact and easier to calculate.

Since Bayes factors can easily become very large or very small, it makes sense to find and quote logarithms thereof. The procedure thus becomes that of choosing one reference model, then to find a Bayes factor for every other model with respect to $\mathcal{M}_b$, and then to compare

$$\mathcal{E}(m,b) \equiv \log \mathcal{B}_{m,b} = \log \frac{p(\mathcal{D} \,|\, \mathcal{M}_m)}{p(\mathcal{D} \,|\, \mathcal{M}_b)}. \tag{4.8}$$

The model with the largest $\mathcal{E}(m,b)$ is then the "best model".

## 4.4 The Evidence as Occam's Razor

As we have just seen, the evidence plays a central role in model comparison. Before discussing it, we must clarify the use of the word *evidence*. Strictly speaking, we should call $p(\mathcal{D} \,|\, \mathcal{M}_m)$ the *model likelihood* in the context of section 4.2. We use the term evidence because it is appropriate in the context of parameter inference, as detailed in section 4.7.

Returning to the main issue: the evidence is the probability for $\mathcal{D}$ for all possible values $\Omega(\boldsymbol{\alpha})$ of the parameters $\boldsymbol{\alpha}$ of the model $\mathcal{M}_m$ and therefore to the integral,

$$p(\mathcal{D} \,|\, \mathcal{M}_m) = \int_{\Omega(\boldsymbol{\alpha})} \mathrm{d}\boldsymbol{\alpha}\, p(\mathcal{D}, \boldsymbol{\alpha} \,|\, \mathcal{M}_m) = \int_{\Omega(\boldsymbol{\alpha})} \mathrm{d}\boldsymbol{\alpha}\, p(\mathcal{D} \,|\, \boldsymbol{\alpha}, \mathcal{M}_m)\, p(\boldsymbol{\alpha} \,|\, \mathcal{M}_m). \tag{4.9}$$

We recognise $p(\mathcal{D} \,|\, \boldsymbol{\alpha}, \mathcal{M}_m)$ as the likelihood of the data* which was the subject of chapter 3, while $p(\boldsymbol{\alpha} \,|\, \mathcal{M}_m)$ is the *parameter prior* which will be treated in detail in the following chapter.

Finding the evidence therefore requires computation of the $K$-dimensional integral of equation (4.9). This raises the pitfall of cost of computing and accuracy of numerical results, since every additional parameter requires an additional integral, and since $K$ can easily range into the hundreds or even millions, the numerical challenges become significant. For this reason, analytical answers of the kind found within this text can be very valuable.

We now briefly explain in a qualitative way why computation of the evidence automatically incorporates a penalty for larger $K$ as already discussed in chapter 1 and section 4.1.† Consider two models, $\mathcal{M}_J$ and $\mathcal{M}_K$, with $J$ and $K$ parameters respectively, and with $J < K$. Each model attempts to model possible data with evidences illustrated in figure 4.2. With more parameters at its disposal, $\mathcal{M}_K$ can describe more kinds and more complex data than $\mathcal{M}_J$, so the (qualitative) blue curve for $\mathcal{M}_K$ ranges over a larger set of possible data; in the figure, it is "wider" than the red curve for $\mathcal{M}_J$. Since both evidences must, of course, be normalised to unity, the blue-curve evidence for $\mathcal{M}_K$ must necessarily be mostly (but not always) smaller than that of the red-curve evidence for $\mathcal{M}_J$.

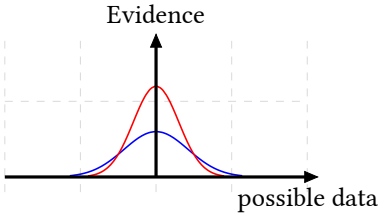



Figure 4.2: Qualitative behaviour of the evidences of $\mathcal{M}_J$ and $\mathcal{M}_K$ as functions of possible data.

* Note that the likelihood merely attempts to *model* the data. The data is only ever processed and analysed, it is *never* modified; tampering with the data is *forbidden*. Data is considered to be immutable! Strictly speaking, we should invent a "data variable" $D$, and the likelihood would be written as $p(D{=}\mathcal{D} \,|\, \boldsymbol{\alpha}, \mathcal{M}_m)$ † Consult Mackay (2003, §28) for a longer, more complete discourse.

Once the actual data $\mathcal{D}$ is measured, it would appear as one particular point on the horizontal axis of figure 4.2. If that actual data falls in the "central" area where the red curve exceeds the blue one, the simpler model $\mathcal{M}_J$ will therefore win. If, however, $\mathcal{D}$ falls somewhere on the left or right tails, then $\mathcal{M}_K$ will win, but only relatively speaking, since the evidence for $\mathcal{M}_J$ and $\mathcal{M}_K$ will both be small.

Essentially, the normalisation requirement functions imposes a penalty on larger $K$ and thereby takes on the function of Occam's razor.

## 4.5 The Ideal Prior

The evidence differs radically from traditional model comparison statistics in that it requires the specification of a parameter prior. While this prior may cancel in specific situations (such as ratios), in general the prior has direct influence on the evidence and therefore on the entire model comparison chain.

For this reason, we aim to list the necessary criteria which priors should fulfil,

1. A prior $p(\boldsymbol{\alpha}|K,\mathcal{M})$ should *reflect the state of knowledge* of the scientist as accurately as possible, where such knowledge is available.
2. Where no such knowledge is available, the prior should be *as impartial* to all values of $\boldsymbol{\alpha}$ as possible, which is the same as saying that it should not represent overconfidence in a particular answer or leave out possible answers.

   Mathematically, the prior should be as uniform as possible over the set of all possible parameter values, $\Omega(\boldsymbol{\alpha})$, but consistent with constraints and prior knowledge.
3. The prior should be as impartial or insensitive as possible regarding the number of parameters $K$, even while the evidence itself should be highly sensitive to $K$.
4. For the purposes of model comparison, the prior should be *proper*, that is, its integral should exist and be normalised,

   $$\int_{\Omega(\boldsymbol{\alpha})} \mathrm{d}\boldsymbol{\alpha}\, p(\boldsymbol{\alpha}|K,\mathcal{M}) = 1. \tag{4.10}$$

   Use of improper priors where this condition is not fulfilled is occasionally convenient, but they are not allowed within the present context of evidence calculation.
5. The ideal prior should be *versatile* in the sense that it should work as well as possible over as wide as wide a class of problems as possible, rather than doing very well in one particular scenario and badly otherwise. That does not mean that such a best prior should always provide the highest evidence for the correct model, but that it should do so *on average* over many different experimental situations and data.
6. Naturally a choice of prior which together with Gaussian likelihoods leads to analytical solutions of the evidence has big advantages over a prior needing numerical calculations, because the solution's behaviour can be studied in detail and applies to an arbitrary number of parameters. Such analytical calculations form an essential part of this text.

## 4.6 Model Comparison & Evidence Within Context

The results derived above are true in general. However, within our context, we set all the model priors equal, $p(\mathcal{M}_m) = \frac{1}{M}$ for all $m$, and the data is given simply by the data vector $\boldsymbol{z}$. A *model* $\mathcal{M}_m$ is made up of a choice of likelihood $p(\boldsymbol{z} \,|\, \boldsymbol{\alpha})$ and of a parameter prior $p(\boldsymbol{\alpha})$.

As a reminder: the likelihood is a model-dependent probability for the data, where calculation of the probability presupposes knowledge of the values of the parameter values. The prior is our assignment of the parameters within this model, prior to the acquisition of data. As is apparent from equation (4.9), the evidence is the prior-weighted average of the likelihood over all possible parameter, and provides a measure as to how a well the model represents the data.

In the present case, the likelihood always takes the form $p(\boldsymbol{z} \,|\, \boldsymbol{\alpha}) = (2\pi)^{-N/2} \exp\big(-\frac{1}{2} N Q(\boldsymbol{\alpha})\big)$ $\big($as given by equation (3.23)$\big)$ with or, once the quadratic form has been diagonalised to yield, $p(\boldsymbol{z} \,|\, \boldsymbol{\beta}) = (2\pi)^{-N/2} \exp\big(-\frac{1}{2} N Q(\boldsymbol{\alpha})\big)$ $\big($as given by equation (3.34)$\big)$.

Since the likelihood always takes on a Gaussian form, and the basis functions entering the design matrix will always remain the same, for the purposes of model comparison we need only keep track of the number of parameters, $K$, in the parameter vector which enter the candidate function (or parametrisation) $\boldsymbol{y}(\boldsymbol{x}) = \sum_{k=1}^{K} f_k(\boldsymbol{x})\, \alpha_k$, hence $K$ represents our measure of model complexity.

For the parameter prior, there will be many variants and choices, which we will track by using $\mathcal{H}_m$ in the conditional for various cases $m$. Hence the models of section 4.2 are denoted by us as

$$\mathcal{M}_m \longrightarrow (K, \mathcal{H}_m)$$

and the parameter prior will be written as $p(\boldsymbol{\beta} \,|\, \mathcal{H}_m)$. We will treat the various priors $\mathcal{H}_m$ as given throughout, so that in our language, the model posterior of equation (4.4) reads

$$p(K \,|\, \boldsymbol{z}, \mathcal{H}_m) = \frac{p(\boldsymbol{z} \,|\, K, \mathcal{H}_m)\, p(K \,|\, \mathcal{H}_m)}{\sum_{K=K_{\min}}^{K_{\max}} p(\boldsymbol{z} \,|\, K, \mathcal{H}_K)\, p(K \,|\, \mathcal{H}_K)} \tag{4.11}$$

where $p(\boldsymbol{z} \,|\, K, \mathcal{H}_m)$ is the evidence and $p(K \,|\, \mathcal{H}_m)$ the prior assignment to $K$ within the prior hypothesis of $\mathcal{H}_m$. For log-Bayes-factor calculations, we select one reference model with $K_b$ parameters and find the log odds for all models between a minimum and maximum $K$-value,

$$\mathcal{E}(K, K_b \,|\, \mathcal{H}_m) = \log \frac{p(\boldsymbol{z} \,|\, K, \mathcal{H}_m)}{p(\boldsymbol{z} \,|\, K_b, \mathcal{H}_m)} \qquad K = K_{\min}, \ldots, K_{\max}. \tag{4.12}$$

For present purposes, however, we are really only interested in finding the single model with the largest evidence rather than calculating the log-odds for all competing models*.

## 4.7 Parameter Inference

Parameter inference plays only a marginal role within this text, and we therefore mention it only in passing. Within context, the underlying model $\mathcal{M}$ always stays the same, and the comparison is between different values of one or more parameters $\boldsymbol{\alpha}$.

* See chapter 6 for details thereof.

Once again making use of the generic equation (4.1) setting $x$ to "the values of the parameters are given by $\boldsymbol{\alpha}$" and $y$ to "the data is $\mathcal{D}$", we obtain a second form of Bayes' rule,

$$p(\boldsymbol{\alpha} \mid \mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D} \mid \boldsymbol{\alpha}, \mathcal{M})\, p(\boldsymbol{\alpha} \mid \mathcal{M})}{p(\mathcal{D} \mid \mathcal{M})} \tag{4.13}$$

in which the evidence appears in the denominator. The probability $p(\boldsymbol{\alpha} \mid \mathcal{D}, \mathcal{M})$ on the left hand side is called the posterior and it plays a central role in fitting routines.

## 4.8 Information Criteria

Chapter 1 section 1.4.1 provided a brief overview of two prominent information criteria in literature which are widely applied to problems of model comparison. Within context, information criteria are equivalent to $-2\mathcal{E}(m, b)$ or in a simpler form $-2\log\big(p(\mathcal{D} \mid \mathcal{M}_m)\big)$.

### 4.8.1 Akaike's Information Criterion

Akaike (1974) formulated the AIC based on model selection on grounds of model similarity between their logarithmic Bayes factors. Within our context, the AIC has the form,

$$\mathrm{AIC} = NQ(\widehat{\boldsymbol{\alpha}}) + 2K. \tag{4.14}$$

We note that as more parameters are added, the chi-squared term tends to decrease, while the number of parameters tend to increase. This can be seen as a compromise between underfitting and overfitting to Occam's razor.

### 4.8.2 Bayesian Information Criterion

Schwarz (1978) formulated the BIC from within the Bayesian framework on grounds of model selection in which the probability of selecting the best model increases with sample size (Burnham and Anderson 2002, p. 286). Within our context, the BIC has the form,

$$\mathrm{BIC} = NQ(\widehat{\boldsymbol{\alpha}}) + N\log K. \tag{4.15}$$

In contrast to BIC, AIC seeks to select the best model at a given sample size, and so, for AIC the best model may vary with $N$; for BIC on the other hand, its best model is independent of $N$. Thus BIC aims to select the best model with accuracy which increases in proportion to the sample size.

Based on the penalty terms, BIC is more tolerant of free parameters than AIC, but less tolerant at higher $N$. As such, between the two criteria, AIC may be more likely to select an overfitting model, while BIC may be more likely to select an underfitting model.

* * *

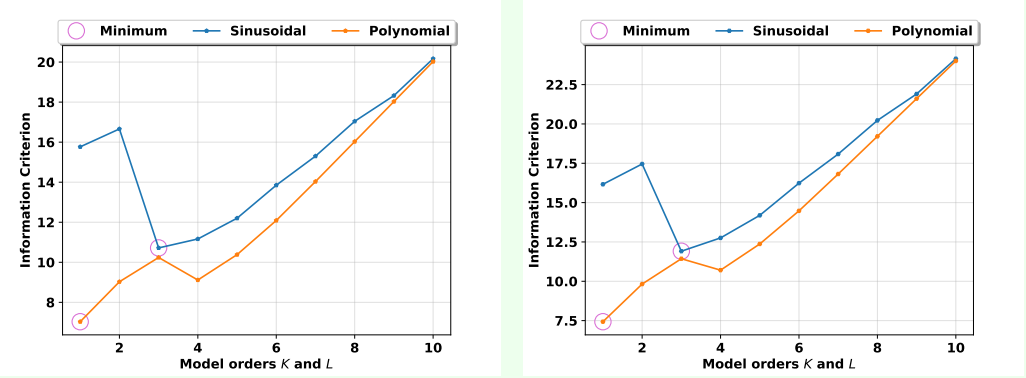
**Example 4.1** **Sally's Seashells *II***

Surmising tides subsequently secured soars in sales of Sally's seashell selling service. In hopes of improving her predictive ability, Sally decides to test her models, and in order to do so constructs 10 submodels for each of her two proposed candidate models,

$$v_n(x_n \mid \boldsymbol{\alpha}, \mathcal{M}_K) = \sum_{k=1}^{K} \alpha_k \, x_n^{k-1}$$

$$w_n(x_n \mid \boldsymbol{\alpha}, \mathcal{M}_L) = \frac{1}{\sqrt{2}} \sum_{\ell=1}^{L} \alpha_\ell \, \sin\left(\frac{\pi(2n+1)(\ell+1)}{2N}\right)$$

Therefore, she has parametrisations of orders $K, L \in \{1, 2 \ldots, 10\}$ for both models, $v$ as well as $w$. In order to assess these models, she considers the AIC and BIC estimates, as modelled by her data. Of course, knowing that higher order polynomials are poor representatives, and knowing about the caveats of overfitting data, she expects higher order models to be penalised. The models are evaluated as,




(i) Progression of the AIC for Sally's models. (ii) Progression of the BIC for Sally's models.

Figure 4.3: Sally's models evaluated by information criteria

Figures 4.3(i) and 4.3(ii) illustrates the performance of Sally's models as evaluated by the AIC and BIC respectively. Interestingly, both information criteria considers the best polynomial to be of order 1, i.e. of the form $\boldsymbol{v}(\boldsymbol{x} \mid \widehat{\boldsymbol{\alpha}}) = \widehat{\alpha}_1$, and we know that this cannot model the tidal variation at all. Despite order 4 (cubic) providing a visually pleasing model of the data, the polynomial model is heavily penalised, therefore she concludes that the polynomial model is inadequate for all $K$.

The sinusoidal model is evaluated to be best when $L = 3$ for both AIC and BIC! She therefore concludes that her best model is $\boldsymbol{w}(\boldsymbol{x} \mid \widehat{\boldsymbol{\alpha}}, \mathcal{M}_3)$.

Sally realises that both of her previous concerns had now been addressed: she was able to model without the pitfall of overfitting, and at the same time penalise models which are not only overfitting, but also poorer (according to her state of knowledge) representatives of the underlying "true" function.

# Spherical Symmetry

## 5.1 Why Spherical Symmetry?

Let us briefly recapitulate the current state of affairs. Given data, $\boldsymbol{z}$, a Gaussian likelihood and minimal prior information on the parameters $\boldsymbol{\alpha}$, the aim is to calculate evidences of the form $p(\boldsymbol{z} \,|\, K, \mathcal{H}_m)$, as set out in section 4.6, in order to find model posteriors for $K$ given $\boldsymbol{z}$ ((see equation (4.11))

$$p(K \,|\, \boldsymbol{z}, \mathcal{H}_m) = \frac{p(\boldsymbol{z} \,|\, K, \mathcal{H}_m)\, p(K \,|\, \mathcal{H}_m)}{p(\boldsymbol{z} \,|\, \mathcal{H}_m)} \qquad K = K_{\min}, \ldots, K_{\max}$$

following which, we find the "best model" for the data from the available collection of model posteriors, compactly represented as the "best $K$", which we may denote as $\widehat{K}$.

As previously stated, we use the symbol $\mathcal{H}_m$ to distinguish between the various choices of parameter prior, $p(\boldsymbol{\beta} \,|\, K, \mathcal{H}_m)$. Given a particular $\mathcal{H}_m$, the desired evidence was previously set out in equation (4.9) in terms of an integral over $\boldsymbol{\alpha}$, however, now that we are using the transformed parameters, $\boldsymbol{\beta}$, the evidence is given by

$$p(\boldsymbol{z} \,|\, K, \mathcal{H}_m) = \int_{\Omega(\boldsymbol{\beta})} \mathrm{d}\boldsymbol{\beta}\, p(\boldsymbol{z} \,|\, \boldsymbol{\beta})\, p(\boldsymbol{\beta} \,|\, K, \mathcal{H}_m), \tag{5.1}$$

where we have shortened notation for the likelihood as given by $p(\boldsymbol{z} \,|\, \boldsymbol{\beta}, K, \mathcal{H}_m)$ to $p(\boldsymbol{z} \,|\, \boldsymbol{\beta})$. We consider it in this way for the reason that if $\boldsymbol{\beta}$ is known, then $K$ is automatically known as well, and the likelihood does not depend on the choice of prior noted in $\mathcal{H}_m$.

The likelihood is given and fixed (as set out in section 3.6) and after transformation of the original $K$-dimensional parameters $\boldsymbol{\alpha}$ to the symmetric $K$-dimensional parameters $\boldsymbol{\beta}$, it is

$$\begin{aligned} p(\boldsymbol{z} \,|\, \boldsymbol{\beta}) &= (2\pi)^{-N/2} \exp\!\left(-\tfrac{1}{2} N Q(\boldsymbol{\beta})\right) \\ \text{with} \qquad Q(\boldsymbol{\beta}) &= Q(\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\mathsf{T}} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \end{aligned}$$

where the parameters are defined by $\widehat{\boldsymbol{\beta}} = \boldsymbol{\Lambda}^{1/2}\, \mathbb{S}\widehat{\boldsymbol{\alpha}}$. By "spherically symmetric", it refers to the likelihood having the same value at all points $\boldsymbol{\beta}$ on the surface of the $K$-dimensional hypersphere centred at $\widehat{\boldsymbol{\beta}}$, for which $\left\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right\|$ is a constant. We call this hypersphere "non-central", since its origin does not coincide with the origin of parameter space.

We now introduce a second spherical symmetry in the $K$-dimensional parameter space, namely the symmetry in that of the parameter prior. Consider the following

$$R_K := \|\boldsymbol{\beta}\| = \sqrt{\textstyle\sum_{k=1}^{K} \beta_k^2} \tag{5.2}$$

for which we have spherical symmetry on the parameter prior such that it depends only on the norm of the parameters (or geometrically, the radius*), and not on the individual components of $\boldsymbol{\beta}$. Thus, it follows that

$$p(\boldsymbol{\beta} \,|\, K, \mathcal{H}_m) \stackrel{!}{=} f(R \,|\, K, \mathcal{H}_m), \tag{5.3}$$

* We will often denote $R_K$ simply as $R$, but we must remember that it is a function of $K$.

for some function $f$. The exact form of the function $f$ is a matter of choice, as long as it is non-negative and its integral over $\mathrm{d}\boldsymbol{\beta}$ is 1.

This assumption of spherical symmetry of the parameter prior on the "central" hypersphere is equivalent to stating that the observer's information remains unchanged under rotation of $\boldsymbol{\beta}$ around the origin; it can be viewed as a "rotational" Principle of Indifference* or "information isotropy", which states that $p(\boldsymbol{\beta} \mid K, \mathcal{H}_m)$ must be uniformly distributed over the *surface* of the hypersphere with radius $\|\boldsymbol{\beta}\|$, and for any such radius, therefore the prior has dependence only on $\|\boldsymbol{\beta}\|$ (De Kock and Eggers 2017b, p. 7).

## 5.2 Spherical & Hyperspherical Coordinates

On the way towards finding acceptable forms of $f$ and expanding its application, we reiterate that we are within the realm of (hyper)spheres, and therefore take a brief detour to consider spherical coordinates in $K$ dimensions. Motivated by the previous section, we now deliberately denote the radius by $R$. For $K = 3$, the transformation from the cartesian coordinates to the spherical coordinates is well known, and defined by

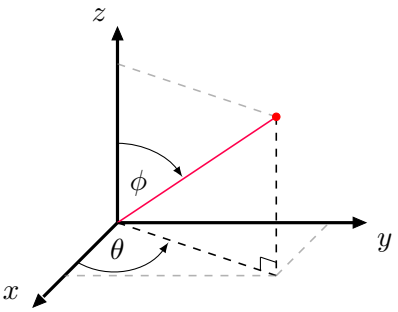



Figure 5.1: Spherical coordinates for $K = 3$.

$$x = R\cos\theta\sin\phi, \quad y = R\sin\theta\sin\phi, \quad z = R\cos\phi.$$

Note that that the azimuthal angle $\theta$ is within the domain $[0, 2\pi]$, while the polar angle, $\phi$, introduces the variation along a new axis, and only needs to span $[0, \pi]$ in order for $(R, \theta, \phi)$ to span the entire $(x, y, z)$ space. It is also well known that the (absolute value of) the Jacobian matrix for the change of variables is

$$|J_3| := \left|\det\left[\frac{\partial(x, y, z)}{\partial(R, \theta, \phi)}\right]\right| = R\sin\phi.$$

For the low $K$, we could still manage with unsubscripted symbols as tabulated below,

| $K$ | Cartesian coordinates | Spherical coordinate transformation |
|---|---|---|
| 2 | $(x, y)$ | $(R\cos\theta,\ R\sin\theta)$ |
| 3 | $(x, y, z)$ | $(R\cos\theta\sin\phi,\ R\sin\theta\sin\phi,\ R\cos\phi)$ |
| 4 | $(x, y, z, w)$ | $(R\cos\theta\sin\phi\sin\varphi,\ R\sin\theta\sin\phi\sin\varphi,\ R\cos\phi\sin\varphi,\ R\cos\varphi)$ |

However, for general $K$ we must use subscripted Cartesian coordinates, $x_k$, and spherical angles, $\theta_\ell$ to describe the transformation,

$$(x_1, x_2, \ldots, x_K) \to (R, \theta_1, \theta_2, \ldots, \theta_{K-1}) \qquad \text{or} \qquad \boldsymbol{x} \to (R, \boldsymbol{\theta})$$

* The Principle of Indifference: when there is no reason to say that one hypothesis is more plausible than another, then all hypotheses should be weighed equally.

with $R$ here given by $\sqrt{\sum_{k=1}^{K} x_k^2}$, and coordinate transformations given by

$$\begin{aligned} x_1 &= R\cos\theta_1 \sin\theta_2 \sin\theta_3 \cdots \sin\theta_{K-1} \\ x_2 &= R\sin\theta_1 \sin\theta_2 \sin\theta_3 \cdots \sin\theta_{K-1} \\ x_3 &= R\cos\theta_2 \sin\theta_3 \cdots \sin\theta_{K-2} \\ &\vdots \qquad\qquad \vdots \\ x_{K-1} &= R\sin\theta_{K-2} \sin\theta_{K-1} \\ x_K &= R\cos\theta_{K-1} \end{aligned} \tag{5.4}$$

with Jacobian matrix

$$J_K = \frac{\partial(x_1, x_2, \ldots, x_K)}{\partial(R, \theta_1, \ldots, \theta_{K-1})} = \begin{bmatrix} \partial_R x_1 & \partial_{\theta_1} x_1 & \ldots & \partial_{\theta_{K-1}} x_1 \\ \partial_R x_2 & \partial_{\theta_1} x_2 & \ldots & \partial_{\theta_{K-1}} x_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial_R x_K & \partial_{\theta_1} x_K & \ldots & \partial_{\theta_{K-1}} x_K \end{bmatrix} \tag{5.5}$$

whose determinant is found to be

$$|J_K| = R^{K-1} \sin\theta_1 \sin^2\theta_2 \sin^3\theta_3 \cdots \sin^{K-2}\theta_{K-2} = R^{K-1} F(\boldsymbol{\theta})$$

In summary, the transformation from a $K$-dimensional Cartesian coordinate system leads to a $K$-dimensional hyperspherical coordinate system which consists of one radius and $K-1$ angles, with $\theta_1 \in [0, 2\pi]$ and $\theta_2, \ldots, \theta_{K-2}, \theta_{K-1} \in [0, \pi]$.

For future use, we note that the integral over all angles of the determinant is given by a product of $R^{K-1}$ and an angular factor $F(\boldsymbol{\theta})$ and is found to be

$$A_K(R) = \int \mathrm{d}\boldsymbol{\theta}\, |J_K| = R^{K-1} \frac{2(\pi)^{K/2}}{\Gamma(\frac{K}{2})}. \tag{5.6}$$

Which is, in fact, the surface area of the $K$-dimensional hypersphere. The derivation of this result is treated at length in appendix C.

## 5.3 Radii & Scales in Fixed Dimension & All Dimensions

### 5.3.1 Spherical Symmetry in Earlier Work

- Zellner (1986) proposed the $g$-prior is, which (in our notation) is defined as

$$p(\boldsymbol{\alpha} \mid g, \sigma, \mathbb{X}, K) = \frac{1}{(\det \mathbb{X}^\intercal\mathbb{X})^{1/2}(2\pi\sigma^2 g/N)^{K/2}} \exp\Big(-\frac{N}{2g\sigma^2}\boldsymbol{\alpha}^\intercal(\mathbb{X}^\intercal\mathbb{X})\boldsymbol{\alpha}\Big) \tag{5.7}$$

  where $\mathbb{X}$ is the same design matrix used in the linear trial function. In other words, $\boldsymbol{\alpha}$ is distributed according to a $K$-dimensional Gaussian with covariance matrix $g\sigma^2(\mathbb{X}^\intercal\mathbb{X})^{-1}$, which means that

$$\boldsymbol{\alpha} \sim \mathcal{N}(\text{mean}, \text{variance}) = \mathcal{N}\big(0, g\sigma^2(\mathbb{X}^\intercal\mathbb{X})^{-1}\big) \tag{5.8}$$

In Zellner's formulation, the parameter $g > 0$ was fixed for analysis but could be chosen freely, while $\sigma$ was a free parameter with its own prior. In our present context of fixed experimental uncertainties, we do not consider $\sigma$ a variable but fixed. It will eventually be set to 1 but we keep it for the moment to keep track of all scale parameters.

Since Zellner's $g$-prior is a Gaussian, the integral equation (1.17) can be solved analytically, making for easy and convenient use.

- One pitfall of the $g$-prior is that the posterior can be reasonable even if $g$ is chosen to be large in an effort to be uninformative. Liang et al. (2007, p. 8) showed that if $g \to \infty$, then comparing Bayes factors will tend to zero and in turn favour the smallest model, irrespective of the information or the data, which is undesired in model selection. Such a phenomenon is known as *Bartlett's Paradox*.

- Liang et al. (2007) introduced hyper-$g$ priors as the successor to Zellner's $g$-prior

$$p(g) = \frac{a-2}{2}(1-g)^{-a/2}, \qquad g > 0 \tag{5.9}$$

which is proper for $a > 2$. Calculations with the hyper-$g$ prior, like the $g$-prior, allowed for computing evidences with analytic results. Within context, the evidence for the hyper-$g$ prior is

$$p(z \mid \mathcal{H}_g) = \frac{(a-2)\Gamma\left(\frac{N}{2}\right)}{2(K+a-2)}\left(\pi \boldsymbol{z}^2\right)^{-N/2} {}_2F_1\left(\begin{matrix}1, N/2 \\ (K+a)/2\end{matrix} \,\middle|\, \frac{\widehat{\boldsymbol{\beta}}^2}{\langle \boldsymbol{z}^2 \rangle}\right) \tag{5.10}$$

- We recognise that spherical symmetry was included within their contexts, for example, observe that equation (5.7) may be expressed by spherically symmetric parameters $\boldsymbol{\beta}$, since $\boldsymbol{\alpha}^{\mathsf{T}}\mathbb{H}\boldsymbol{\alpha} = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}$, and that equation (5.10) incorporates the symmetric parameters.

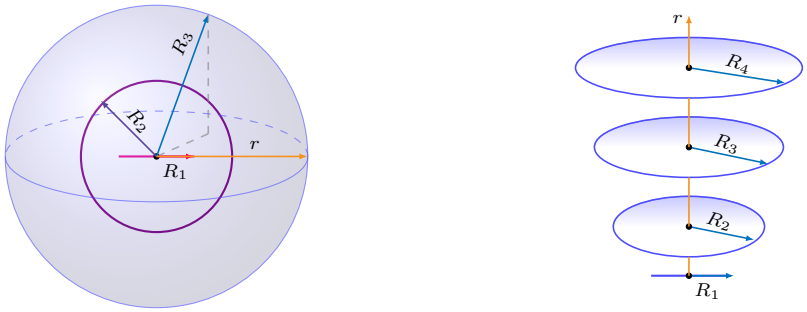
### 5.3.2 Scales as a Guide for Model Comparison

The idea of spherical symmetry as approached in section 5.1 is, of course, nothing new; it was already implicit within the formalisms of both Liang et al. (2007) and Zellner (1986), having made the choice of a Gaussian prior probability on the parameters, in which part of the utility of the Gaussian distribution relied on the spherical symmetry of the distribution itself (De Kock and Eggers 2017b, p. 2).

From the point of view of model comparison, however, spherical symmetry for fixed $K$ is not enough. Our stated aim of comparison of models with *different* $K$ implies that we must now think about comparing, for example, a radius on a $K = 2$ dimensional circle to a radius on a $K = 3$ dimensional sphere, and so on.

In pure mathematics, there is of course no problem: we simply define a radius $r$ and use it for any and all $K$: there is no implicit scale which tells us whether a length $r = 1$ in e.g. two dimensions is any different to $r = 1$ in four dimensions.

However, as soon as there are explicit scales involved, these scales must be taken into account. Therefore, the proposed question is: *What sets the scales for the radii in differently dimensioned parameter spaces within context of our Bayesian linear regression?*




(i) Superimposed spherical spaces of dimensions $K = 1$, $K = 2$, and $K = 3$, with their respective radii of $R_k$ as well as the overarching radius $r$.

(ii) Visualising planes in each dimension, the overaching radius spans all dimensions and establishes a common link between each respective dimension (i.e. model).

Figure 5.2: Dimension specific radii $R_K$ and the overarching $r$.

Figure 5.2 illustrates the generic difference between $K$-specific radii, $R_K$, and an overarching mathematical radius $r$, which applies to all dimensions, in two distinct, yet related ways.

The question of scale-setting will be sought below for two different regimes: scales as set by the likelihood and scales as set by the parameter priors. As previously explained, we do assume that there is no other prior knowledge which separately sets a scale or imposes a constraint.

### 5.3.3 Scales for Fixed $K$

First we consider the likelihood, for which scales are conceptually intuitive. For the likelihood, it is always the *squared scale* which is of interest, therefore we shall reference squares. Moreover, the radius can always be found by taking the square root of the scales that we define.

1. The squared scale in *data space* is set by the squared length of the data vector, such that $\boldsymbol{z}^\mathsf{T}\boldsymbol{z} = \|\boldsymbol{z}\|^2 = \sum_{n=1}^{N} z_n^2$. To track dependence on $N$, however, it is better to use the $N$-average

$$\langle \boldsymbol{z}^2 \rangle = \frac{1}{N}\sum_{n=1}^{N} z_n^2 \tag{5.11}$$

   which will generally tend to a constant for large $N$, and to write $\boldsymbol{z}^2 = N\langle \boldsymbol{z}^2 \rangle$.

2. The scale in *model space* is given by the model vector, $\boldsymbol{\zeta}$, whose minimum chi-squared criterion can be written in terms of the projector of the data $\big($equation (3.28)$\big)$,

$$\widehat{\boldsymbol{\zeta}} = \mathbb{X}\widehat{\boldsymbol{\alpha}} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{z} = \mathbb{P}\boldsymbol{z} \tag{5.12}$$

   Since $\mathbb{P}\mathbb{P} = \mathbb{P}$, the squared model space scale is,

$$\widehat{\boldsymbol{\zeta}}^2 = \boldsymbol{z}^\mathsf{T}\mathbb{P}\boldsymbol{z}; \tag{5.13}$$

   Considering the $N$-normalised version which converges to a constant for large $N$ we have,

$$\langle \widehat{\boldsymbol{\zeta}}^2 \rangle = \frac{\boldsymbol{z}^\mathsf{T}\mathbb{P}\boldsymbol{z}}{N}. \tag{5.14}$$

3. The squared scale in parameter space is given by the squared length of the maximum likelihood vector $\big($which we could call $\widehat{R}_K^2$ in analogy to equation (5.2)$\big)$,

$$\widehat{\boldsymbol{\beta}}^2 := \|\widehat{\boldsymbol{\beta}}\|^2 = \sum_{k=1}^{K} \widehat{\beta}_k^2. \tag{5.15}$$

   The squared scale of the discrepancy between data space and model space is simply the well-known minimum chi-squared criterion,

$$\widehat{\boldsymbol{\chi}}^2 = NQ(\widehat{\boldsymbol{\beta}}) = N\big(\langle \boldsymbol{z}^2 \rangle - \widehat{\boldsymbol{\beta}}^2\big) \tag{5.16}$$

   and is fully determined by $\langle \boldsymbol{z}^2 \rangle$ and $\widehat{\boldsymbol{\beta}}^2$.

4. We note that the maximum likelihood point

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\Lambda}^{1/2}\, \mathbb{S} \widehat{\boldsymbol{\alpha}} \;=\; \boldsymbol{\Lambda}^{1/2}\, \mathbb{S} (\mathbb{X}^{\mathsf{T}} \mathbb{X})^{-1} \mathbb{X}^{\mathsf{T}} \boldsymbol{z} \tag{5.17}$$

   is a function both of the data and of the model, so the parameter space scale is set by a combination of data and model information. Likewise, $\widehat{\boldsymbol{\zeta}}$ is determined both by the data and the MLE. At this point something remarkable happens: the ($N$-normalised) model space squared scale and the (unnormalised) parameter space squared scales are the same, even though they live in different spaces! From $\widehat{\boldsymbol{\zeta}} = \mathbb{X}\widehat{\boldsymbol{\alpha}}$ and equation (3.19), we have

$$\big\langle \widehat{\boldsymbol{\zeta}}^2 \big\rangle = \frac{\widehat{\boldsymbol{\zeta}}^{\mathsf{T}} \widehat{\boldsymbol{\zeta}}}{N} = \frac{\widehat{\boldsymbol{\alpha}}^{\mathsf{T}} \mathbb{X}^{\mathsf{T}} \mathbb{X} \widehat{\boldsymbol{\alpha}}}{N} = \widehat{\boldsymbol{\alpha}}^{\mathsf{T}} \mathbb{H} \widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\beta}}^2 \tag{5.18}$$

   Because $\big\langle \widehat{\boldsymbol{\zeta}}^2 \big\rangle$ is a sum over $N$ terms and $\widehat{\boldsymbol{\beta}}^2$ is a sum over $K$ terms, this equality links the scales between model space and parameter space.

### 5.3.4 An Overarching Radial Scale for All $K$

In the previous subsection, we have sought and found the appropriate scales for model space as well as parameter space for some fixed $K$. As previously suggested, there is need for a higher level of thinking about the problem. For every $K$ the parameter prior $p(\boldsymbol{\beta} \mid K, \mathcal{H}_m)$ is spherically symmetric about the origin; we have many spherical symmetries at the same time.

On the next level, the question is therefore: is there a radius $r$ which applies to *all* models of *all* dimensions $K$? In other words, can model comparison across different $K$ be written in such a way that *all* the spherical symmetries are taken into account, without favouring any one $K$? In other words, can we design a radius $r$ which is as insensitive as possible to the implicit scales set by all the different $\widehat{\boldsymbol{\beta}}^2$ scales for different $K$?

The question of an overarching radius was first considered by De Kock and Eggers (2017b), who showed that the spherical symmetry in the the formalisms of Zellner (1986) and Liang et al. (2007) with regards to the $g$- and hyper-$g$ priors, are in fact special cases of a formulation which introduces a dimension-transcendent radius, $r$, that is common to all models.

The benefit of a single $r$ is an astonishingly simple formula for the so-called $r$-conditioned likelihood which we rederive and extend in section 5.4.

Before doing so, let us remain with the issue of scaling for a moment. We first note that $R^2 = \|\boldsymbol{\beta}\|^2 = \sum_{k=1}^{K} \beta_k^2$ implies that $R$ scales as $\sqrt{K}$ because each term in the sum is a square and therefore positive. The scaling of $R$ will not be exact but depend on the magnitude of each additional $\widehat{\beta}_k^2$ as $K$ is increasedNevertheless, on average it should be true that

$$R \approx \sqrt{K}. \tag{5.19}$$

Conceptually, we may think of this scaling as superimposing the dimensions of all $K$ via the radius $r$ — in much the same way as visualised in figure 5.2. We therefore consider a prior for $\boldsymbol{\beta}$ given $r$,

$$p(\boldsymbol{\beta} \,|\, r, K, \mathcal{H}_m) = c \cdot \delta(R - \ell r) \tag{5.20}$$

with $c$ a normalisation constant determined by $1 \stackrel{!}{=} \int \mathrm{d}\boldsymbol{\beta}\, p(\boldsymbol{\beta} \,|\, r, K, \mathcal{H}_m)$, and multiplied by the Dirac delta*. Following this prior, we consider a transformation to hyperspherical coordinates and use of equation (5.6) gives us

$$1 \stackrel{!}{=} c \int \mathrm{d}R \int \mathrm{d}\boldsymbol{\theta}\, \delta(\ell r - R)\, |J_K| = c\,(\ell r)^{K-1} \int \mathrm{d}\boldsymbol{\theta}\, F(\boldsymbol{\theta}) = c\,(\ell r)^{K-1}\, \frac{2(\pi)^{K/2}}{\Gamma(\frac{K}{2})}$$

and so (inserting $\ell$ into the notation for the prior)

$$p(\boldsymbol{\beta} \,|\, r, K, \ell, \mathcal{H}_m) = \frac{\Gamma(\frac{K}{2})}{2(\pi)^{K/2}}\, \frac{\delta(R - \ell r)}{(\ell r)^{K-1}}. \tag{5.21}$$

The above discussion would motivate the choice such that $\ell = \sqrt{K}$. Of course, we are not limited to this particular choice of radial scale and we note that when $\ell = 1$, then we recover previous, unscaled results as used by De Kock and Eggers (2017b). We therefore postulate that $\ell$ could be defined as a function of $K$, such that $\ell \in \{K^0, K^{1/2}, K^1, \ldots\}$, or any suitable function of $K$ which can maintain the desired scale relation between $r$ and $R$.

### 5.3.5 General Form for the Spherically Symmetric Prior

Priors are supposed to reflect one's state of knowledge before the data is taken. Often, translating that knowledge into a mathematical form is neither clear nor easy, so typically there is a choice of prior. While choice of a prior may not necessarily be unique, at least there are some basic criteria which should be applied as listed in section 4.5. The use of spherical symmetry in general, and a spherically symmetric prior is consistent with that list of criteria.

In equation (5.3), the statement $p(\boldsymbol{\beta} \,|\, K, \mathcal{H}_m) \stackrel{!}{=} f(R \,|\, K, \mathcal{H}_m)$ had been one of spherical symmetry on a $K$-specific radius $R$; now, we extend that assumption to state that the prior for $\boldsymbol{\beta}$ should be symmetric also for every given $r$ as already tested in equation (5.20). We now generalise to find all possible forms for a spherically symmetric prior for $\boldsymbol{\beta}$ given fixed $r$ by demanding that the prior be spherically symmetric (i.e. a function of $\|\boldsymbol{\beta}\|$ only) for every $r$,

$$p(\boldsymbol{\beta} \,|\, r, K, \ell, \mathcal{H}_m) \stackrel{!}{=} f(R \,|\, r, K, \ell, \mathcal{H}_m). \tag{5.22}$$

* Note that there are two levels of Dirac delta"constraints". The first being that which arises from equation (5.2), which is originates from the coordinate transformation, and the second being the scaling of the radii of superimposed dimensions, in other words it arises as a scaling relation between radii.

Again, the function $f$ in equation (5.22) must be some positive function of $R$; it can be written as a probability density function (PDF) by introducing a normalisation constant,

$$p(\boldsymbol{\beta} \,|\, r, K, \ell, \mathcal{H}_m) \stackrel{!}{=} c\, p(R \,|\, r, K, \ell, \mathcal{H}_m). \tag{5.23}$$

We emphasise that this is a *statement of symmetry* and not the result of a transformation to a spherical coordinate system. However, the statement of a transformation of variables (as in equation (5.4)) is still valid, and we will make use of it.

The new prior $p(R \,|\, r, K, \ell, \mathcal{H}_m)$ clearly has the function of relating the $K$-specific radius $R$ to the $K$-overarching radius, and we are free to choose whatever functional form we like, as long as it is properly normalised over $0 \leq R < \infty$. In addition, $p(\boldsymbol{\beta} \,|\, r, K, \ell, \mathcal{H}_m)$ must be properly normalised over the domain $\Omega(\boldsymbol{\beta}) = \mathbb{R}^K$,

$$1 \stackrel{!}{=} \int_{\Omega(\boldsymbol{\beta})} \mathrm{d}\boldsymbol{\beta}\, p(\boldsymbol{\beta} \,|\, r, K, \ell, \mathcal{H}_m) = \int_{\Omega(\boldsymbol{\beta})} \mathrm{d}\boldsymbol{\beta}\, c\, p(R \,|\, r, K, \ell, \mathcal{H}_m) \tag{5.24}$$

Note that the $K$-fold integral is over parameter space $\Omega(\boldsymbol{\beta})$, not just one variable $R$. At this point, we transform coordinates from $\boldsymbol{\beta}$ to the $K$-spherical coordinates of equation (5.4):

$$\mathrm{d}\boldsymbol{\beta} = \mathrm{d}R\, \mathrm{d}\boldsymbol{\theta} \left| \frac{\partial(\boldsymbol{\beta})}{\partial(R, \boldsymbol{\theta})} \right| = \mathrm{d}R\, \mathrm{d}\boldsymbol{\theta}\, |J_K| \tag{5.25}$$

and so

$$1 \stackrel{!}{=} c \int_0^\infty \mathrm{d}R\, p\big(R \,|\, r, K, \mathcal{H}_m\big) \int \mathrm{d}\boldsymbol{\theta}\, |J_K| \tag{5.26}$$

with $|J_K| = R^{K-1} F(\boldsymbol{\theta})$, and once again using the result of equation (5.6), it follows that

$$1 \stackrel{!}{=} c \cdot \frac{(2\pi)^{K/2}}{\Gamma\left(\frac{K}{2}\right)} \int_0^\infty \mathrm{d}R\, R^{K-1} p(R \,|\, r, K, \mathcal{H}_m). \tag{5.27}$$

and inserting this into equation (5.23) we finally obtain

$$p(\boldsymbol{\beta} \,|\, r, K, \mathcal{H}_m) = \frac{\Gamma\left(\frac{K}{2}\right)}{(2\pi)^{K/2}} \frac{p(R \,|\, r, K, \mathcal{H}_m)}{\int_0^\infty \mathrm{d}R\, R^{K-1}\, p(R \,|\, r, K, \mathcal{H}_m)} \tag{5.28}$$

This result is the most general form for a parameter prior consistent with spherical symmetry for all possible $K$. We are free to choose any $p(R \,|\, r, K, \mathcal{H}_m)$ as long as it is properly normalised. The special case equation (5.21) is the result of the choice

$$p(R \,|\, r, K, \mathcal{H}_m) = \delta(R - \ell r).$$

Note that the general form of equation (5.28) establishes a general projection onto a spherically symmetric surface. The choice of the Dirac delta, as consistent with De Kock and Eggers (2017b), corresponds to the projection onto the surface of a sphere*.

* This projection can be thought of as the projection of the prior $p(\boldsymbol{\alpha})$, even though the prior in parameters $\boldsymbol{\alpha}$ is not considered explicitly, since we establish the prior in *bprmb* directly.

## 5.4 The *r*-Conditioned Likelihood

**Identities 5.1**

$$\delta(x-y) = 2x\,\delta(x^2-y^2) \tag{5.29}$$

$$\delta(x^2-y^2) = \frac{1}{2\pi i}\int_C \mathrm{d}s\, \exp(sx^2 - sy^2) \tag{5.30}$$

Where $C$ is along the imaginary line from $(c-i\infty)$ to $(c+i\infty)$ (Watson 1922).

We now recalculate the evidence $p(\boldsymbol{z}\,|\,K,\mathcal{H}_m)$ taking into account the assumption of extended spherical symmetry and the overarching radius $r$. Note, once more, that $K$ is the dimension of our model and $\mathcal{H}_m$ is our choice in parameter assignment on the prior (for various $m$). Recall that The first step is to expand the evidence in terms of $r$,

$$p(\boldsymbol{z}\,|\,K,\mathcal{H}_m) = \int_0^\infty \mathrm{d}r\, p(\boldsymbol{z},r\,|\,K,\mathcal{H}_m) \;=\; \int_0^\infty \mathrm{d}r\, p(\boldsymbol{z}\,|\,r,K,\mathcal{H}_m)\, p(r\,|\,K,\mathcal{H}_m) \tag{5.31}$$

The "$r$-prior", $p(r\,|\,K,\mathcal{H}_m)$, will be treated extensively later. The other factor $p(\boldsymbol{z}\,|\,r,K,\mathcal{H}_m)$ is the "$r$-conditioned likelihood" or simply $r$-likelihood which we now derive for the specific scaling choice $p(R\,|\,r,K,\mathcal{H}_m) = \delta(R-\ell r)$.

Expanding in terms of $\boldsymbol{\beta}$ similar to that in equation (5.1), the $r$-likelihood, as inclusive of $\ell$, is defined as

$$p(\boldsymbol{z}\,|\,r,K,\ell,\mathcal{H}_m) = \int \mathrm{d}\boldsymbol{\beta}\, p(\boldsymbol{z}\,|\,\boldsymbol{\beta},r,\ell)\, p(\boldsymbol{\beta}\,|\,r,K,\ell,\mathcal{H}_m)$$

We express the $\boldsymbol{\beta}$-likelihood and its exponents of equations (3.34) and (3.36) in the form

$$p(\boldsymbol{z}\,|\,\boldsymbol{\beta},r,\ell) = C\,\exp\!\left(-\tfrac{1}{2}\boldsymbol{z}^2\right)\exp\!\left(-\tfrac{N}{2}(\boldsymbol{\beta}^2 - 2\langle\boldsymbol{\beta},\widehat{\boldsymbol{\beta}}\rangle)\right),$$

with $C = (2\pi)^{-N/2}$, and so the $r$-likelihood follows as

$$\begin{aligned}
p(\boldsymbol{z}\,|\,r,K,\ell,\mathcal{H}_m) &= \int \mathrm{d}\boldsymbol{\beta}\, p(\boldsymbol{z}\,|\,\boldsymbol{\beta},r,\ell)\, p(\boldsymbol{\beta}\,|\,r,K,\ell,\mathcal{H}_m) \\
&= C\,\exp\!\left(-\tfrac{1}{2}\boldsymbol{z}^2\right)\int \mathrm{d}\boldsymbol{\beta}\,\exp\!\left(-\tfrac{N}{2}(\boldsymbol{\beta}^2 - 2\langle\boldsymbol{\beta},\widehat{\boldsymbol{\beta}}\rangle)\right) \\
&\qquad \cdot \frac{2\ell r\,\delta\!\left((\ell r)^2 - \boldsymbol{\beta}^2\right)}{A_K(\ell r)} \\
&= C\frac{2\ell r}{A_K(\ell r)}\,\exp\!\left(-\tfrac{N}{2}\boldsymbol{z}^2\right)\int \mathrm{d}\boldsymbol{\beta}\,\exp\!\left(-\tfrac{N}{2}(\boldsymbol{\beta}^2 - 2\langle\boldsymbol{\beta},\widehat{\boldsymbol{\beta}}\rangle)\right) \\
&\qquad \cdot \int_C \mathrm{d}s\,\frac{1}{2\pi i}\exp\!\left(s(\ell r)^2 - s\boldsymbol{\beta}^2\right) \\
&= C\frac{2\ell r}{A_K(\ell r)}\,\exp\!\left(-\tfrac{1}{2}\boldsymbol{z}^2\right)\frac{1}{2\pi i}\int_C \mathrm{d}s\,\exp\!\left(s(\ell r)^2\right) \\
&\qquad \cdot \int \mathrm{d}\boldsymbol{\beta}\,\exp\!\left(-\tfrac{1}{2}(N+2s)\boldsymbol{\beta}^2 + N\langle\boldsymbol{\beta},\widehat{\boldsymbol{\beta}}\rangle\right) \\
&= C\frac{2\ell r}{A_K(\ell r)}\,\exp\!\left(-\tfrac{1}{2}\boldsymbol{z}^2\right)\frac{1}{2\pi i}\int_C \mathrm{d}s\,\exp\!\left(s(\ell r)^2\right) \\
&\qquad \cdot \left[\left(\frac{2\pi}{N+2s}\right)^{K/2}\exp\!\left(\frac{N^2\widehat{\boldsymbol{\beta}}^2}{2(N+2s)}\right)\right]
\end{aligned} \tag{5.32}$$

$$= C\, \exp\!\big(-\tfrac{1}{2}(\boldsymbol{z}^2 + N(\ell r)^2)\big)\, {}_0F_1\!\left(\begin{matrix} \\ K/2 \end{matrix}\,\middle|\, \left(\frac{N\ell r\,\widehat{\boldsymbol{\beta}}}{2}\right)^2\right) \tag{5.33}$$

With this result we have realised how *all models* are resolved in terms of the overarching radius. In other words, the $K$-dimensional system in parameter space has been reduced to a 1-dimensional system in the radial dimension (De Kock and Eggers 2017b, p. 5). It is an effective illustration of how the overarching radius encompasses all models.

## 5.5 The *r*-Prior

Having resolved the consequences of spherical symmetry of $\boldsymbol{\beta}$ for a fixed $r$, we now consider the radial prior $p(r \,|\, K, \mathcal{H}_m)$, for which we need to complete the calculation of the evidence in terms of equation (5.31). The introduction of the overarching radius was motivated by the need to connect one variable in that of the radii of models with different $K$. Therefore, any prior for $r$ should preferably not depend on $K$ at all.

Naturally the $r$-prior will be written in terms of one of the usual probability densities. This will involve specifying one or more *hyper-parameters*, which are parameters that determine the form of $p(r \,|\, \mathcal{H}_m)$. To keep track of these hyper-parameters, we will be adding them to the notation of the $r$-prior as we proceed.

In principle, we can choose any functional form we like for the $r$-prior as long as $0 \leq r < \infty$ is allowed. It helps, however, to think of it as also being the result of projecting a hyper-prior onto a sphere, in the same way that the Gaussian for $\boldsymbol{z}$ became a Gaussian for the parameters $\boldsymbol{\beta}$ which were then projected onto the surface of the hypersphere with radius $R$.

Thus, if we assume $r$ to be the radius of a hyper-prior in a $\overline{K}$-dimensional spherically symmetric Gaussian for variables $\boldsymbol{b} = (b_1, b_2, \ldots, b_{\overline{K}})$ centred at the origin, with variance $\Delta^2$, then projection onto the radial* probability follows the same path taken in previous sections. Doing the calculation as usual, while also keeping track of the new hyperparameters $\overline{K}$ and $\Delta$,

$$\begin{aligned} p(r \,|\, \Delta, \overline{K}, \mathcal{H}_c) &= \int \mathrm{d}\boldsymbol{b}\, p(\boldsymbol{b} \,|\, r, \Delta, \mathcal{H}_{\overline{K}})\, p(\boldsymbol{b} \,|\, \mathcal{H}_{\overline{K}}) \\ &= \int \mathrm{d}\boldsymbol{b}\, (2\pi\Delta^2)^{-\overline{K}/2}\, \exp\!\big(-\tfrac{1}{2\Delta^2}\boldsymbol{b}^2\big)\, 2r\, \delta(r^2 - \boldsymbol{b}^2) \end{aligned}$$

we obtain a probability density which can be understood as a so-called "central $r^2$-Gamma distribution" which is related to the usual Gamma distribution by transforming to the variable $x = r^2$,

$$p(r \,|\, \overline{K}, \Delta, \mathcal{H}_c) = \frac{\sqrt{2}}{\Delta}\left(\frac{r}{\sqrt{2}\Delta}\right)^{\overline{K}-1} \frac{1}{\Gamma\left(\frac{\overline{K}}{2}\right)} \exp\!\left(-\frac{r^2}{2\Delta^2}\right) \tag{5.34}$$

This result is closely related to the chi-squared distribution used in statistics. We have also for the first time introduced the generic hypothesis symbol $\mathcal{H}_m$ as the specific $\mathcal{H}_c$ which as shorthand for "central $r^2$ Gamma distribution".

What values should we assign to the two new hyperparameters, $\overline{K}$ and $\Delta$? For the time being, we just consider $\overline{K}$, for the moment. From equation (5.34), its role is clearly to specify at what

* This corresponds to moving from the spherical surface to within the spherical volume.

power $r$ grows, and we observe that the radius scales as $\overline{K}-1$, which is essential for maintaining spherical symmetry within a spherical volume. A larger $\overline{K}$ also implies a larger mean $r$ and a wider shape.

Since $r$ is supposed to be invariant with respect to model order, there are no direct links between the values of $\overline{K}$ and the set of $K = \{K_{\min}, \dots, K_{\max}\}$ models considered during model comparison, but it makes sense to not let $\overline{K}$ differ massively from this set, for that reason we refer to $\overline{K}$ as the overarching model order.

This result may be extended when parametrising the $r$-prior according to a $\overline{K}$-dimensional Gaussian centred a distance $\boldsymbol{\gamma}$ away from the origin $\mathcal{N}(\boldsymbol{b} \,|\, \boldsymbol{\gamma}, \Delta^2)$, which yields the following,

$$p(r \,|\, \boldsymbol{\gamma}, \overline{K}, \Delta, \mathcal{H}_\gamma) = \frac{\sqrt{2}}{\Delta}\left(\frac{r}{\sqrt{2}\Delta}\right)^{\overline{K}-1} \frac{1}{\Gamma(\frac{\overline{K}}{2})} \exp\left(-\frac{(r^2+\boldsymbol{\gamma}^2)}{2\Delta^2}\right) {}_0F_1\left(\overline{K}/2 \;\middle|\; \left(\frac{r\boldsymbol{\gamma}}{2\Delta^2}\right)^2\right) \tag{5.35}$$

We call the PDF in equation (5.34) the *central* gamma $r$-prior, and equation (5.35) the *non-central* gamma $r$-prior (De Kock and Eggers 2017a, p. 9). We may observe that when $\boldsymbol{\gamma}^2 = 0$, then the non-central prior simplifies to the central one. Once again, we note that both priors scale as $r^{\overline{K}-1}$, which is necessary for maintaining spherical symmetry in a *volume* in spherical space*.

## 5.6 *r*-Priors & *r*-Likelihoods: Summary of Dependencies

The framework of spherical symmetry and the acquired results up to this point can be summarised as

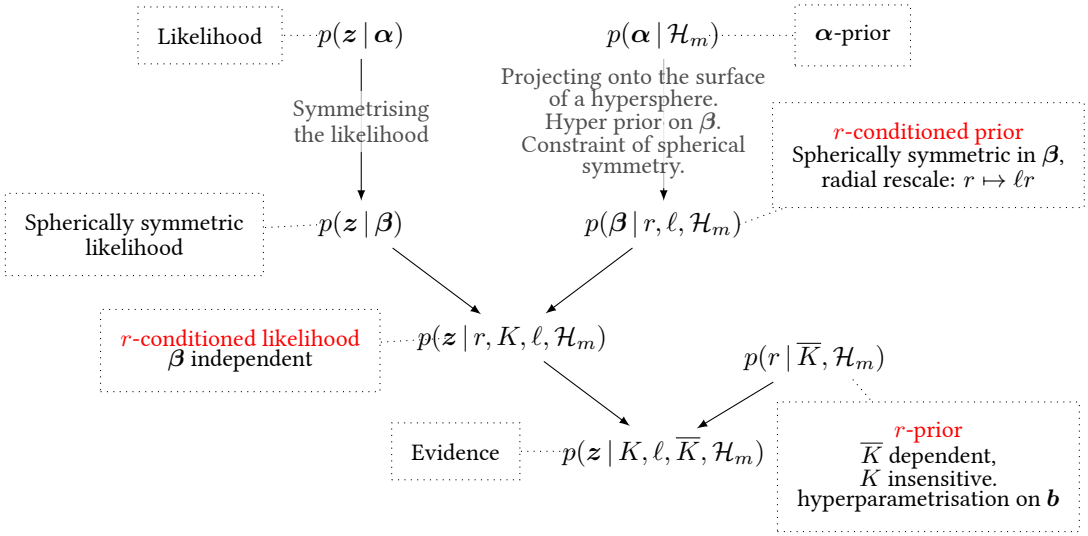



Figure 5.3: Scheme of dependencies

In selecting an $r$-prior, we have free choice provided that we maintain the symmetry in spherical space, this is in contrast to the frameworks of Zellner and Liang et al. which are restricted to mixtures of conjugate priors (De Kock and Eggers 2017b, p. 2).

* When considering the prior on the radius, we are moving from the surface to within the volume, and we require spherical symmetry in both cases, for the surface as well as within the volume. These two details are conceptualised through sampling in spherical space in appendix C, first as sampling symmetrically on a spherical surface, and then symmetrically within a spherical volume.

## 5.7 The Resulting Evidences

With an analytical expression for the $r$-likelihood in hand in equation (5.33) and with one of the $r$-priors, either central or non-central, we can now find the evidence as an $r$-integral, as already done by example in equation (5.31) (now including the hyperparameters inherited from the prior),

$$p(\boldsymbol{z} \,|\, K, \ell, \overline{K}, \mathcal{H}_m) = \int_0^\infty \mathrm{d}r \, p(\boldsymbol{z} \,|\, r, K, \ell, \mathcal{H}_m) \, p(r \,|\, \overline{K}, \mathcal{H}_m) \tag{5.36}$$

The fact that we are left with just a one-dimensional integral highlights the importance in utilising the framework of spherical symmetry. Here we have the evidence as expressed by a 1-dimensional integral in contrast to the $K$-dimensional integral as before. Moreover, what we have in terms of a $K$ sensitive likelihood and a $K$ insensitive prior may lead to a system with a $K$ sensitive evidence which results from a $K$ insensitive prior. In other words, a prior which is unbiased with respect to our candidate models, but an evidence which is sensitive to the candidate models.

In determining the evidence, we first consider the central $r$-prior of equation (5.34) and find

$$\begin{aligned} p(\boldsymbol{z} \,|\, K, \ell, \overline{K}, \mathcal{H}_m) &= \int \mathrm{d}r \, p(\boldsymbol{z} \,|\, r, K, \ell, \mathcal{H}_m) \, p(r \,|\, \overline{K}, \Delta, \mathcal{H}_c) \\ &= \frac{C \exp\left(-\frac{1}{2}\boldsymbol{z}^2\right)}{\left(N\ell^2\Delta^2 + 1\right)^{\overline{K}/2}} \, {}_1F_1\left(\begin{matrix} \overline{K}/2 \\ K/2 \end{matrix} \,\middle|\, \frac{N^2(\ell\Delta)^2}{2(N\ell^2\Delta^2+1)}\widehat{\boldsymbol{\beta}}^2\right) \end{aligned} \tag{5.37}$$

where again $C = (2\pi)^{-N/2}$. This result may be referred to as the *central evidence*. The asymptotic form is found from the identity for $x \gg 1$

$${}_1F_1\left(\begin{matrix} a \\ c \end{matrix} \,\middle|\, x\right) \simeq \frac{\Gamma(c)}{\Gamma(a)} \, x^{a-c} \, e^x$$

which means in our case that for $N \gg 1$

$$p(\boldsymbol{z} \,|\, \mathcal{H}_m) \simeq \frac{C \exp\left(-\frac{1}{2}\left(\boldsymbol{z}^2 - N\widehat{\boldsymbol{\beta}}^2\right)\right)}{\left(N\ell^2\Delta^2 + 1\right)^{\overline{K}/2}} \, \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(\frac{\overline{K}}{2}\right)} \left(\frac{N\widehat{\boldsymbol{\beta}}^2}{2}\right)^{(\overline{K}-K)/2} \tag{5.38}$$

so that, recognising the minimum-chisquared criterion of equation (5.16) in the exponent, we get for the log-evidence

$$\begin{aligned} -2\log p(\boldsymbol{z} \,|\, \mathcal{H}_m) = \widehat{\boldsymbol{\chi}}^2 + \overline{K}\log(1 + N\Delta^2\ell^2) + (K - \overline{K})\log\left(\tfrac{1}{2}N\widehat{\boldsymbol{\beta}}^2\right) \\ - 2\log\Gamma\left(\tfrac{K}{2}\right) + 2\log\Gamma\left(\tfrac{\overline{K}}{2}\right) - 2\log C \end{aligned} \tag{5.39}$$

Which forms the central $r$-information criterion. For the purposes of model comparison, we are interested only in the $K$-dependence, so the non-dependent terms may be omitted.

The *non-central evidence* follows by integrating the $r$-likelihood with the non-central $r$-prior of equation (5.35),

$$p(\boldsymbol{z} \,|\, \boldsymbol{\gamma}, \overline{K}, \mathcal{H}_m) = \int \mathrm{d}r \, p(\boldsymbol{z} \,|\, r, \ell, K, \mathcal{H}_m) \, p(r \,|\, \boldsymbol{\gamma}, \overline{K}, \Delta, \mathcal{H}_\gamma)$$

$$
\begin{aligned}
&= \frac{C \exp\left(-\frac{1}{2\Delta^2}(\boldsymbol{\gamma}^2 + \Delta^2 \boldsymbol{z}^2)\right)}{(N\ell^2\Delta^2 + 1)^{-\overline{K}/2}} \\
&\quad \cdot \Psi_2\left(\overline{K}/2 : \overline{K}/2, K/2 \;\middle|\; \frac{\boldsymbol{\gamma}^2}{2\Delta^2(N\ell^2\Delta^2 + 1)}, \frac{(N\ell\Delta)^2 \widehat{\boldsymbol{\beta}}^2}{2(N\ell^2\Delta^2 + 1)}\right).
\end{aligned}
\tag{5.40}
$$

Since the Humbert $\Psi_2$ function lacks an asymptotic form, we simply consider the quantity $-2\log p(\boldsymbol{z} \,|\, \boldsymbol{\gamma}, \overline{K}, \mathcal{H}_m)$ for the result in equation (5.40) to form the non-central $r$-information criterion.

Note that in each case, for both central and non-central evidences, the results are fully analytic*. In this sense the problem associated with the curse of dimensionality in evidence calculations is overcome via the framework of spherical symmetry, at least within the context of Gaussian linear regression.

Once again, note that when $\boldsymbol{\gamma} \to \mathbf{0}$, the non-central variant reduces to the central case. With $\boldsymbol{\gamma}^2$ being indefinite, it may be of consideration to marginalise this quantity from the evidence, in which case assigning a uniform, but proper prior on $\boldsymbol{\gamma}^2$ is considered in the form of a half-Gaussian (De Kock and Eggers 2017a, p. 10),

$$
p(\boldsymbol{\gamma}^2 \,|\, \sigma_\gamma) = \sqrt{\frac{2}{\pi\sigma_\gamma^2}} \exp\left(-\tfrac{1}{2\sigma}\boldsymbol{\gamma}^2\right) \tag{5.41}
$$

It is worth noting that $\boldsymbol{\gamma}$ should remain indeterminate unless our state of knowledge on the system may assign a value to it. When marginalised from the evidence in our current framework, the result yields

$$
\begin{aligned}
p(\boldsymbol{z} \,|\, \mathcal{H}_m) &= \int \mathrm{d}\boldsymbol{\gamma}\, p(\boldsymbol{z} \,|\, \boldsymbol{\gamma}, \mathcal{H}_m)\, p(\boldsymbol{\gamma}^2 \,|\, \sigma_\gamma) \\
&= \frac{\exp\left(-\frac{1}{2}\boldsymbol{z}^2\right)}{(N\ell^2\Delta^2 + 1)^{-\overline{K}/2}} \\
&\quad \cdot \Psi_1\left(\overline{K}/2 : {}^{1}\!/_{2} \atop \overline{K}/2, K/2 \;\middle|\; \frac{\sigma_\gamma}{2\Delta^2(N\ell^2\Delta^2 + 1)(N\sigma_\gamma^2 + \Delta^2)}, \frac{(N\ell\Delta)^2 \widehat{\boldsymbol{\beta}}^2}{2(N\ell^2\Delta^2 + 1)}\right)
\end{aligned}
\tag{5.42}
$$

Similarly to the Humbert $\Psi_2$ function, the Humbert $\Psi_1$ function also lacks an appropriate asymptotic form, therefore we consider $-2\log p(\boldsymbol{z} \,|\, \mathcal{H}_m)$ to define the marginalised non-central criterion.

In contrast to this result, we shall elaborate on a setting which includes $\boldsymbol{\gamma}$, instead of opting for the marginalised evidence as seen here, thus we shall consider equation (5.40) to be the primary result of the non-central evidence.

Note that the central and non-central evidences of equations (5.37) and (5.40) respectively simplify to an exponential function and a Humbert $\Psi_1$ function if $\overline{K} = K$. If we set $\ell = 1$, then these results simplify further and yield those of De Kock and Eggers (2017a,b).

* This is an important consequence since high dimensional integrals such as those of evidences are generally solved numerically.

CHAPTER 6

# Simulation Analysis

## 6.1 Multiple Model Linear Regression

Up to this point, we have focused on the *inference* using different models, which in our case were defined by a fixed set of basis functions but including successively more of them; a $K$-dimensional model consisted of $K$ such functions. We now first show that all these models can be put into a single vector-matrix equation representing a system in which we could respectively model the data by $K_n$ distinct models of orders $K \in [K_{\min}, K_{\max}]$, where $K_n = K_{\max} - K_{\min}$. For this we first revisit two results that we previously derived, the Moore-Penrose inverse as well as the projection matrix,

$$\widehat{\boldsymbol{\alpha}} = (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\boldsymbol{z}$$
$$\mathbb{P} = \mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}$$

For the $K_{\max}$-dimensional vector $\widehat{\boldsymbol{\alpha}}$, we generally generated the $N$-dimensional vector of the best model, $\widehat{\boldsymbol{\zeta}}$. Created in this way, this is essentially the model defined by $K_{\max}$ parameters and the full $N \times K_{\max}$ design matrix. In order to produce models of the subspaces thereof, i.e. lesser order models, we make use of the *projector of rank* $K$, $\mathbb{P}_K$. This means that for the $N \times N$ projection matrix $\mathbb{P}$, which is full rank, the matrix $\mathbb{P}_K$ is of rank $K$, yet of dimension $N \times N$, essentially a consideration of $\mathbb{P}$ up to $K$ columns, while the successive $N - K$ columns are zero. We may thus create models of arbitrary orders of $K$ by means of a single formula with multiple projection matrices,

$$\begin{aligned}\widehat{\boldsymbol{\zeta}}_K &= \mathbb{X}\mathbb{P}_K\widehat{\boldsymbol{\alpha}} \\ &= \mathbb{X}_K\widehat{\boldsymbol{\alpha}}.\end{aligned} \tag{6.1}$$

The resultant formulation then includes $K_n$ distinct models* of orders $K \in [K_{\min}, K_{\max}]$. For any modelled data we will have $K_n$ distinct $\boldsymbol{\chi}^2$ estimates, in other words we generate a vector containing elements of $\boldsymbol{\chi}^2_K$, which is the chi-squared criterion for *each* respective model,

$$\boldsymbol{\chi}^2_K = (\boldsymbol{z} - \boldsymbol{\zeta}_K)^{\mathsf{T}}(\boldsymbol{z} - \boldsymbol{\zeta}_K). \tag{6.2}$$

In summary, we generate $K_n$ models of orders $K \in [K_{\min}, K_{\max}]$ by considering the projection of $\boldsymbol{z}$ of not only the model order $K_{\max}$ (i.e. design matrix of rank $K_{\max}$), but with respect to all submodels of order $K$, thus resulting in models of all orders from $K_{\min}$ ranging up to $K_{\max}$. Essentially extending linear regression as previously formulated to that of multiple models.

### 6.1.1 Regression with Orthogonal Design Matrices

When concerned with simulation, and numerical computing in general, a choice in orthonormal basis functions make part of an important part of numerical computation: simplifying linear systems and numerical efficiency. Orthonormal basis functions yield an orthonormal design

* This is essentially an $N \times K_n$ matrix with each column representing a model of ascending order.

matrix, which in turn results in $\mathbb{X}^{\mathsf{T}}\mathbb{X} = \mathbb{H} = \mathbb{I}$. Within context, this assignment is of high importance, because the inverse of any high dimensional matrix may be non-trivial to compute, in terms of both numerical complexity as well as time expenditure*

Since the Fourier basis forms an orthogonal basis for $L_2$, we consider the desired basis functions to be as defined by the terms of the discrete cosine transform (Ahmed et al. 1974, p. 90),

$$f_k(x_n) = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \frac{\sqrt{2}}{\sqrt{N}} \cos\left(\frac{\pi(2n+1)K}{2N}\right) & k > 1 \end{cases} \tag{6.3}$$

and similarly we may also consider the basis functions as defined by the terms of the discrete sine transform (Kekre and Solanki 1978, p. 308),

$$f_k(x_n) = \frac{\sqrt{2}}{\sqrt{N}} \sin\left(\frac{\pi(2n+1)(K+1)}{2N}\right) \tag{6.4}$$

With these two definitions, we fulfil the orthogonality condition in having an orthonormal design matrix such that $\mathbb{X}^{\mathsf{T}}\mathbb{X} = \mathbb{I}$. Under this condition many quantities that we have previously encountered reduce to simpler forms, for example the Moore-Penrose pseudo-inverse simplifies to $\widehat{\boldsymbol{\alpha}} = \mathbb{X}^{\mathsf{T}}\boldsymbol{z}$ and the best model vector becomes $\widehat{\boldsymbol{\zeta}} = \mathbb{P}\boldsymbol{z} = \mathbb{X}\mathbb{X}^{\mathsf{T}}\boldsymbol{z}$, among others.

## 6.2 Generating Simulated Data

This far we have focused exclusively on aspects of *inference*, where data is given. We shall now explore a system for simulation of data. Chapter 3 detailed data and the levels thereof extensively. We encountered three different aspects pertaining to data as

1. Physical data and its origin.
2. The description of existing data
3. The numerical simulation of artificial data

For the purposes of generating simulated data, we are considering aspect (3) as the simulation of artificial data, where simulated data $\equiv$ artifical data. Moreover, we considered the concept of data from *additive error*, which for data $z_n$, was represented by

$$z_n = y(x_n) + a_n + \rho_n$$

where $a_n$ accounts for systematic distortion (which is the same under repeated experiments), although in simulation there is no consideration of systematic error. On the other hand, $\rho$ accounts for noise random variation (which is different under repeated experiments). Within all analysis of data, the underlying "true" law is always unknown — but not within the realm of simulation!

Beyond the inference involved with the analysis of data, when dealing with simulation our state of knowledge now includes *all aspects of data*, meaning that we *know* the variation which the "true" law experience with experiments, but most importantly, we *know* the "true" law, or rather the "true" function.

* This is particularly important if numerical precision has to be taken into account, in that precision may represent zero by a value of really small order, typically $10^{-20}$, and the identity matrix with these terms may not be seen in trivality as the identity matrix, and may thus not necessarily be trivially inverted.

### 6.2.1 The Construct

To avoid philosophical discussions regarding the nature of "truth", we shall henceforth refer to the "true" function as the *construct*, denoted by $\widetilde{\boldsymbol{z}}$. Therefore we may speak of simulated data generated according to the construct. We may consider the mathematics of specifying the construct as the same as which we have for specifying candidate models, i.e. design matrix, parametrisations, etc.

Of course, it is important to carefully distinguish between quantities of inference, such as the model order $K$, and equivalent quantities of simulation. We introduce the the quantities $C, C_{\min}$, and $C_{\max}$ as representative of the order of the construct (in simulation).

We assign to the construct the same definition as we had with the models, for $C \in [C_{\min}, C_{\max}]$, that is a linear system of parameters and basis functions,

$$\begin{aligned} \widetilde{z}(x_n \,|\, \widetilde{\boldsymbol{\alpha}}) &= \sum_{c=1}^{C} \widetilde{f}_c(x_n)\, \widetilde{\alpha}_c \\ \widetilde{\boldsymbol{z}} &= \widetilde{\mathbb{X}}\widetilde{\boldsymbol{\alpha}} \end{aligned} \tag{6.5}$$

The construct's design matrix, $\widetilde{\mathbb{X}}$, is of dimension $N \times C$, and the construct is paramtrised by $C$ parameters, $\widetilde{\boldsymbol{\alpha}}$, and so on. We maintain a distinction between $C, \widetilde{\boldsymbol{z}}, \widetilde{\mathbb{X}}, \widetilde{\boldsymbol{\alpha}}$ from their inference counterparts $K, \boldsymbol{z}, \mathbb{X}, \boldsymbol{\alpha}$.

In much the same way in how we have defined candidate models of differing orders (with the projector), we may define constructs of differing orders, and for constructs of orders $C \in [C_{\min}, C_{\max}]$ we have that,

$$\widetilde{\boldsymbol{z}}_C = \widetilde{\mathbb{X}}_C \widetilde{\boldsymbol{\alpha}}_C \tag{6.6}$$

We now have a set of $C_n = C_{\max} - C_{\min}$ constructs, from which we are able to produce a data set from *each* construct, i.e. we have would have $C_n$ distinct data sets. Since we are considering the data to be generating according to additive noise, we thus have

$$\boldsymbol{z}_C = \widetilde{\boldsymbol{z}}_C + \boldsymbol{\rho}, \tag{6.7}$$

where $\rho$ is sampled from $\mathcal{N}(0, \sigma_\rho^2)$. Having $C_n$ constructs, each of differing orders, we subsequently have $C_n$ different data sets, each generated by a construct of ascending order. We emphasise that $\boldsymbol{\rho}$ is generated anew for each of the $C_n$ data sets; this generates unique data for each $C$ and thereby avoids unwanted correlations between data sets generated with constructs of differing $C$. For the purposes of the inference part, each simulated data set is deemed immutable, as always.

### 6.2.2 Parametrising the Construct

The construct parameters $\widetilde{\boldsymbol{\alpha}}$ serve as a control of the signal-to-noise ratio. If the design matrix $\widetilde{\mathbb{X}}$ is defined according to either the cosine or the sine basis functions as in equations (6.3) and (6.4), then $\widetilde{\boldsymbol{\alpha}}$ can be thought of as an amplitude, which makes a signal strong or weak. Given that $\sigma_\rho$ is a constant, that also determines the signal-to-noise ratio.

To create a large variety of data, $\widetilde{\boldsymbol{\alpha}}$ itself is made to vary. De Kock and Eggers (2017a) and Liang et al. (2007) parameterise $\widetilde{\boldsymbol{\alpha}}$ according to a normal distribution, and we follow their example in spirit by proposing the following options for the construct parameters, with $\phi_C$ representing the random variable and $a$ and $b$ representing tuning parameters which determine whether the simulated data has either a weak or strong signal-to-noise ratio. The options investigated were

$$\text{Option A:} \quad \widetilde{\alpha} = a + b\phi_C \qquad \phi_C \sim \mathcal{N}(a, b^2\sigma_\phi^2) \tag{6.8}$$

$$\text{Option B:} \quad \widetilde{\alpha} = a\sigma_\phi + b\phi_C \qquad \phi_C \sim \mathcal{N}(a\sigma_\phi, b^2\sigma_\phi^2) \tag{6.9}$$

$$\text{Option C:} \quad \widetilde{\alpha} = a + b\frac{\phi_C}{\sigma_\phi} \qquad \phi_C \sim \mathcal{N}(a, b^2) \tag{6.10}$$

In assessing these options, we make use of dimensional analysis. To keep different concepts pertaining to $N$ and $C$ separate, we momentarily denote the dimension associated with $\phi$ by $D$ and that of the parameters $\widetilde{\alpha}$ by $P$. We also assign a "dimension" to the basis functions associated with the design matrix $\widetilde{X}$, even though they are often dimensionless. The "dimensions" of these three quantities are then constrained by $D = FP$, considering that

$$\dim(\phi_C) = \dim(\sigma_\phi) = \dim(\widetilde{\alpha}_C) = P$$
$$\dim(\rho_n) = \dim(\sigma_\rho) = \dim(z_n) = D$$

For dimensionless basis functions, the dimensions D and P are equal, and the scales of $\sigma_\phi$ and $\sigma_\rho$ can be compared directly. For option A in equation (6.8), control parameters $a$ and $b$ have different dimensions, which is undesirable in terms of comparison. Additionally, option C in equation (6.10) would leave $\widetilde{\alpha}_C$ dimensionless even while $\dim(\phi_C) = P$, therefore option B in equation (6.9) provides the only option in which both control parameters are dimensionless. Hence we settle on

$$\widetilde{\alpha}_C = a\sigma_\phi + b\phi_C \qquad \phi_C \sim \mathcal{N}(0, \sigma_\phi^2) \tag{6.11}$$

$$p(\widetilde{\alpha}_C) = \frac{1}{b\sigma_\phi\sqrt{2\pi}} \exp\left(-\frac{(\widetilde{\alpha}_c - a\sigma_\phi)^2}{2b^2\sigma_\phi^2}\right) \qquad \widetilde{\alpha}_C \sim \mathcal{N}(a\sigma_\phi, b^2\sigma_\phi^2) \tag{6.12}$$

Any simulated data generated according to equation (6.7) would then have the form

$$z_n = \sum_{c=1}^{C} \widetilde{f}_C(x_n)(a\sigma_\phi + b\phi_c) + \rho_n \tag{6.13}$$

Although expectation values of any real data may never be taken within the Bayesian framework*, however, within the context of simulated data, our state of knowledge permits expectation values, although this is not directly over the data, rather abstracted to the levels of randomness contained within $\widetilde{\alpha}_C$, i.e. $\phi_C$, as well as $\rho_n$.

Given that $E(\phi_C^2) = \sigma_\phi^2$ and $E(\phi_C) = 0$ for the signal and for the noise $E(\rho_n^2) = \sigma_\rho^2$ and $E(\rho_n) = 0$, we interpret expectation values over simulated data as

$$E(z_n) = \int \mathrm{d}\boldsymbol{\phi} \int \mathrm{d}\rho_n \, z_n(\boldsymbol{\phi}, \rho_n) \, p(\boldsymbol{\phi}) \, p(\rho_n)$$

* We consider the data to be immutable; the idea of integrating over all $\boldsymbol{z}$ is contrary to the idea of fixed data.

$$= a\sigma_\phi \sum_{c=1}^{C} \widetilde{f}_C(x_n) \tag{6.14}$$

$$E(z_n^2) = (a^2 + b^2)\,\sigma_\phi^2 \left[ \sum_{s,t=1}^{C} \widetilde{f}_s(x_n) f_t(x_n) \right] + \sigma_\rho^2 \tag{6.15}$$

$$\mathrm{var}(z_n) = b^2 \sigma_\phi^2 \left[ \sum_{s,t=1}^{C} \widetilde{f}_s(x_n) \widetilde{f}_t(x_n) \right] + \sigma_\rho^2 \tag{6.16}$$

Since we are simulating with orthonormal basis vectors, it follows that

$$E(\boldsymbol{z}^2) = \sum_{n=1}^{N} E(z_n^2) = C(a^2 + b^2)\sigma_\phi^2 + N\sigma_\rho^2 \tag{6.17}$$

If we consider the simple case when $\sigma_\phi = 1$, then all the options for the tuning parameters become identical and we recover the statement made by De Kock and Eggers (2017a, p. 12). In much the same way, the simplest assignment for the noise distribution is $\mathcal{N}(0, 1)$, where $\sigma_\rho = 1$.

In accordance with Liang et al. (2007, p. 19), higher amplitudes result in a strong signal-to-noise ratio, while lower amplitudes result in a weak signal to noise ratio. With the current specification, for any fixed $b$, we may tune the signal-to-noise ratio with $a$, and in doing so, we will use below the terms "weak signal" for $a = 1$ and "strong signal" for $a = 5$.

## 6.3 Modelling Simulated Data

Having established our understanding of the construct and the conceptual underpinning of data generated according to additive noise, with $\rho$, and also knowing that we model data according to additive noise*, w.r.t. $\varepsilon$ as previously formalised (as in sections 3.3 to 3.5), we are now interested in modelling any of the $C_n$ data sets, $\boldsymbol{z}_C$, with any of the available models, $\boldsymbol{\zeta}_K$.

It is essential to understand that the aim of modelling data is to correctly predict, by means of the model posterior $p(K \mid \boldsymbol{z}_C)$, the order of the construct, $C$, in each data set, and this is the importance of *why* we have $C_n$ distinct sets of data.

Of course, if we have $K_n$ models, then *each (!)* data set $\boldsymbol{z}_C$ is modelled by all (!) $K_n$ models. In other words, our analysis produces a $C_n \times K_n$ matrix of minimum chi-squared estimates (one for each model and for each respective data set). The minimum chi-squared criterion for any given model which models any given data set becomes,

$$\widehat{\chi}^2_{CK} = \boldsymbol{z}_C^2 - \widehat{\boldsymbol{\beta}}^2_{CK} \tag{6.18}$$

with which we are able to evaluate the respective information criteria against each model for all data sets.

### 6.3.1 Finite Accuracy in Predictions

We learned from chapter 3 that when overfitting data, the candidate model tends to regard noise in the data as signal and the chi-squared criterion tends to zero in that case, i.e. it models each

* Note that our state of knowledge includes all these principles and we know them to be true for the data that is to be studied.

data point (with $N$ parameters). With the established definition of the construct and the general idea behind generating simulated data, we may now get an idea in seeing how we are able to predict the construct by considering the accuracy in our prediction.

If we set the noise to zero in the simulated data such that ($\sigma_\rho = 0$), then $\boldsymbol{z} = \widetilde{\boldsymbol{z}}$, then we are modelling the construct *directly*, and with the understanding of the projection, we note that a construct of arbitrary order $C$ can be modelled as

$$\widehat{\boldsymbol{\zeta}} = \mathbb{P}\widetilde{\mathbb{X}}\widetilde{\boldsymbol{\alpha}} \tag{6.19}$$

The only way for any model, $\boldsymbol{\zeta}$, to model the construct accurately, irrespective of $C$, would be to assume that there exists some construct such that $\widetilde{\boldsymbol{z}} = \boldsymbol{\zeta} = \mathbb{X}\boldsymbol{\alpha}$. Therefore, there exists a candidate model of order $K$, which models the construct such that we have the following,

$$\begin{aligned} \boldsymbol{\zeta} &= \mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}(\mathbb{X}\alpha) \\ &= \mathbb{X}\boldsymbol{\alpha} \end{aligned} \tag{6.20}$$

What this means is that the only model which could ever return the construct is the one which models *all* data points, and *not* the model with $K = C$. Recall that this example models the construct *without* the presence of noise!

Of course, from a theoretical perspective the desired model is indeed the one in which the order of the candidate model matches the construct which modelled the data: $K = C$, and for this, the importance of model selection via evidence and information criteria is only reinforced.

### 6.3.2 Model Evaluation

With knowledge of the construct, we follow Liang et al. (2007, p. 19) in determining the theoretical squared error between the construct and the model, with each construct considered against each respective model. The squared error loss between construct and model is defined as,

$$SE_{CK} = \|\widetilde{\boldsymbol{z}}_C - \boldsymbol{\zeta}_K\|^2 \tag{6.21}$$

Given that every quantity has been set up as a matrix, this also yields a matrix of dimension $C_n \times K_n$, and we consider the models evaluated by information criteria against the squared error, for which a lower squared error ensures a better result. When the order of the candidate model matches the order of the construct, $C = K$, the candidate would be considered to be the $ideal$ prediction of the construct when modelling the data, so these are the ideal estimates when evaluating a model. This ideal is known as the *Oracle*, which tells when a prediction may be correct (or true).

The Oracle estimates lie on the main diagonal when $C_n = K_n$. For cases when $C_n \neq K_n$, we consider two cases. The first case occurs when the orders of all $C_n$ constructs are contained within the orders of the $K_n$ models and the second when the orders of the constructs are not fully contained within those of the models.

For the first case we would have that $K_{\min} < C_{\min}$ and $C < K$; as such the Oracle occurs on the $C_{\min}$-superdiagonal. This generalises to whenever $C_{\min} < K$. For the second case we consider $C_{\min} < K_{\min} < C_{\max}$, which will yield an Oracle on the $C_{\min}$-subdiagonal.


(i) $C \in \{6,7,8\}$ and $K \in \{5,6,7,8,9\}$ (ii) $C \in \{1,2,3\}$ and $K \in \{3,4,5,6,7\}$

Figure 6.1: *SE* for data generated according to constructs of order $C$ and modelled according to models of order $K$. Oracle estimates are indicated in red, either on the sub- or superdiagonal.

If there are no instances in which $C = K$, then there is nothing to compare the squared error result to, since such an analysis would have no Oracle. For the sake of our current study, we shall consider the case in which $C_{\min} = K_{\min}$ and $C_{\max} = K_{\max}$, thus resulting in a square matrix, since $C_n = K_n$, which in turn eliminates the complication of modelling data without having a baseline estimate for the evaluated models. For this we have a squared error assignment for each model against which we assess their respective information criteria.

In addition, it is important to note that the squared error analysis can be performed for any system where our state of knowledge provides enough information on $\widetilde{z}$, such as in simulation, the Oracle estimate is only useful in cases when $\widetilde{\mathbb{X}} = \mathbb{X}$. Therefore we define our construct according to the same basis functions as the model.

The procedure detailed up to this point outlines *one complete simulation* in which $K_n$ models (each of ascending order) are applied to $C_n$ datasets (each generated according to constructs of ascending order). Models are assessed in terms of the available information criteria, based and weighed against the theoretical squared error as given by equation (6.21). We consider the *mean squared error* (MSE) for $T$ trials of each possible model-dataset pair, i.e. we perform $T$ complete simulations in total.

$$MSE_{CK} = \langle SE_{CK} \rangle = \frac{1}{T} \sum_{t=1}^{T} (SE_{CK})_t \tag{6.22}$$

In summary, what is achieved with this matrix formulation is the ability to analyse $C_n$ distinct data sets with $K_n$ distinct models in one fell swoop!

Note that the matrix formulation developed here differs from past occurrences in literature in that we have generalised the square matrix system, with $C_n = K_n$, to any arbitrary matrix system of $C_n \times K_n$; furthermore, we make the difference between $C$ and $K$ explicit, which in turn provides us with freedom to perform inferences on systems where $\widetilde{\mathbb{X}} \neq \mathbb{X}$.

## 6.4 Simulation Study

### 6.4.1 Initialising the Simulation

Our simulation was performed with the following settings:

| $N$ | $T$ | $C_{\min}$ | $C_{\max}$ | $K_{\min}$ | $K_{\max}$ | $f_k(x)$ |
|---|---|---|---|---|---|---|
| 100 | 1000 | 1 | 25 | 1 | 25 | cosine |

Simply stated, we generate 25 unique sets of data, each of 100 points, model each of the 25 data sets with 25 models, perform the analysis and repeat the procedure over 1000 trials. We shall consider two cases of the simulation with respect to the signal to noise ratio, both strong and weak signal to noise ratios, with $a = 5$ and $a = 1$ respectively. At present, we set $\Delta = 1$ in equations (5.37) and (5.40), since we cannot make a meaningful assignment at present, the scale of $\Delta$ in particular, is unknown, therefore $\Delta = 1$ remains as the choice at present.

### 6.4.2 The Central $r$-prior: Investigating the Radial Scale & the Overarching Model

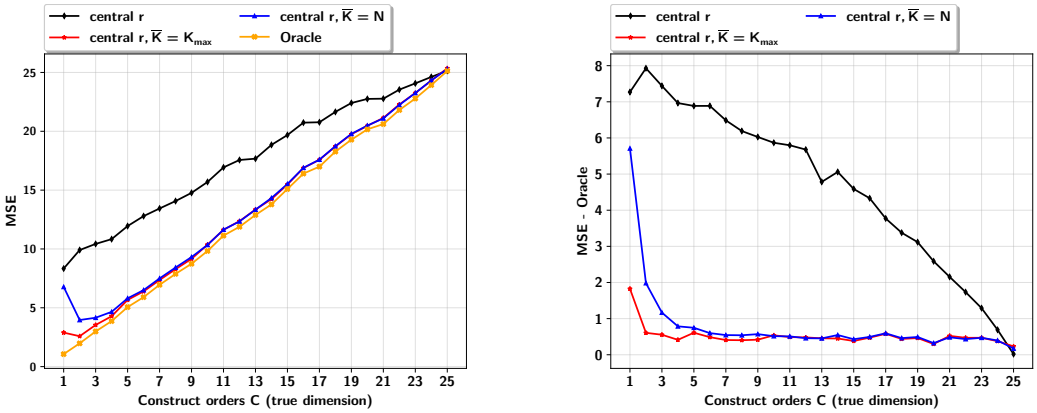
We first turn to the investigation of the effect that the radial scale factor $\ell$ and the overarching model order $\overline{K}$ may have on the central information criterion that resulted from the evidence of equation (5.37). With the theory for these quantities outlined in section 5.3, we have considered the ideal representation of the $\ell = \sqrt{K}$, while the overarching model's order, $\overline{K}$ was set to $\overline{K} = K_{\max}$ at minimum (recall that we choose it to be similar to a possible K), since all $K$ form subspaces of $K_{\max}$.

Alternative choices, such as $\ell = \{K^0, K^{1/2}, K^1, \ldots\}$ and $\overline{K} = K_{\max} + 1, \ldots, N$, also form part of the investigation and analysis.

In testing *any* of these possibilities, we compared the results with those of the original as per De Kock and Eggers (2017b), for which (in our framework) $\ell = 1$ and $\overline{K} = K$. We refer to the De Kock and Eggers (2017b) settings as the *classical* central information criterion. The best combination between $\ell$ and $\overline{K}$ shall be chosen as the desired central information criterion.

#### The Overarching Model Order

We first consider the case when $\ell = 1$ throughout, since this allows us to investigate $\overline{K}$ without the influence of the radial rescale. We first consider the postulated quantities of $\overline{K}$, namely $K$ and $N$,



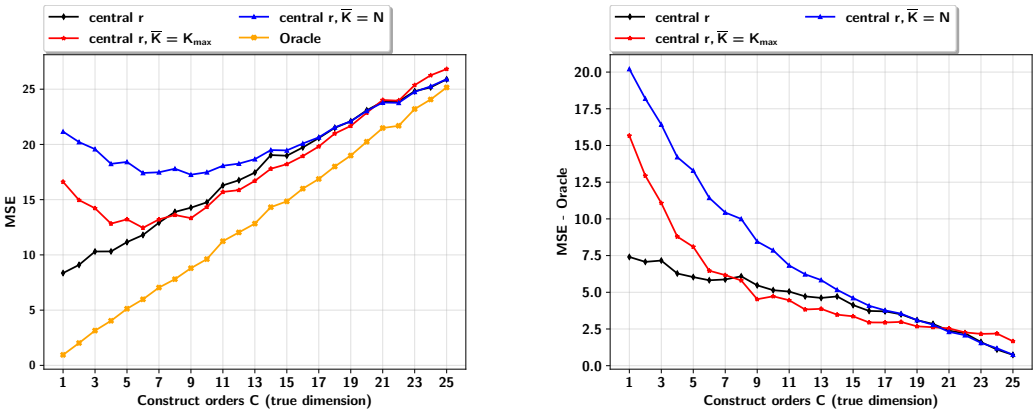

(i) MSE of central information criteria

(ii) Difference between MSE and Oracle

Figure 6.2: Investigating $\overline{K}$: strong signal (lower is better), with $\ell = 1$ for all criteria.

Figure 6.2(i) illustrates that setting $\overline{K}$ separately from the model's $K$ results in a substantial performance improvement. We also see that the overarching model order of $K$ initially outperforms that of $N$, and partial convergence is achieved for predicting constructs of higher orders. Figure 6.2(ii) Illustrates a more detailed version of figure 6.2(i) by subtracting the Oracle estimates

from the MSE.




(i) MSE of central information criteria (ii) Difference between MSE and Oracle

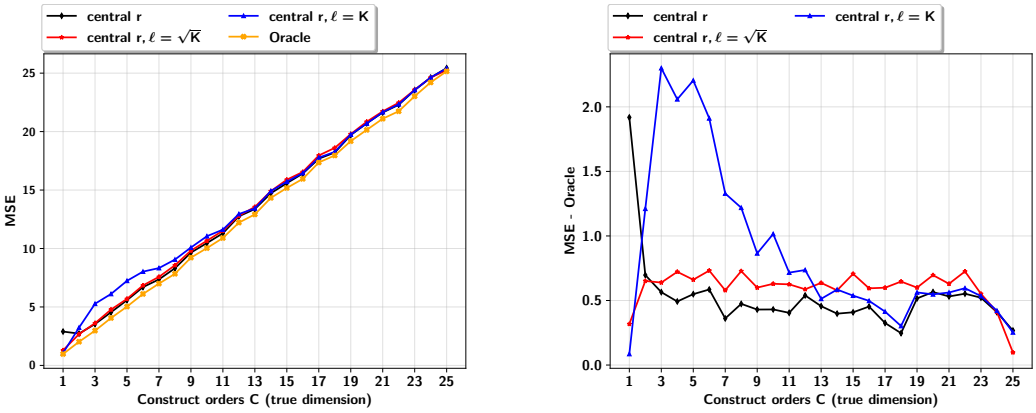
Figure 6.3: Investigating $\overline{K}$: weak signal (lower is better), with $\ell = 1$ for all criteria.

As illustrated in figure 6.3, under weak signal analysis under the same condition, the classical criterion is better than $\overline{K} = N$ but a little worse than $\overline{K} = K_{\max}$. When predicting constructs of higher orders, the $N$-based criterion converges to the classical and the $K$-based criterion falls marginally short in comparison, even though being the best midway.

Similar analyses were performed for alternate postulates of $\overline{K}$, which include both indexed and fixed orders, and the results which returned the best performance overall, are those that were initially postulated, and displayed by figures 6.2 and 6.3. As such, we settle on the overarching model order of $\overline{K} = K_{\max}$, which fixes the order to the order of the maximum model in our set of candidate models. This choice is based on the achieved results in which it was overall the most consistent between the regimes of weak and strong signal to noise ratios.

### The Radial Scale Factor

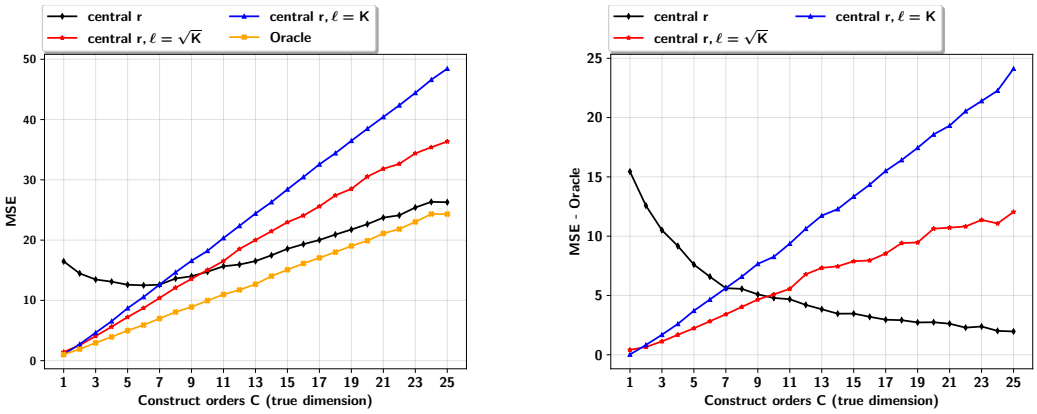
Having established the overarching model order, we no longer use the classical criterion as our baseline but rather the case with $\overline{K} = K_{\max}$. Setting $\ell = 1$ would provide us with the overarching criterion of the previous section, which we know outperforms the classical central criterion. Therefore we are now searching for the scale factor which will improve on the performance of the overarching criterion. Once again, we begin with the postulated scales of $\ell = \sqrt{K}$ and $\ell = K$,




(i) MSE of central information criteria (ii) Difference between MSE and Oracle

Figure 6.4: Investigating $\ell$: strong signal (lower is better), with $\overline{K} = K_{\max}$ for all criteria.

For strong signal, the performance when $\ell = 1$ is evidently better than either $\ell = \sqrt{K}$ or $\ell = K$ over nearly all 25 sets of data, however it is only marginally so compared to $\ell = \sqrt{K}$ which shows more stability in comparison (particularly for lower $C$). $\ell = K$ displays poor performance as well as instability relative to the other two criteria.




(i) MSE of central information criteria

(ii) Difference between MSE and Oracle

Figure 6.5: Investigating $\ell$: weak signal (lower is better), with $\overline{K} = K_{\max}$ for all criteria.

For weak signal $\ell = \sqrt{K}$ provides the best performance for low $C$, but as $C$ increases, the performance of the $\ell = 1$ case improves, for which it has the best performance for higher $C$.

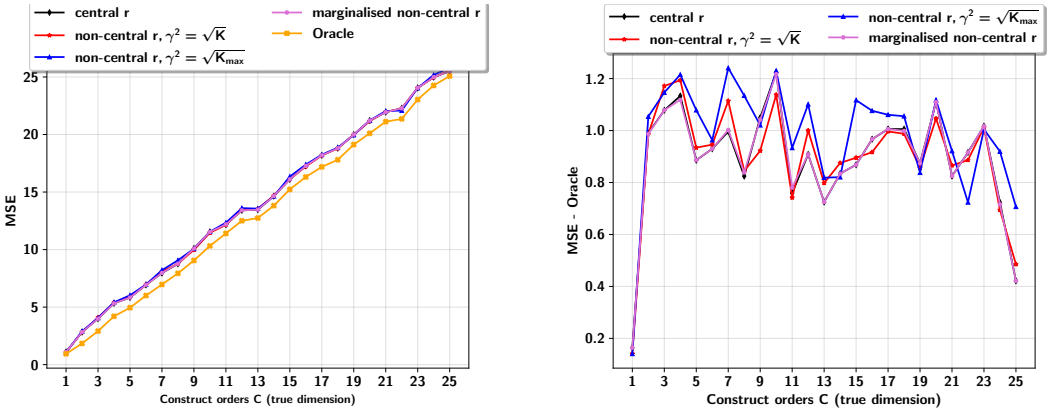
Various regimes were tested, which include both indexed and fixed scenarios for $\ell$, such as indexed assignments including $\{K, K^2, \ldots\}$ and fixed assignments including $\{K, N\}$. In each case the results were similar to figures 6.4 and 6.5, in which the $\ell = 1$ criterion had the best performance for weak signal while the $\sqrt{K}$ criterion had the most stability.

Here the result is not as clear-cut as with the overarching model order; we obtain varying results for any given radial scale $\ell$ depending on the regime of weak or strong signal-to-noise ratio. However, we choose $\ell = \sqrt{K}$ based on the stability which it provides over the unit $\ell$. We shall henceforth consider any central criterion as scaled by $\ell = \sqrt{K}$ and with overarching model order $\overline{K} = K_{\max}$.

### 6.4.3 The Non-central $r$-prior: Investigating the Location Parameter

For the non-central information criterion based on the evidence of equation (5.40), we consider all the previous results to be *inherited* from the best case central criterion, therefore $\ell = \sqrt{K}$ and $\overline{K} = K_{\max}$.

The first issue that we considered was $\boldsymbol{\gamma}^2$ as a location parameter and the second was whether $\sqrt{K}$ may be a suitable location in terms of how this relates to the maximum likelihood in spherical space. In reference to a known result, we consider this to be the information criterion of the non-central evidence in which $\boldsymbol{\gamma}^2$ is marginalised, as in equation (5.42) and per De Kock and Eggers (2017a), which we refer to as the marginalised non-central criterion.

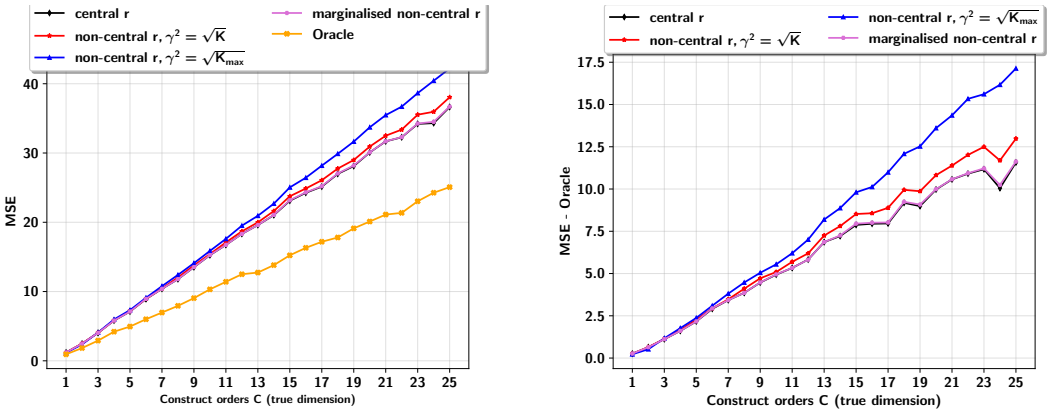



(i) MSE of central information criteria

(ii) Difference between MSE and Oracle

Figure 6.6: Investigating $\gamma^2$: strong signal (lower is better)

From figure 6.6, we first note that the marginalised non-central criterion closely resembles the central criterion, while the best overall performance is achieved by $\gamma^2 = K$ more so than $\gamma^2 = \sqrt{K}$, but both of the latter outperform the former.




(i) MSE of central information criteria

(ii) Difference between MSE and Oracle

Figure 6.7: Investigating $\gamma^2$: weak signal (lower is better)

For the weak signal shown in figure 6.7, we once again have that the marginalised non-central criterion performs similarly to the central criterion, while the criterion with $\gamma^2 = \sqrt{K}$ performs marginally worse for higher order $C$, and the criterion with $\gamma^2 = K$ diverges fairly early in comparison, with relatively poor performance for higher order $C$.

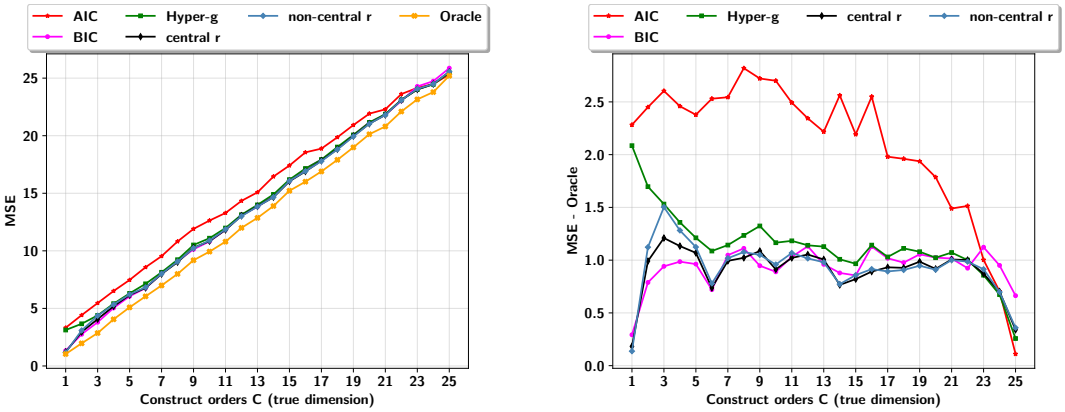
The fact that the marginalised non-central criterion is so similar in performance to that of the central criterion, for both weak and strong signal cases, is a real validation to the effect of the radial scaling and the overarching model (and the choices made thereof), because we have managed to match the performance of the modern (non-central) criterion by introducing these quantities to the classical (central) criterion. In addition to these cases, an investigation was launched for the case $\gamma^2 = \sqrt{K_{\max}}$, with relation to the overarching model order, which was convergent to the indexed case in figures 6.6 and 6.7.

For this reason we no longer need to consider the marginalised non-central criterion, instead we shall settle on the non-central criterion of $\gamma^2 = \sqrt{K}$, since it provides decent performance of the investigated scenarios for the non-central criterion, while still differing enough from the

central criterion — and we desire this difference. We shall henceforth consider any non-central result to have the setting $\gamma^2 = \sqrt{K}$.

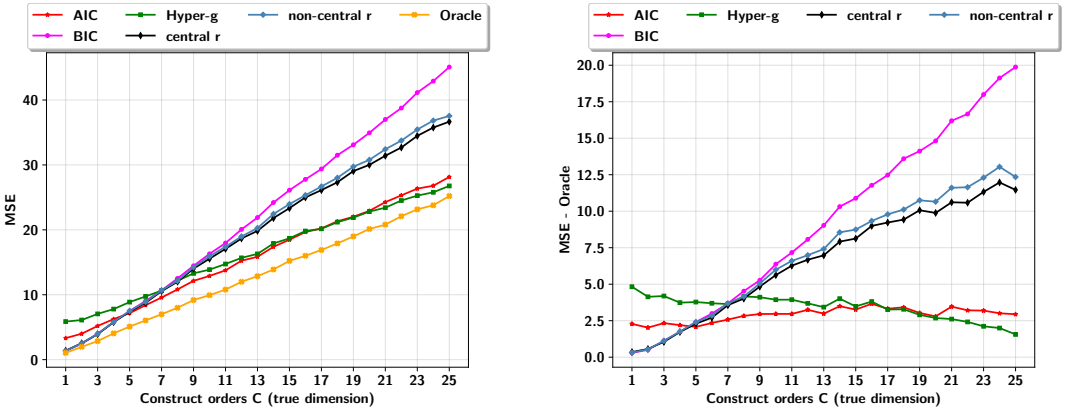
## 6.5 Comparative Results Versus Known Information Criteria

With the established criteria for the central and the non-central information criteria in context of the introduced framework of radial scaling, overarching model order, and the non-central location, we now study the performance of the central and non-central information criteria in comparison to some well-known information criteria, namely AIC, BIC and the hyper-$g$.




(i) MSE of information criteria (ii) Difference between MSE and Oracle

Figure 6.8: Strong signal performance of various information criteria (lower is better).




(i) MSE of information criteria (ii) Difference between MSE and Oracle

Figure 6.9: Weak signal performance of various information criteria (lower is better).

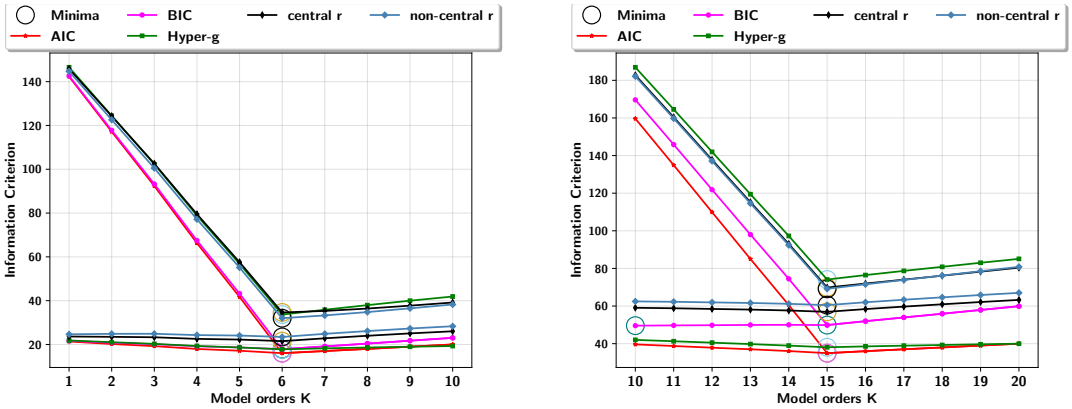
The performance of the non-central and the central criteria are similar in both strong and weak signal regimes, with the non-central criterion marginally outperforming the central within the strong signal regime, while the central criterion marginally outperforms the non-central in the weak signal regime. This is the case over all constructed data sets, except for data generated by a construct of lower order in the weak signal regime, in which they are similar.

Compared to the performance of the other information criteria, within the strong signal regime both offer better performance than than all criteria, while also being on par with the BIC, and offering better performance for certian data sets over the midrange. For the weak signal regime,

however, this is inverted: performance of the BIC and both central and non-central criteria worsens with data of ascending construct order.

Therefore in each case the behaviour of both the central and non-central criteria follow a trend similar to that of the BIC, with significantly better performance in the weak signal regime.

In order to explore the behaviour between regimes in more detail, we consider the performance of the criteria over a *single* data set with fixed $C$ for both signal strength regimes as a function of $K$. In doing so we consider the candidate models of orders in a range surrounding the order of the construct which generated the data to provide a detailed view*.




(i) Data generated by construct of order 6, modelled by models of orders $K \in [1, 10]$.

(ii) Data generated by construct of order 15, modelled by models of orders $K \in [10, 20]$.

Figure 6.10: Mean progressions of the information criteria for specific data sets.

In both figures 6.10(i) and 6.10(ii) the upper curves result from the strong signal regime, while the lower curves result from the weak signal regime. The best model (for the given fixed $C$) is the one at which each curve reaches its minimum.

We note that within both cases AIC and BIC converge once $K \geq C$, meaning that it makes no distinction of the signal-to-noise ratio. The hyper-$g$ as well as the central and non-central criteria maintain the distinction between weak and strong signal.

The distinction is important since we hope to maintain accuracy in predictions *irrespective* of the regime, which may be unknown, because we then have a framework which may provide meaningful results irrespective of the regime, and when considering real data, the regime of strong or weak signal-to-noise ratios is unknown. This is unlike the BIC, which underfits in figure 6.10(ii) for weak signal and loses quality in predictions. Therefore the verification that the distinction is maintained, as with the hyper-$g$ criterion, is a good indication of the assessments made by the $r$-criteria, both central and non-central.

* The low number of candidate models aids in the readability, since the models of order not near that of the construct provide little useful information within this context.

# Conclusion

**The Bayesian Theory of Probability**

We provided an introduction to systems of deductive and inductive reasoning in the respective forms of propositional logic and uncertainty logic, after which uncertainty logic was formally investigated and shown to be described by the framework of Bayesian probability theory. This was proven (in full) on the grounds of a new formalism as accounted in appendix A.

**Data Modelling & Model Comparison**

The setting in which our primary research questions arise was developed by introducing the concepts behind data and data modelling, for which we modelled systems of data via least squares and linear regression. We showed that the fundamental idea behind modelling data requires greater insights than are provided by these methods and the chi-squared criterion.

These insights are found, in both, information theory (in the form of information criteria), and within the framework of Bayesian probability theory. Both formalisms aim to solve the primary dilemma in data modelling, namely knowing which model best describes a given set of data.

**Spherical Symmetry**

The framework of spherical symmetry was employed within our setting of data modelling and model comparison, for which a generalised projection onto a spherically symmetric surface was developed. In addition, critical insights into radial scales between model order and respective radii were obtained. Furthermore, it was shown and conceptualised how models may be enveloped by an overarching model for which an overarching radius spanning all models exists. These insights were incorporated into the framework and shown to be an essential part of $r$-priors and their resulting evidences. These insights are shown to have extended the existing body of work on spherical symmetry.

**The Radial Scale and the Overarching Model**

The principal new ideas introduced into the framework of spherical symmetry were the radial scale factor $\ell$, and the concept of the overarching model, leading to the parameter $\overline{K}$ (the overarching model order). These two ideas provided us with a means of tuning the information criteria in ways that had not been achieved before.

Investigations into how $\overline{K}$ affected the performance of information criteria with regards to model selection provided insights which showed that this new parameter delivered improvements to what had already been established in the literature of $r$-priors.

The radial scale factor, on the other hand, induces some variation between the strong- and weak-signal regimes. This is a hint that knowledge of the signal strength may influence the performance of the criteria based on the radial scale. For this reason, we know that there is a need for more information with regards to the scope and effects of the radial scale factor as well as potential choices thereof. The variation with signal strength provides a hint that the radial scale may best be left indeterminate and marginalised from the evidence, since our state of knowledge may be inadequate in an appropriate assignment for $\ell$ without reference to the signal-to-noise ratio — however this is still open to investigation.

Regarding the effectiveness of introducing a non-centrality (location) parameter $\boldsymbol{\gamma}$, we found

that the non-central information criterion performs on par with the central information criterion, with stability in performance for both strong- and weak-signal regimes in our assignment of $\gamma^2$. Moreover, we discovered that when marginalised from the evidence, the marginalised non-central and the central information criteria provide remarkably similar results, therefore we relax the necessity on marginalising it from the evidence, as was the case in established literature.

## Future Outlook

One natural extension to the formalism for both the central and the non-central criteria would be that which was introduced by De Kock and Eggers (2017a), namely considering an extended model with a noise model which parametrises the noise as an extension to the candidate model. Known results involving this extended model indicate an improvement over the performance of the current non-central information criterion, particularly in the weak signal regime.

This is of course on the basis that we know and understand the current base model, that results based on the marginalised non-central information criterion of the past have been matched or exceeded as indicated in this study.

Therefore, provided with adequate improvements to the current state of knowledge regarding the radial scaleindexScale!Radial scale, the extended model may improve with the introduction of both the overarching model as well as the radial scale factor which were introduced within this study.

Specifically, the formalism of the radial scale factor and the overarching model would find application to the noise space models, thus introducing *two* unique radial scales and overarching models — one pair in model space and one pair in noise space.

In addition to these advancements, we are inclined to explore two facets of the prior, first the parameter prior, and second the $r$-prior. With the generalised constraint on a prior projected onto a spherically symmetric surface we have greater freedom in choice of prior, therefore we have greater freedom in the outcome of the $r$-likelihood.

When constructing the $r$-prior, we assigned it through a hyper parametrisation, however, we know that we are conditioned on the prior scaling according to $r^{K-1}$, and with this knowledge we also have a certain choices for suitable $r$-priors.

These two prior considerations may lead to noteworthy investigations regarding the framework of spherical symmetry.

APPENDIX A

# Foundations of Bayesian Probability Theory

## A.1 Axioms of Uncertainty Logic

This section aims to formalise the framework of uncertainty logic, therefore the exposition is technical and mathematically demanding. The material presented here was sourced and synthesised from (Jaynes 2003), (Cox 1961) and (Cox 1946).
Deviating from the formalism highlighted by Jaynes (2003) and many others, who opt for a functional justification of proposed axioms based on the desiderata, we follow guidelines from the formalism of Terenin and Draper (2017) who propose axioms justified by the desiderata, which are in turn used to produce the foundation for the framework of uncertainty logic. However, unlike their formalism which is asserted through set theory, we detail an analytic formalism within the framework of propositional logic and assert its equivalence to the functional formalism.

First, we begin by constructing a set of suitable axioms,

> Axiom A.1: Real number representation
> States of uncertainty are represented by real numbers.

Axiom A.1 follows directly from our first desideratum. The implication is that we now have a means of ordering our convictions in states of uncertanty, since real numbers are transitive. Following this ordering, we consider "more believable/reasonable/plausible" to represent *our conviction* in a state of uncertainty — meaning that we can represent greater conviction with a greater number and lesser conviction with a lesser number. In other words, the more convinced we are, the greater the outcome of an evaluated state of uncertainty.

Suppose we have propositions $w$, $y$, and $z$ with states of uncertainty in each, $\psi(w)$, $\psi(y)$, and $\psi(z)$, then if $\psi(w)$ is more convincing than $\psi(y)$, and $\psi(y)$ is more convincing than $\psi(z)$, it follows that the convictions in our states of uncertainty may be ordered as $\psi(w) \geq \psi(y) \geq \psi(z)$.

Real numbers also possess the property of *density*, which means that there *always* exists a number between any two numbers. Therefore, if we introduce a new proposition, $x$, with the state of uncertainy $\psi(x)$, of which this state may, for example, be more convincing than $\psi(y)$ but less than $\psi(w)$, then it naturally implies that $\psi(w) \geq \psi(x) \geq \psi(y) \geq \psi(z)$. This is true for any collection of propositions. Furthermore, this extends the reasoning of the second desideratum, providing a representation that small changes in the knowledge of a proposition may indeed yield small changes in its state of uncertainty.

Additionally, we may extend axiom A.1 by asserting postulates* which aim to formalise and expand upon the axiom,

> Postulate A.1: Logical propositions exist within a field
> Let $\Omega$ be a space of logical propositions and let $\mathbb{F}$ be a field with $\sigma$-algebra on $\Omega$.

* The terms postulate and axiom are often used interchangeably, however we consider postulates to be a slightly weaker form of an axiom, because we consider positing a postulate to depend on having established an axiom or axioms.

Here $\Omega$ is a space containing propositions in the field $\mathbb{F}$ under $\sigma$-algebra. We consider the field to be consistent with the operators `AND` as well as `OR`. This algebra implies that for any element contained in $\Omega$, its negation is also contained within $\mathbb{F}$, therefore `NOT` is included in this system, meaning that our entire framework of propositional logic is consistent within this setting. This algebra is also closed under countable disjunctions*, meaning that if countably many propositions exist in $\mathbb{F}$, then their disjunctions also exist in $\mathbb{F}$.

Note that $\sigma$-algebra is usually defined as an algebra of sets, but we are able to formulate it in terms of propositional logic given that we have highlighted the equivalences of set theory and propositional logic; both are underpinned by Boolean algebra. In the context of propositional logic, the space $\Omega$ is not a necessary requirement, since such a collection of elementary propositions is not always necessary — however, we impose that propositional logic always comprises of the two trivial propositions, namely 1 (true) and 0 (false), therefore this space may always be constructed within this framework of logical reasoning.

> Postulate A.2: Measure on $\mathbb{F}$
>
> Let $\psi$ be a measure on $\mathbb{F}$ such that $\psi : \mathbb{F} \times (\mathbb{F} \setminus 0) \to R \subseteq \mathbb{R}$, with representation $\psi(x \mid y)$ for any $x \in \mathbb{F}$ and $y \in (\mathbb{F} \setminus 0)$.

Note that axiom A.1 is extended by postulates A.1 and A.2. Following this we may state that our convictions in states of uncertainty in propositions are represented by the set of real numbers, $\mathbb{R}$, or the subset $R$. Furthermore, note that $y \in (\mathbb{F} \setminus 0)$, implying that when considering the state of uncertainty in a proposition given another, then the given proposition is in a *known* state of truth — hence the semifield $(\mathbb{F} \setminus 0)$ excludes the false proposition.

> Axiom A.2: Sequential continuity
>
> Conviction in states of uncertainty may tend to an assurity.

Axiom A.2 follows from the second desideratum, noting that this may provide a system in which small changes in the knowledge of a proposition yield small changes in its state of uncertainty. This axiom may be trivially justified within the context of set theory, since set theory provides an inherent understanding of ordering in the form of subsets. Propositional logic has no such structure, therefore the statement made by this axiom requires more care within our framework.

Following axiom A.1 we noted that we *now* have a means of ordering, in other words, it is essential to note that ordering only follows once we have established that we are concerned with our conviction in a state of uncertainty — propositions themselves have no ordering. It is crucial to note that it is our *conviction* in a state of uncertainty that enables us to deduce what may be considered "believable". In other words, given different constructs of knowledge, states of uncertainty may have different orderings of plausibility for *the same* proposition.

> Postulate A.3: Ordering in states of uncertainty
>
> For any collection of known propositions $y_1, y_2, \ldots, y_K$ our conviction in the state of

* Set theory refers to countable unions.

> uncertainty in the proposition $x$ depends on the conditioned information provided by any given proposition. Therefore, if $y_{k+1}$ is more descriptive than $y_k$, we assert that
> $$\psi(x \mid y_1) \leq \psi(x \mid y_2) \leq \cdots \leq \psi(x \mid y_K)$$
> for any successive increase in conviction given new knowledge/information.

We note that following axiom A.2 and postulate A.3, we are able to make a statement on our conviction in the state of uncertainty of a proposition given the reverse ordering, meaning that any successive decrease in our conviction, given information, would eventually tend to reduce our conviction in the state of uncertainty to zero — making it implausible and not worth any consideration. This implies consistency according to the third desideratum, and additionally, implies continuity at zero: $\psi(x \mid y_1) \to 0$.

It is noteworthy to observe that conviction in a state of uncertainty need not depend on the proposition being the same given different states of knowledge; states of uncertainty may be ordered even when concerned with different propositions, however the states of uncertainty are then required to be applicable to the same logical conclusion.

> Postulate A.4: Falsities in uncertainty logic
> Following postulates A.2 and A.3, and the consequences thereof, we may assert that the state of uncertainty equivalent to the false proposition is, in fact, $0 \in R$.

This postulate invokes the third desideratum, while following the logical progression of the aforementioned axioms and postulates. Consequently, we have that $R$ has a *well-defined* lower bound of zero.

> Axiom A.3: Decomposibility
> For some operator $\circ : R \times R \to R$, $\psi(x, y \mid z)$ can be expressed as $\psi(x \mid z) \circ \psi(y \mid x, z)$.

> Axiom A.4: Negation
> There exists a function $N : R \to R$ such that $\psi(\overline{x}) = N\big(\psi(x)\big)$, for all $x \in \mathbb{F}$.

Axiom A.3 follows from the third desideratum, since consistency with propositonal logic implies that compound logical propositions should decompose into simpler ones. Similarly, axiom A.4 also follows from the third desideratum, and in terms of states of uncertainty any proposition being mapped to its negation means that knowledge on the conviction of a proposition's state of uncertainty implies knowledge on the conviction of the state of uncertainty of it's negation.

Furthermore, following this reasoning, we may assert the existence of an upper bound on $R$; this is evident given that we know of the existence of a lower bound — meaning that there exists a negation to the falsity (lower bound), namely the truism. We may thus posit the following,

> Postulate A.5: Triusm in uncertainty logic
> There exists some upper bound in $R$ for $\psi$, such that the negation of the falsity is that

upper bound: $\psi(\overline{0}) = t$, where $t$ represents the truism in uncertainty logic.

The assertion of this postulate is in accordance with the third desideratum, since certainty has to comply with (known) truth. Therefore, as the conviction in a state of uncertainty increases to the point of absolute certainty, then there can only be one correspondence in our state of uncertainty, namely that of truth (where we have the logical equivalence of $t \equiv$ true).

Intuitively, this postulate is necessary when considering conviction in a state of uncertainty to be definitive (meaning no uncertainty), i.e. it has to be definite, thus the representation cannot be that of an infinity, since it is not possible to have infinite conviction on a state of uncertainty (let alone infinite conviction on anything), bearing in mind that conviction is conditioned on knowledge/information.

As a result of postulates A.4 and A.5, we have two invariants in uncertainy logic, both of which correspond to the two invariants of propositional logic: true and false. Furthermore, we may represent our convictions in these two states of uncertainty as follows,

$$\psi(x \,|\, x) = t \tag{A.1}$$

$$\psi(x \,|\, \overline{x}) = 0 \tag{A.2}$$

The first equation states that our conviction in a state of uncertainty in a proposition, given that knowledge about the proposition is known (i.e. true), is a certainty (which thus maps to the logical certainty). The second equation states that our conviction in a state of uncertainty of a proposition, given that knowledge about the negation of that proposition is known, has to be false (which thus maps to the logical falsity).

Axiom A.5: Consistency under extension

If $(\Omega,\, \mathbb{F},\, \psi)$ satisfies axioms A.1 to A.4 as well as postulates A.1 to A.3, then $(\Omega \times \Omega,\, \mathbb{F} \otimes \mathbb{F},\, \psi \circ \psi)$ must as well, meaning that $\psi(w, y \,|\, x, z) = \psi(w \,|\, x) \circ \psi(y \,|\, z)$ is consistent.

This axiom also follows from the third desideratum, and since propositional logic is consistent under extension, we (naturally) expect uncertainty logic to maintain this consistency*.

Postulate A.6: Structure of uncertainty logic

If $(\Omega,\, \mathbb{F},\, \psi,\, R,\, \circ,\, N)$ satisfies axioms A.1 to A.5 and postulates A.1 to A.3, then this is the algebraic structure underlying uncertainty logic.

Having established the underlying structure for the framework of uncertainty logic, we may now consider investigating some necessary details and properties regarding the nature hereof. In all cases, consider the conviction in any state of uncertainty to be non-trivial, unless stated otherwise.

* The justification here succeeds that of Terenin and Draper (2017), since consistency under extension follows from the framework of propositional logic and the desiderata, while they assert it through reasoning of repetitions.

Lemma A.1: The operation $\circ$ is monotonic

*Proof*

For any propositions $x$ or $z$, the conviction in a state of uncertainty in either proposition, given information of the form of known logical propositions $y_1, y_2, \ldots, y_K$, is, in accordance with axiom A.2 (continuity) and postulate A.3 (states may be ordered), confirmed to conform to transitive ordering.

For any $y_i$ in this collection, axiom A.3 (decomposibility) asserts that $\psi(x, z \,|\, y_i) = \psi(x \,|\, y_i) \circ \psi(z \,|\, x, y_i)$. Given that we have established that $\psi(x \,|\, y_1) \leq \psi(x \,|\, y_2) \leq \cdots \leq \psi(x \,|\, y_K)$, the same ordering, evidently, applies to the quantity $\psi(z \,|\, x, y_i)$. Hence, it follows that the operation $\circ$ is monotonic in this setting.

Consider the case in which we apply axiom A.5 (consistency under extension), for which we have $\psi(x, z \,|\, y_i, y_j) = \psi(x \,|\, y_i) \circ \psi(z \,|\, y_j)$. We have that monotonicity already holds for both, $\psi(x \,|\, y_i)$ and $\psi(z \,|\, y_j)$, hence the operation $\circ$ is also monotonic within this setting. Consequently, we conclude that the operation is monotonic in both arguments. □

Lemma A.2: The operation $\circ$ is cancellative

Cancellativity implies that if $\psi(x) \circ \psi(y) = \psi(x) \circ \psi(z)$ then $\psi(y) = \psi(z)$, and also if $\psi(x) \circ \psi(y) = \psi(z) \circ \psi(y)$ then $\psi(x) = \psi(z)$.

*Proof*

Consider $\psi(x, z \,|\, y) = \psi(x \,|\, y) \circ \psi(z \,|\, x, y) = \psi(z \,|\, y) \circ \psi(x \,|\, z, y)$, then if our convictions in the states of uncertainty in $x$ and $z$ are such that $\psi(x \,|\, y) = \psi(z \,|\, y)$, with $\circ$ being monotonic, it implies that, $\psi(z \,|\, x, y) = \psi(x \,|\, z, y)$. The alternative follows by the same construction. □

Lemma A.3: The $\circ$ operator has a unique identity element $e \in R$

*Proof*

Concerning convictions any state of uncertainty, the existence of an identity element, $e \in R$, implies that $e \circ \psi(x) = \psi(x) \circ e = \psi(x)$. Suppose that there exists an element $f \in R$ such that $f \circ \psi(x) = \psi(x) \circ f = \psi(x)$, then by lemma A.2 (cancellativity) we have that

$$e \circ \psi(x) = f \circ \psi(x)$$
$$e \overset{!}{=} f$$

Lemma A.2 also holds for operating from the left. □

Lemma A.4: Associativity

The operation $\circ$ is associative.

*Proof*

Logical propositions are associative with respect to `AND`, however, there are two scenarios in which we need to consider the operation $\circ$, the first is that of axiom A.3 (decomposibiliy): be-

haviour under decomposition, and the second is that of axiom A.5 (consistency under extension): behaviour under extension.

*Part 1: Associativity with respect to decomposition*

Invoking axiom A.3, we have that

$$\begin{aligned}\psi(w,x,y\,|\,z) &= \psi\big(w,(x,y)\,|\,z\big)\\ &= \psi(w\,|\,z)\circ\psi(x,y\,|\,w,z)\\ &= \psi(w\,|\,z)\circ\big(\psi(x\,|\,w,z)\circ\psi(y\,|\,w,x,z)\big)\end{aligned}$$

And similarly,

$$\begin{aligned}\psi(w,x,y\,|\,z) &= \psi\big((w,x),y\,|\,z\big)\\ &= \psi(w,x\,|\,z)\circ\psi(y\,|\,w,x,z)\\ &= \big(\psi(w\,|\,z)\circ\psi(x\,|\,w,z)\big)\circ\psi(y\,|\,w,x,z)\end{aligned}$$

We may thus conclude that the operation $\circ$ is associative under decomposition. ◪

*Part 2: Associativity with respect to extension*

When invoking extension with axiom A.5, we then have,

$$\begin{aligned}\psi(w,x,y\,|\,z) &= \psi\big((w,x),y\,|\,z\big)\\ &= \psi\big((w,x)\,|\,z\big)\circ\psi(y\,|\,w,x,z)\\ &= \big(\psi(w\,|\,z)\circ\psi(x\,|\,w,z)\big)\circ\psi(y\,|\,w,x,z)\\ &= \psi(w\,|\,z)\circ\big(\psi(x,y\,|\,w,z)\big)\\ &= \psi(w\,|\,z)\circ\big(\psi(x\,|\,w)\circ\psi(y\,|\,z)\big)\end{aligned}$$

For which the result is inclusive of $\psi\big(w,(x,y)\,|\,z\big)$. This affirms that the operation $\circ$ is associative under extension. Furthermore, associativity holds in both cases when extended to collections of $K$ propositions via the use of recursion in either case. □

Lemma A.5: Multiplicativity

The structure $(\Omega,\,\mathbb{F},\,\psi,\,\circ)$ is isomorphic to the structure $(\Omega,\,\mathbb{F},\,\psi,\,\times)$, where $\times$ is multiplication.

*Proof*

Aczel (2006, pp. 254–268) details that for any given function that is monotonic, continuous, cancellative, associative, all within a closed interval, then there exists a continuous function, $f$, such that

$$\alpha\circ\beta = f\big(f^{-1}(\alpha)+f^{-1}(\beta)\big)$$

We have these properties by lemmas A.1 to A.4, of which a closed interval is asserted by postulates A.4 and A.5 (the existence of a truism and a falsity). By axiom A.3, we have that,

$$f^{-1}\big(\psi(x\,|\,z)\circ\psi(y\,|\,x,z)\big) = f^{-1}\big(\psi(x\,|\,z)\big)+f^{-1}\big(\psi(y\,|\,x,z)\big)$$

Following which, we may assert

$$\begin{aligned}
\log\big(\psi(x,y\,|\,z)\big) &= \log\big(\psi(x\,|\,z)\circ\psi(y\,|\,x,z)\big)\\
&= \log\big(\psi(x\,|\,z)\big)+\log\big(\psi(y\,|\,x,z)\big)\\
\exp\Big(\log\big(\psi(x,y\,|\,z)\big)\Big) &= \exp\Big(\log\big(\psi(x\,|\,z)\big)+\log\big(\psi(y\,|\,x,z)\big)\Big)\\
&= \exp\Big(\log\big(\psi(x\,|\,z)\big)\Big)\times\exp\Big(\log\big(\psi(y\,|\,x,z)\big)\Big)\\
\implies \psi(x,y\,|\,z) &= \psi(x\,|\,z)\times\psi(y\,|\,x,z)
\end{aligned}$$

The logarithm being monotonic implies that the composition of the logarithm with $\psi$ is also monotonic, and the exponential is monotonic as well as bounded above and below within a closed interval. □

Lemma A.6: Normalisation
The structure $(\Omega,\,\mathbb{F},\,\psi,\,\times,\,R)$ is isomorphic to the structure $(\Omega,\,\mathbb{F},\,\psi,\,\times,\,[0,1])$

*Proof*
Postulates A.4 and A.5 (the existence of a truism and a falsity) assert lower and upper bounds on $R$, namely $R=[0,t]$. Suppose $[0,1]\subseteq[0,t]$, then there exists a $\psi(x)$ such that $1\leq\psi(x)\leq t$, then for the statement $\psi(x\,|\,x)=t$, we have that $1\leq t$.

Conversely, suppose $[0,t]\subseteq[0,1]$, then there exists a $\psi(x)$ such that $t\leq\psi(x)\leq 1$. Once again, consider the statement that $\psi(x\,|\,x)=t$, which implies that $t\leq 1$, therefore, it follows (by contradiction) that $t=1$. Following this, we may invoke the Cantor-Bernstein-Schröder theorem* to assert equality between $R$ and $[0,1]$. □

This proof illustrates that there exists some freedom of choice in deciding the magnitude of the truism, which implies that there exists some arbitrary scale in the definition. We establish $t=1$ as a convention based on the representation of the truism in propositional logic, in addition to it being the simplest scale (least assumptions).

Lemma A.7: Scaling
The structure $(\Omega,\,\mathbb{F},\,\psi,\,\times,\,[0,1],\,N)$ is isomorphic to the structure $(\Omega,\,\mathbb{F},\,\psi,\,\times,\,[0,1],\,\widetilde{N})$ with $\widetilde{N}(\frac{1}{2})=\frac{1}{2}$

*Proof*
Following axiom A.4 (negation) and lemma A.6 (normalisation), we may assert that $N\big(\psi(0)\big)=1$ and $N\big(\psi(1)\big)=0$, which is consistent with our framework of propositional logic and implies that $N$ is strictly decreasing. Axiom A.2 (continuity), additionally, implies continuity of $N$. Furthermore, $N$ has to be self-negating, meaning that $N\big[N\big(\psi(x)\big)\big]=\psi(x)$.

Suppose $N$ is not strictly decreasing, then we may assert that there exists some $\psi(x)<\psi(y)$, which results in $N\big(\psi(x)\big)=N\big(\psi(y)\big)$, then by self-negation we have that $\psi(x)=\psi(y)$, which is a contradiction.

* Cantor-Bernstein-Schröder theorem states that if a bijection exists between two sets, then the cardinality of those sets are equal. Here the identity map is a sufficient bijection since we have equal bounds.

From axiom A.4 we also have that $N$ maps the entirety of $R$ to itself, following which we may invoke Brouwer's fixed point theorem* to assert that there exists some $\psi(\alpha) \in [0,1]$, such that $N\big(\psi(\alpha)\big) = \psi(\alpha)$. Monotonicity of $N$ implies that this fixed point is unique.

Therefore, there exists a $k > 0$ such that

$$\psi^k(\alpha) = \tfrac{1}{2}$$

Where $k$ is chosen such that $N(\frac{1}{2}) = \frac{1}{2}$ remains consistent. Jaynes (2003, p. 33) asserts that the actual value of $k$ is of no importance to provide consistent results within the framework, because all previous lemmas are satisfied irrespective of $k$. □

Lemma A.8: Additivity of negation
The structure $\big(\Omega,\, \mathbb{F},\, \psi,\, \times,\, [0,1], N\big[\psi(x)\big]\big)$ is isomorphic to the structure $\big(\Omega,\, \mathbb{F},\, \psi,\, \times,\, [0,1], 1-\psi(x)\big)$ for any $x \in \mathbb{F}$.

*Proof*

*Part 1: Functional equation of $N$*

According to Paris (1995, p. 28) and De Kock (2014, p. 133), and following from lemma A.7 (scaling), let $\alpha = \psi(x, y \,|\, z)$ and $\beta = \psi(y \,|\, z)$, we may then examine the negations of $\alpha$ and $\beta$ as

$$\begin{aligned} N(\alpha) &= \psi(\overline{x}, y \,|\, z) \\ N(\beta) &= \psi(\overline{y} \,|\, z) = \psi\big((\overline{x}+y, x+\overline{y}) \,|\, z\big) \end{aligned}$$

We note that the composed $\alpha$, may be decomposed (according to axiom A.3), which yields

$$\begin{aligned} \psi(\overline{x}, y \,|\, z) &= \psi(y \,|\, z)\psi(\overline{x} \,|\, y, z) \\ &= \psi(y \,|\, z) N\big(\psi(x \,|\, y, z)\big) \\ &= \beta N\big(\tfrac{\alpha}{\beta}\big) \end{aligned}$$

However, there is an additional factor of consideration, namely the negated proposition. We therefore consider the alternate case of negation, in which we may assert the following,

$$\begin{aligned} \psi(x, \overline{y} \,|\, z) &= \psi(x, \overline{y}, \overline{x}+\overline{y} \,|\, z) \\ &= N(\alpha) N\big(\psi(\overline{x}+y \,|\, \overline{x}+\overline{y}, z)\big) \\ &= N(\alpha) N\Big(\tfrac{N(\beta)}{N(\alpha)}\Big) \end{aligned}$$

From which it follows that

$$\beta N\big(\tfrac{\alpha}{\beta}\big) = N(\alpha) N\Big(\tfrac{N(\beta)}{N(\alpha)}\Big) \qquad , \qquad 0 < \alpha \leq \beta \leq 1 \tag{A.3}$$

We now have a functional equation which $N$ must satisfy. It is of importance to note that the arguments to this functional equation are states of uncertainty, i.e. $\psi$. Bearing in mind that this equation has two arguments, we introduce the binary operator $*$, such that the operator is an extension of equation (A.3).

* Brouwer's fixed point theorem states that every continuous function from a closed disc which maps to itself has at least one fixed point.

Henceforth let any state of uncertainty have a representation of the form $\mu \equiv \psi(x)$, $\nu \equiv \psi(y)$, and $\upsilon \equiv \psi(z)$. With this representation we have a *parsable* means of employing the defined operator $*$. We now consider the assignment: $N(\beta) = \nu$ and $\alpha = \mu$, then by taking the negation of equation (A.3), we have an operation which is defined as

$$\mu * \nu = N\Big(N(\mu)N\big(\tfrac{\nu}{N(\mu)}\big)\Big) \tag{A.4}$$

Consequently, we have two functional results originating from the negation of a state of uncertainty, namely equations (A.3) and (A.4), of which the latter is in the form of a newly introduced binary operator. ◪

*Part 2: Commutativity of* $*$

By applying the functional given by equation (A.3) to the inner argument of equation (A.4), we observe that

$$\begin{aligned}\mu * \nu &= N\Big(N(\mu)N\big(\tfrac{\nu}{N(\mu)}\big)\Big)\\ &= N\Big(N(\nu)N\big(\tfrac{\mu}{N(\nu)}\big)\Big)\\ &= \nu * \mu\end{aligned}$$

Thus, it follows that the operator is commutative. ◪

*Part 3: Inversion*

Considering the operator to be applied recursively, that is, the operator acting upon two states of uncertainty as one argument of equation (A.3), we have that

$$\begin{aligned}(\mu * \nu)N\big(\tfrac{\mu}{\mu*\nu}\big) &= N(\mu)N\Big(\tfrac{N(\mu*\nu)}{N(\mu)}\Big)\\ &= N(\mu)N\Big(\tfrac{1}{N(\mu)}\Big[N(\mu)N\big(\tfrac{\nu}{N(\mu)}\big)\Big]\Big)\\ &= \nu\end{aligned}$$

Following which, it appears that the $*$ operator is an (argument-wise) inverse of the functional equation. Alternatively, letting equation (A.3) act as an argument within the operation of equation (A.4), we are essentially inverting* the order of operations, leading to,

$$\begin{aligned}\Big[\mu N\big(\tfrac{\nu}{\mu}\big)\Big] * \nu &= N\Big[N(\nu)N\Big(\frac{\mu N\left(\frac{\mu}{\nu}\right)}{N(\nu)}\Big)\Big]\\ &= N\Big(N(\nu)N\Big(\tfrac{1}{N(\nu)}\Big[N(\nu)N\big(\tfrac{N(\mu)}{N(\nu)}\big)\Big]\Big)\Big)\\ &= \mu\end{aligned}$$

Consequently, we have that the $*$ operator is the (argument-wise) inverse of the the functional given by equation (A.3). ◪

*Part 4: Associativity*

* Equivalent to first performing $f^{-1}\big(f(x)\big) = x$. and now performing $f\big(f^{-1}(x)\big) = x$

In order to test for associativity, we first consider

$$\begin{aligned}\mu * (\nu * \upsilon) &= N\left[N(\mu)N\left(\frac{N\left[N(\nu)N\left(\frac{\upsilon}{N(\nu)}\right)\right]}{N(\mu)}\right)\right] \\ &= N\left[N(\nu)N\left(\tfrac{\upsilon}{N(\nu)}\right)N\left(\frac{\mu}{N(\nu)N\left(\frac{\upsilon}{N(\nu)}\right)}\right)\right]\end{aligned}$$

Secondly, when comparing this with the alternative, we have that

$$\begin{aligned}(\mu * \nu) * \upsilon &= N\left[N(\nu)N\left(\tfrac{\mu}{N(\nu)}\right)N\left(\frac{\upsilon}{N(\nu)N\left(\frac{\mu}{N(\nu)}\right)}\right)\right] \\ &= N\left[N(\nu)N\left(\tfrac{\upsilon}{N(\nu)}\right)N\left(\frac{\mu}{N(\nu)N\left(\frac{\upsilon}{N(\nu)}\right)}\right)\right]\end{aligned}$$

From this, we are able to conclude that the operation $*$ is associative. ◪

*Part 5: Distributivity*

We may expand the inversion equation using $a\mu$, defined as,

$$\begin{aligned}a\mu &= a(\mu * \nu)\, N\left(\tfrac{\nu}{\mu * \nu}\right) \\ &= a(\mu * \nu)\, N\left(\tfrac{a\nu}{a(\mu * \nu)}\right)\end{aligned}$$

and applying the operation of $*$ between $a\mu$ and $a\nu$, we have

$$\begin{aligned}a\mu * a\nu &= \left(a(\mu * \nu)\, N\left(\tfrac{a\nu}{a(\mu * \nu)}\right)\right) * a\nu \\ &= a(\mu * \nu)\end{aligned}$$

We have that the operator $*$ is distributive. ◪

*Part 6: Induction on $*$*

For the fixed point in the states of uncertainty, we know that $N(\frac{1}{2}) = \frac{1}{2}$. Furthermore, we may evaluate $\frac{1}{2} * \frac{1}{2} = 1$. Using induction and the distributive property, we have (for $n > 0$),

$$\frac{1}{2^n} * \frac{1}{2^n} = \frac{1}{2^{n-1}}$$

Let $*_m(\frac{1}{2^n})$ represent the operation applied to $\frac{1}{2^n}$ $m$ times recursively, i.e. $\frac{1}{2^n} * \frac{1}{2^n} * \cdots * \frac{1}{2^n}$. Then if $m = 2^n$, this notation is (by induction) well defined since it would equal 1 in that case. Suppose $*_m(\frac{1}{2^n}) < \frac{m}{2^n}$, then choosing a number between them, say $\frac{1}{\sqrt[q]{2^p}}$ with $p, q \in \mathbb{N}$, meaning

$$*_m\left(\frac{1}{2^n}\right) < \frac{1}{\sqrt[q]{2^p}} < \frac{m}{2^n}$$

Raising the power by a magnitude of $q$ then yields

$$\left[*_m\left(\frac{1}{2^n}\right)\right]^q < \frac{1}{2^p} < \frac{m^q}{2^{nq}}$$

The last part of the inequality results in $m^q > 2^{nq-p}$, and we consequently have,

$$\left[*_m\left(\frac{1}{2^n}\right)\right]^q > (*_{2^{nq-p}})\left(\frac{1}{2^{nq-p}}\right) = \frac{2^{nq-p}}{2^{nq}} = \frac{1}{2^p}$$

Which contradicts our initial statement. In a similar fashion, the same contradiction occurs when assuming $*_m(\frac{1}{2^n}) > \frac{m}{2^n}$, therefore, it follows that

$$*_m\left(\frac{1}{2^n}\right) = \frac{m}{2^n}$$

Which completes the inductive process on $*$. ◪

*Part 7: Additivity of* $*$

Choosing $\mu = \frac{m_1}{2^n}$ and $\nu = \frac{m_2}{2^n}$, we have that

$$\begin{aligned}
\mu * \nu &= \frac{m_1}{2^n} * \frac{m_2}{2^n} \\
&= *_{m_1}\left(\frac{1}{2^n}\right) *_{m_2} \left(\frac{1}{2^n}\right) \\
&= *_{m_1+m_2}\left(\frac{1}{2^n}\right) \\
&= \frac{m_1 + m_2}{2^n} \\
&= \mu + \nu
\end{aligned}$$

Thus the binary operation $*$ is, in fact, the binary operation $+$. ◪

*Part 8: Negation principle*

Following the established parts, and using the functional identity (equation (A.3)), we have that

$$\begin{aligned}
(\mu + \nu)N\left(\tfrac{\mu}{\mu+\nu}\right) &= \nu \\
&= \mu + \nu - \nu
\end{aligned}$$

For which we are finally able to conclude that for any state of uncertainty in any proposition, we have that $N\big(\psi(x)\big) = 1 - \psi(x)$. □

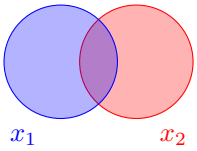
Lemma A.9: Finite additivity

For any collection of $K$ mutually exclusive propositions, $x_1, x_2, \ldots, x_K$, the state of uncertainty between their disjunctions is given by

$$\psi\left(\bigoplus_{k=1}^{K} x_k\right) = \sum_{k=1}^{K} \psi(x_k)$$

*Proof*

The statement is trivially true for $K = 1$, with $\psi(x_1) = \sum_{k=1}^{1} \psi(x_k) = \psi(x_1)$. Therefore we consider our basis step to be for the case $K = 2$. We may consider Venn diagriams as an illustrative means* to reason the state of uncertainty between a disjunction of propositions. When we consider the state of uncertainty of the disjunction, the result is the state of uncertainty in each proposition as well as two times the state of uncertainty in the conjunction. However, since we our collection of propositions




* A formal proof is given in section 2.2.4 (page 16)

is mutually exclusive, the conjunction between any propositions is false, therefore the state of uncertainty in the conjunction yields zero and the result simplifies,

$$\psi(x_1 + x_2) = \psi(x_1) + \psi(x_2) = \sum_{k=1}^{2} \psi(x_k)$$

Which is true. Now assume that this lemma holds for all $k$ up to $K$. We are then required to prove that this remains true for $K + 1$.

$$\begin{aligned}\psi\Big(\bigoplus_{k=1}^{K+1} x_k\Big) &= \psi\Big(\bigoplus_{k=1}^{K} x_k + x_{K+1}\Big) \\ &= \sum_{k=1}^{K} \psi(x_k) + \psi(x_{K+1}) \\ &= \sum_{k=1}^{K+1} \psi(x_k)\end{aligned}$$

The result thus follows that finite additivity is true for all $K$. □

Lemma A.10: Countable additivity

Extending finite additivity to any *non*-finite collection of mutually exclusive propositions $x_1, x_2, x_3, \ldots$ asserts that the framework satisfies

$$\psi\Big(\bigoplus_{k=1}^{\infty} x_k\Big) = \sum_{k=1}^{\infty} \psi(x_k)$$

*Proof*

By lemma A.8 (additivity of negation), we have that, for any state of uncertainty

$$N\big(\psi(x)\big) = \psi(\overline{x}) = 1 - \psi(x)$$

From lemma A.9 (finite additivity) we assert that the state of uncertainty for a limiting case of a collection of propositions is given by

$$\begin{aligned}\psi\Big(\bigoplus_{k=1}^{\infty} x_k\Big) &= \psi\Big(\lim_{K\to\infty} \bigoplus_{k=1}^{K} x_k\Big) \\ \sum_{k=1}^{\infty} \psi(x_k) &= \lim_{K\to\infty} \sum_{k=1}^{K} \psi(x_k)\end{aligned}$$

Since the magnitude of the truism is finite, it follows from lemma A.8 that the following remains consistent,

$$\psi\Big(\bigoplus_{k=1}^{\infty} \overline{x}_k\Big) = 1 - \psi\Big(\lim_{K\to\infty} \bigoplus_{k=1}^{K} x_k\Big) \tag{A.5}$$

However, from the assertion of axiom A.2 (sequential continuity) and given lemma A.9 (finite additivity), we have, from consistency of equation (A.5), i.e. the result is defined, that

$$\psi\Big(\lim_{K\to\infty} \bigoplus_{k=1}^{K} x_k\Big) = \lim_{K\to\infty} \psi\Big(\bigoplus_{k=1}^{K} x_k\Big)$$

$$= \lim_{K \to \infty} \sum_{k=1}^{K} \psi(x_k)$$
$$= \sum_{k=1}^{\infty} \psi(x_k)$$

From which it follows that countable additivity is satisfied. □

## A.2 From Uncertainty Logic to Probability Theory

### A.2.1 Equivalent Frameworks

The desiderata presented in definition 2.1 are related to axioms formulated by Cox (1946, 1961, pp. 4–34), who sought to provide a quantitative framework for inductive reasoning which extends classical logic, appeals to *common sense*, and is simultaneously based on *simple* axioms which make no large leaps in logic.

Axiom A.6: Cox's axioms of inductive reasoning

- C1 Knowledge of a proposition $x$ implies equivalent knowledge of its converse: $\overline{x}$.
- C2 Knowledge of the conjunction $(x, y)$ can be quantified from knowledge of a single proposition, e.g. $x$, combined with knowledge of the conditional $(x \mid y)$, provided that inference on the conditioned proposition, $y$, is true.
- C3 There exists numerical precision on the inferences of propositions, which implies that if an inference of a proposition $x$ is more believable than that of $y$, and inference of $y$ is more believable than that of $z$, then $x$ is necessarily more believable than $z$.

Cox (1961) reasoned that the first two axioms are the only* necessities needed to provide a platform for the development of a framework for inductive reasoning.

Pólya (1941, p. 2) reasoned that inductive reasoning should be subjective, unlike any forms of deductive reasoning, asserting that the nature of such a theory should, in principle, *measure one's degree of belief* — contending that reasonable (or reasonably accepted) beliefs may influence our state of knowledge. Pólya (1941) further reasoned that one's degree of belief may be subjected to variation; if evidence yields results on the contrary to one's expectations, then one's degree of belief in a proposition weakens, whereas if evidence yields results in favour of one's expectations, then one's degree of belief strengthens.

Cox (1961, pp. 1, 94) asserted that inductive reasoning is a form of *probable inference* (given an appropriate premise) and that it is relative; one person may accept an inferential conclusion as more or less credible than another person may — even though it is the same conclusion, meaning that it is subjective (and grounded in prior knowledge/evidence on what is being inferred). In addition, Cox (1961) stated that in any subjective sense, one's conclusions may be subject to change by new experience, thus probable inference in a hypothesis may gradually become more

* Cox (1961, p. 34) does not state the third axiom formally, reasoning that numerical precision may often be included without second thought and that it cannot be neglected.

probable provided that favourable instances of inference increase or otherwise decrease with contrary instances.

In both cases, that of Cox (1961) and Pólya (1941), their qualitative foundations were fulfilled by the *calculus of probabilities*, and both posited that probability theory succeeds the frequency formalism. Cox (1946, 1961) went one step further in quantifying the formalism of a framework for inductive reasoning which extends propositional logic *without* making any reference to frequencies (and games of chance), and in turn, resulted in the following,

**Theorem A.1** **Cox's Theorem**

Any subjective theory of reasoning which extends classical logic is isomorphic to the calculus of probabilities.

Interestingly, just as Cox's axioms lead to the calculus of probabilities, they may also be deduced in reverse: from within the framework of probabilities. As a result, Cox's axioms behave as a fail-safe which maintain consistency with any alternative formalism, even if these axioms are not incorporated by design (Skilling 2005, p. 4). An example of an alternative formalism is Kolmogorov's formalism of probability theory — which, from the calculus of probabilities, lead to Cox's axioms, thus asserting logical equivalence (Jaynes 2003, p. 652; Skilling 2005, p. 5).

When considering our formalism of *uncertainty logic*, we note that it is, on the premise of our desiderata, consistent with Cox's axioms. It is now of interest to show that this framework is equivalent to that of probability theory. Before doing so, it is worthwhile to first reconcile conceptual differences within the two frameworks (without making any assumption of equivalence).

The subjectivity in uncertainty logic is contained within the definition of $\psi(x)$: our conviction in the state of uncertainty in the proposition $x$, which was described as the level of how *convincing* a state of uncertainty may be. Pólya (1941) and Jaynes (2003) refer to *probability as a degree of belief* and this measure is exactly how convinced one is by any given probability, in other words, one's conviction.

What Cox (1961) and Jaynes (2003) respectively refer to as (the observer's) information and prior knowledge, are both considered to be one's knowledge on a known proposition (recall: a true statement) in uncertainty logic, which we denote by $\kappa$, and as $\psi(x \mid y, \kappa)$ in a conditioned state of uncertainty (where knowledge on $y$ is true). Knowledge naturally influences conviction and an initial conviction in a state of uncertainty is premised on initial knowledge thereof.

The advantage of referring to the terms as defined by uncertainty logic is that it provides a less abstract form of reasoning about these quantities, while maintaining consistency and meaning.

It is for more than merely aesthetic reasons that we shall continue the use of these quantities as defined by uncertainty logic; "knowledge" cannot be abstracted from an observer in the same way that "information" can. Moreover, preferring "conviction" more so than "degree of belief", since conviction is inherently more subjective than belief; one's belief can subsist beyond death, whereas one's conviction cannot.

At this point our formalism of uncertainty logic appears to be consistent with the philosophies of Pólya (1941), Cox (1961), and Jaynes (2003). We now give consideration to an alternative formu-

lation of the theory of probability — that of Kolmogorov (1956), who independently provided an axiomatic foundation for the theory. Kolmogorov (1956, pp. 1–16) decided upon his axioms as such based on simplicity (much like Cox (1946)) and phrased it in the framework of set theory,

Axiom A.7: Kolmogorov's axioms of probability theory

- K1 $\mathbb{F}$ is a field which contains the set $\Omega$, a collection of elementary elements, $e_i$, with subsets $\omega_k$.
  - I The complement for any set $\omega_k$ with respect to $\Omega$ is contained in $\mathbb{F}$: $\Omega - \omega_k \in \mathbb{F}$.
  - II $\mathbb{F}$ is closed under countable unions, meaning that if countably many $\omega_k$ are in $\mathbb{F}$, then their union is also in $\mathbb{F}$.
- K2 There exists a measure $p$ on $\mathbb{F}$ where
  - I A non-negative real number is assigned to each set $\omega_k$ in $\mathbb{F}$ as $p(\omega_k)$, which we call the probability of $\omega_k$.
  - II $p$ is normalised such that $p(\Omega) = 1$.
  - III $p$ is additive such that if $\omega_1, \omega_2, \ldots, \omega_K$ are disjoint (share no common elements), then $p(\bigcup_k \omega_k) = \sum_k p(\omega_k)$ is their union.
- K3 If a decreasing sequence $\omega_1 \supseteq \omega_2 \supseteq \cdots \supseteq \omega_K \supseteq \cdots$ tends to the empty set, then $\lim p(\omega_k) \to 0$ as $k \to \infty$.

Axiom K1 essentially asserts that $\mathbb{F}$ is a field with $\sigma$-algebra on $\Omega$. Kolmogorov (1956, p. 3) asserted that axioms K1 and K2 are valid and consistent for all finite systems of probability, and when considering such finite systems, axiom K2.III refers to *finite* additivity.

Kolmogorov (1956, p. 14) extended his formalism to encompass infinite sets (noting that axioms K1 and K2 remain valid), upon which continuity at zero followed in the form of axiom K3. Interestingly, this continuity was reasoned to be valid for both finite and infinite systems, even though asserted from the onset of infinite sets. For infinite sets, additivity in axiom K2.III refers to *countable* additivity.

We are able to rephrase these axioms in terms of propositions, and in doing so, we have that $\Omega$ is a collection of propositions ($w, x, y$, etc.), where all $\omega_k$ correspond to propositions with the inclusion of the core operators (`NOT`, `AND`, `OR`) in our framework*. With a propositional rephrase of these axiom, we are able to draw parallels to uncertainty logic.

Our formalism of uncertainty logic satisfies axiom K1 by postulate A.1 (logical propositions exist within a field, $\mathbb{F}$). Axiom K2 is in accordance with postulate A.2 (there exists a measure on $\mathbb{F}$), where non-negativity in axiom K2.I follows from postulate A.4 (falsity in uncertainty logic) defining a lower bound, following which axiom K2.II follows from postulate A.5 (truism in uncertainty logic), which is justified by lemma A.6 (normalisation) with the assertion that the truism in uncertainty logic may be defined (up to an arbitrary scale) as $\psi(x \mid x) = 1$.

Axiom K2.III is asserted by lemmas A.9 (finite additivity) and A.10 (countable additivity), expressed as $\psi\left(\bigoplus_k x_k\right) = \sum_k \psi(x_k)$. Axiom K3 follows from postulate A.3 (ordering in states of uncertainty) and it is worth noting that we do not need to invoke a collection of infinite propo-

* With note that the notion of subsets is not inherent to propositional logic.

sitions to reach this conclusion. In agreement with Jaynes (2003, p. 653), we are fundamentally concerned with an infinite sequence of different *states of knowledge*, which may be about a single proposition, as was illustrated in postulate A.3. It follows that uncertainty logic is also consistent with Kolmogorov's philosophy.

We may now reflect on how Cox's axioms are indeed simpler in nature than those of Kolmogorov's axioms; Cox's axioms were formulated as a basis for reasoning which resulted in probability theory, while Kolmogorov's axiom were framed as the basis for probability theory and make no reference to knowledge/information. Even though their results are equivalent, they approached it from different directions, which allowed Cox to make less preliminary assumptions. This is why Cox's axioms are appropriately chosen as the basis of our desiderata, moreover now that we have asserted uncertainty logic through Kolmogorov's axioms via Cox's theorem — showing that their independent formalisms compliment each other.

What we have achieved by our formalism of uncertainty logic was to find a *middle ground* between the formalisms of Cox and Kolmogorov. The goal was not to be rigorous simply for the sake of rigour, but rather to show that it can be done in this way*. We may thus give credence to probability theory being a fully functional *framework of reasoning*.

### A.2.2 Logic of Probabilities

Having established consistency between uncertainty logic and probability theory, we may henceforth express our *conviction* in a state of uncertainty in a proposition as the probability of that proposition: $\psi(x) \equiv p(x)$, in other words probability meaures (or evaluates) our conviction. From this point onward uncertainty logic may be known as probability theory.

⁂

*What is a probability?*

Instinctively, we may consider probability to be synonymous with *chance*. Offhand, we may consider this *chance* to be the possibility of some event happening. The assignment of this possibility would then be the probability of the said event actually happening. In other words, probability may be considered to be an indicator of how probable an occurrence of a possible event may be. If we consider this in practise, then we may, for example, flip a coin once and (instinctively) assign probabilities of $\frac{1}{2}$ for both heads and tails.

**Example A.1** **States of Uncertainty: The Glass Urn**

A *transparent (!)* glass urn contains three balls: one blue, one green, and one red. We may represent the state of each ball with the following propositions,

- $r \equiv$ the red ball is inside the urn
- $g \equiv$ the green ball is inside the urn
- $b \equiv$ the blue ball is inside the urn

* Jaynes (2003, p. 674) remarked that authors often make the mistake of confusing mathematical rigour with "correctness of results", and in doing so pay more attention to the appearance of mathematical rigour rather than the validity thereof.

If raising the question "Is the red ball inside the urn?", we may evaluate this question using standard propositional logic: $\overline{r} \equiv 0$, meaning $r \equiv 1$, i.e. the red ball is indeed inside the urn. However, since our third desideratum imposes consistency with propositional logic, we expect our state of uncertainty in the proposition $r$ to reflect the same, i.e. $\psi(\overline{r}) = 0$. To make this statement more succinct, we consider our conviction in the state of uncertainty to be conditioned on the knowledge we have at hand, being that we know the urn contains all balls, therefore

$$\psi(r \,|\, \text{the urn contains all balls}) \equiv 1$$

Here we introduced the proposition "the urn contains all the balls" which we considered as "our knowledge (or information)" — this is also consistent with our third desideratum which states that all (necessarily relevent) propositions should be considered.

To reiterate, note that propositional statements have two states of certainty: true and false. As such, our conviction in states of uncertainty in any proposition should reflect these two states of certainty when knowledge affects our conviction in states of uncertainty to become certain.

If we are to select a ball from the urn while closing our eyes (thus make a random selection), we may evaluate our conviction in the state of uncertainty in selecting a particular ball by first proposing the propositions of selecting a ball,

- $R \equiv$ the red ball is selected
- $G \equiv$ the green ball is selected
- $B \equiv$ the blue ball is selected

Then we may evaluate these propositions given our knowledge at hand, that all balls are contained within the urn,

$$\begin{aligned}\psi(R \,|\, \text{the urn contains all balls}) &= \tfrac{1}{3}\\ \psi(G \,|\, \text{the urn contains all balls}) &= \tfrac{1}{3}\\ \psi(B \,|\, \text{the urn contains all balls}) &= \tfrac{1}{3}\end{aligned}$$

Suppose the red ball has been selected and the urn now contains two balls, how do we evaluate our conviction in the state of uncertainty when next selecting a green ball?

We can reason that the urn contains one green ball and two balls total, therefore $\psi(G \,|\, R)$, the state of uncertainty in selecting a green ball given that a red ball has been selected is $\frac{1}{2}$. Of course, our state of uncertainty in then selecting a blue ball is a certainty, since the urn contains one remaining ball which we know to be blue, meaning $\psi(B \,|\, G, R) = 1$.

Firstly, note that we still refer to *state of uncertainty in a proposition* in this example, knowing that it may be exchanged with *probability of a proposition*, with $\psi \mapsto p$. Secondly, we reiterate

that conditioned knowledge (be it a proposition or other) is in a state of truth; it is known to us. Lastly, our (conditioned) state of knowledge influences our conviction and this becomes apparent when changes in our state of knowledge bring forth changes in our conviction.

To address the arbitrariness in the scale of how the truism is defined, we give it consideration in terms of lemma A.8 (additivity of negation) and axiom A.3 (decomposibility). If certainty is represented as $\psi(x \,|\, \overline{x}) = t$ (where $t$ may be anything), then lemma A.8 and axiom A.3 have to be scaled accordingly,

$$t = \psi(x \,|\, \kappa) + \psi(\overline{x} \,|\, \kappa) \qquad\qquad t\,\psi\big((x, y) \,|\, \kappa\big) = \psi(x \,|\, y, \kappa)\,\psi(y \,|\, \kappa)$$

To put this into perspective, let $t = 100$ and then reconsider example A.1. Let $u \equiv$ "the urn contains all balls", then probability for each ball being inside the urn while $u$ is true, is given by

$$\psi(R \,|\, u) = \psi(G \,|\, u) = \psi(B \,|\, u) = 33.\dot{3}$$

Which sums to 100 (i.e. $t$). The probability for selecting a green ball after the red ball has been selected is $p(G \,|\, R) = 50$, for which it then follows that the probability of selecting a blue ball (given the previous two selections) is $p(B \,|\, G, R) = 100$. These two quantities can be restated via the axiom A.3 (decomposibility),

$$\psi\big((B, G) \,|\, R\big) = \psi(B \,|\, G, R)\,\psi(G \,|\, R)$$

And knowing that the right hand side is equal to 5000, it is not a far reach to say that the left hand side has to be scaled by 100, because the joint probability (before decomposed) has to be 50 to maintain consistency with the lemma A.8! So the left hand side of this decomposition is scaled as $100\psi(B, G \,|\, R)$. This rescale is the reinterpretation of probabilities as percentages.

Even more generally, probabilities may be rescaled by any monotonic function, but then consistency requires that the sum and product rules are adapted accordingly, otherwise the *content* of the theory changes (Jaynes 2003, p. 652). It is evident that the choice of $t = 1$ is indeed the quintessential choice of scale.

* * *

*Can probabilities be unrelated to events of chance*?
We have established probability to be a measure of our conviction, and our convictions need not be weighed according to a numerical assignment (as illustrated in example A.1); it may very well be weighed based on *indeterminate* assignments.

One example hereof is embedded in *human experience*, which we consider from RiRi's perspective: RiRi was acquainted with two people, Jay and JamJam. After becoming familiar enough to form bonds of friendship, RiRi felt both of them to be trustworthy. RiRi's state of knowledge was based on his perception of trustworthiness: $s \equiv$ the person is trustworthy. Let $a$ and $b$ be Jay's and JamJam's respective impressions on RiRi, meaning that $p(s \,|\, a)$ and $p(s \,|\, b)$ determined his conviction to trust or distrust either of them.

A breach in trust by Jay meant that perception of Jay's trustworthiness declined, thus $p(s \,|\, a)$ resulted in lesser conviction in Jay's trustworthiness. Improved relations with JamJam resulted

in an improved perception of JamJam's trustworthiness, which meant that $p(s \mid b)$ resulted in greater conviction in JamJam's trustworthiness. Of course, any repeated breach of trust would yield an even lesser conviction while a strengthening bond would yield a stronger conviction (corresponding to negative/positive changes in states of knowledge).

Even without numerical assignments we may reach conclusions on indeterminate probabilistc assignments, having asserted that $p(s \mid a)$ has less plausibility than $p(s \mid b)$.

APPENDIX B

# The Theory of Hypergeometric Functions

## B.1 Introduction

Hypergeometric functions, also known as hypergeometric series, historically arose during a study of quadrature conducted by John Wallis during the $17^{\text{th}}$ century, however it was not formalised until it was later discovered and studied by Euler. Eventually, the hypergeometric function matured into its modern form under the helm of Gauss (as well as Pfaff), where it was also shown to be a solution to certain sescond order differential equations (Dutka 1984). The study in differential equations was continued by Riemann and Kummer, where Riemann observed that a characterisation for second-order differential equations with three regular singularities produced a powerful, efficient technique for obtaining hypergeometric functions (Andrews et al. 1999, p. 61).

Hypergeometric functions are also ubiquitous in the physical sciences; Bessel functions, orthogonal polynomials (e.g. associated Legendre polynomials or Jacobi polynomials), etc. can all be represented in terms of hypergeometric functions.

Hypergeometric functions are of the most important special functions in mathematics. It has become essential in the study of special functions due to its robust ability to generalise special functions across many different families. In essence, the hypergeometric function developed into a *catch-all* function for special functions. In this respect, it was surpassed by Meijer G-functions, of which hypergeometric functions are also a special case. Although mentioned, but not detailed, we are forced to consider one important question: *what is a special function?*

Special functions are functions that are ubiquitous in several (if not all) fields that make use of mathematical techniques to solve problems. For example, the trigonometric functions $\sin x$, $\cos x$ and $\tan x$ arise in all contexts, from solutions to differential equations describing waves, to approximations of functions (or irrational numbers), and the like.

However, it is not only the ubiquity that make these type of functions special. Special functions generally have representations which are defined by power series, (definite, indefinite, contour) integrals, generating functions, infinite products, orthogonal polynomials, etc. The aforementioned trigonometric functions have power series representations, consider one of which is,

$$\sin x = \sum_{k \geqslant 0} \frac{1}{(2k+1)!}(-1)^k x^{2k+1}$$

Special functions usually have series representations (in addition to other representations, if defined as such), and are analytic in their domains, but more importantly may be classed and defined by hypergeometric functions, in other words, most special functions share an underlying structure which is defined by hypergeometric functions — which is what makes hypergeometric functions as special and powerful as they are.

In order to make sense of hypergeometric functions, allow us to first consider the well-known geometric series,

$$S = \sum_{k \geqslant 0} ar^k$$

Within a geometric progression, the ratio between successive terms is a constant term known as the common ratio, namely $r$. To clarify this statement, for any geometric series $\sum_k t_k$, the ratio $\frac{t_{k+1}}{t_k} = r$. The behaviour of the series depends on this term, for example if $|r| < 1$, then the series converges. The extension of the geometric series is known as the *hyper*geometric series, where the ratio between successive terms is no longer a constant, but a rational function of $k$.

## B.2 Generalised Hypergeometric Functions

**Identities B.1**

$$\binom{-n}{k} = (-1)^k \binom{n+k-1}{k} \tag{B.1}$$

$${}_2F_1\left(\begin{matrix}1/2,\ 1\\ 3/2\end{matrix}\,\middle|\, -z^2\right) = \frac{1}{2z}\log\left(\frac{1+z}{1-z}\right) \tag{B.2}$$

The generalised hypergeometric function is a power series in $z$ with $p$ parameters in the numerator and $q$ parameters in the denominator, and it is defined in terms of rising powers,

$${}_pF_q\left(\begin{matrix}a_1,\ a_2,\ \ldots,\ a_p\\ b_1,\ b_2,\ \ldots,\ b_q\end{matrix}\,\middle|\, z\right) = \sum_{k\geqslant 0} \frac{a_1^{\overline{k}}\, a_2^{\overline{k}} \cdots a_p^{\overline{k}}}{b_1^{\overline{k}}\, b_2^{\overline{k}} \cdots b_q^{\overline{k}}} \frac{1}{k!}\, z^k \tag{B.3}$$

The parameters in the numerator, $a_i$, are known as *upper* parameters while the parameters in the denominator, $b_i$, are known as *lower* parameters, and the quantity $z$ is known as the argument. None of the lower parameters may be negative or zero, otherwise we encounter division by zero (Graham et al. 1989, p. 205).

As previously mentioned, many functions appear as special cases of equation (B.3), for example, the Bessel and modified Bessel functions of the first kind are defined by ${}_0F_1\left(\begin{matrix}\\ a\end{matrix}\,\middle|\, z\right)$, for particular $a$ and $z$. Additionally, since $1^{\overline{k}} = k!$, we may characterise the previously mentioned geometric series as a hypergeometric series with one upper and zero lower parameters, namely

$${}_1F_0\left(\begin{matrix}1\\ \end{matrix}\,\middle|\, z\right) = \sum_{k\geqslant 0} z^k$$

Having returned to the geometric series, we are reminded once more about the aforementioned characteristic criterion of a geometric series: successive terms have a constant ratio. Expanding upon this, for any hepergeometric series given by equation (B.3), $\sum_k t_k$, we have that

$$\frac{t_{k+1}}{t_k} = \frac{(k+a_1)(k+a_2)\cdots(k+a_p)}{(k+b_1)(k+b_2)\cdots(k+b_q)} \frac{1}{k+1}\, z \tag{B.4}$$

Equation (B.4) presents the ratio as a rational function of $k$, in other words, hypergeometric series are precisely those series with a first term of unity (i.e. $t_0 = 1$) and a ratio of successive terms that is a rational function of $k$ (i.e. a ratio of polynomials in $k$). If, in certain cases, the denominator does not contain the term $(k+1)$, it is useful to include this term in both the numerator and denominator, to compensate for the factorial term in the definition (Graham et al. 1989, p. 208). Note that the parameters of equation (B.3) appear *explicitly* in this ratio.

To familiarise ourselves with the aforementioned content, we may examine it in practise with the following example.

**Example B.1 Ubiquity of Hypergeometric Functions**

Consider the indefinite integral of $\sec x$, one that most encounter within their first course in calculus. Traditionally, we would multiply by a clever fraction and employ $u$-substitution. While this technique is remarkable, it is, nevertheless, merely a trick. Here we consider the manner in which it is traditionally solved in such courses to be rather unsatisfactory. Alternatively, we opt for a solution which involves *no trickery*.

$$\begin{aligned}
\int \mathrm{d}x \sec x &= 2\int \mathrm{d}x\, e^{ix}(1+e^{2ix})^{-1} \\
&= 2\int \mathrm{d}x\, e^{ix} \sum_{k=0}^{\infty} \binom{-1}{k} (e^{2ix})^k (1)^{-1-k} \\
&= 2\int \mathrm{d}x\, e^{ix} \sum_{k=0}^{\infty} (-1)^k \binom{1+k-1}{k} e^{2ixk} \\
&= 2\sum_{k=0}^{\infty} (-1)^k \int \mathrm{d}x\, e^{(1+2k)ix} \\
&= 2\sum_{k=0}^{\infty} (-1)^k \frac{1}{(1+2k)i} e^{(1+2k)ix} \\
&= -2ie^{ix} \sum_{k=0}^{\infty} \frac{1}{1+2k} \left(-e^{2ix}\right)^k
\end{aligned}$$

Ignoring the constant of integration for the time being, we observe that we have an infinite series, and we may therefore determine the ratio of successive terms in order to rewrite this series as a hypergeometric function.

$$\begin{aligned}
\frac{t_{k+1}}{t_k} &= \frac{1}{1+2(k+1)} \left(-e^{2ix}\right)^{k+1} \Big/ \frac{1}{1+2k} \left(-e^{2ix}\right)^k \\
&= \frac{1+2k}{3+2k} (-e^{2ix}) \\
&= \frac{(k+\frac{1}{2})(k+1)}{k+\frac{3}{2}} \frac{1}{k+1} (-e^{2ix})
\end{aligned} \tag{B.5}$$

Note that, as stated, the parameters of the hypergeometric appear *explicitly* in this ratio. Therefore, we have that

$$\begin{aligned}
\int \mathrm{d}x \sec x &= -2ie^{ix}\, {}_2F_1\left(\begin{matrix}\frac{1}{2}, 1 \\ \frac{3}{2}\end{matrix} \,\middle|\, -e^{2ix}\right) + c \\
&= -\log\left|\frac{1+ie^{ix}}{1-ie^{ix}}\right| + c \\
&= \log\left|\frac{2-i(e^{ix}-e^{-ix})}{i(e^{ix}+e^{-ix})}\right| + c \\
&= \log\left|\sec x + \tan x\right| + c
\end{aligned} \tag{B.6}$$

This example takes the unconventional approach to illustrate the practicality of utilising hypergeometric functions. In doing so, it enables us to do mathematics *without* involving clever trickery. In other words, we may opt for an approach involving greater dependence on logic and intuition when facing problems such as these. This does, unfortunately, come with a costly caveat, being that one is required to have a much greater background in a broader body of work*. In addition to this, example B.1 illustrates why it is so common for mathematica to produce hypergeometric results for integrals.

### B.2.1 Hypergeometric Functions of Finite Series

A negative integer as upper parameter causes the infinite series to become finite, since $(-a)^{\overline{k}} = 0$ whenever $k > a \geq 0$ (Graham et al. 1989, p. 206). The simplest case may be considered with a single upper parameter,

$$ {}_1F_0\left(\begin{matrix}-a\\ \end{matrix}\middle|\, z\right) = \sum_{k\geqslant 0} \frac{(-a)^{\overline{k}}}{k!}\, z^k = \sum_{k\geqslant 0} \frac{(-1)^k a^{\underline{k}}}{k!}\, z^k = \sum_{k\geqslant 0} (-1)^k \binom{a}{k} z^k = (1-z)^a \tag{B.7} $$

which is merely the binomial theorem and known to be finite. Suppose we consider the Chu-Vandermonde convolution, which is well-defined as

$$ \sum_{k=0}^{n} \binom{s}{k}\binom{t}{n-k} = \binom{s+t}{n} \tag{B.8} $$

This is "well-defined" in the sense that it has a determinate closed form. If, however, we consider the *alternating* Chu-Vandermonde convolution, there exists no well-known, succinct close form as with equation (B.8)

$$ S(n) = \sum_{k=0}^{n} (-1)^k \binom{s}{k}\binom{t}{n-k} \tag{B.9} $$

With our current state of knowledge, we may evaluate the ratio of successive terms as

$$ \frac{t_{k+1}}{t_k} = \frac{(s-k)(n-k)}{t-n+1+k}\frac{1}{k+1}(-1) $$

Which in turn yields a hypergeometric function! We express this hypergeometric function as,

$$ S(n) = \sum_{k=0}^{n} (-1)^k \binom{s}{k}\binom{t}{n-k} = \binom{t}{n} {}_2F_1\left(\begin{matrix}-s,\ -n\\ t-n+1\end{matrix}\middle|\, -1\right) \tag{B.10} $$

Which holds for $t \geq n$. With equation (B.10), we have thus determined a closed form solution for the alternating Chu-Vandermonde convolution. It is worth noting that equation (B.10) acquires the coefficient since the first term of the hypergeometric series is 1.

## B.3 Prominent Hypergeometric Functions

Literature (by and large) tend to focus on these functions. Their prominence is historically pervasive, due to their versatility and prevalence in mathematics and other fields.

* As seen in example B.1, where it was required to know Euler's identity, the binomial expansion for negative powers along with binomial identities, etc.

### B.3.1 Gauss' Hypergeometric Function

When involved with hypergeometric functions, Gauss' hypergeometric function is considered to be *the* hypergeometric function. This is arguably the most important hypergeometric function, while at the same time, being the most prominently studied.

The following differential equation was discovered by Euler, then investigated by both him and Pfaff, and it was Gauss that eventually provided remarkable insight into the characteristics thereof, while Riemann later expanded upon Gauss' findings (Andrews et al. 1999, p. 75).

$$z(1-z)\frac{\mathrm{d}^2w}{\mathrm{d}z^2} + \big(c-(a+b+1)\big)\frac{\mathrm{d}w}{\mathrm{d}z} - abw = 0$$

This equation is known as Euler's differential equation. Gauss provided the following solution, which is now known as *Gauss' hypergeometric function*,

$${}_2F_1\left(\begin{matrix}a,\,b\\c\end{matrix}\,\middle|\,z\right) = \sum_{k\geqslant 0}\frac{a^{\overline{k}}\,b^{\overline{k}}}{c^{\overline{k}}}\frac{1}{k!}\,z^k \tag{B.11}$$

Additionally, an independent solution is also given by $z^{1-c}{}_2F_1(a+1-c,\,b+1-c;\,2-c\,|\,z)$. Provided $\mathrm{Re}(c) > \mathrm{Re}(b) > 0$, the integral representation is given by,

$${}_2F_1\left(\begin{matrix}a,\,b\\c\end{matrix}\,\middle|\,z\right) = \frac{1}{B(b,c-b)}\int_0^1 \mathrm{d}x\, x^{b-1}(1-x)^{c-b-1}(1-zx)^a \tag{B.12}$$

The primary reason that Gauss' hypergeometric maintains a higher status is due to the versatility of the function depending on the parameters, being able to encompass many special functions (as encountered first in example B.1), as well as generalising many famous mathematical relations, for example, we may take note of the following,

$${}_2F_1\left(\begin{matrix}1/2,\,1/2\\3/2\end{matrix}\,\middle|\,-z^2\right) = \frac{1}{z}\log\big(z+\sqrt{1+z^2}\big) \qquad {}_2F_1\left(\begin{matrix}1/2,\,1\\3/2\end{matrix}\,\middle|\,-z^2\right) = \frac{1}{z}\tan^{-1}(z)$$

### B.3.2 Confluent Hypergeometric Functions

Within the study of the following second-order differential equation, known as Kummer's differential equation,

$$z\frac{\mathrm{d}^2w}{\mathrm{d}z^2} + (b-z)\frac{\mathrm{d}w}{\mathrm{d}z} - aw = 0$$

Kummer introduced the ${}_1F_1$ function in 1836, as a solution to his differential equation

$${}_1F_1\left(\begin{matrix}a\\b\end{matrix}\,\middle|\,z\right) = \sum_{k\geqslant 0}\frac{a^{\overline{k}}}{b^{\overline{k}}}\frac{1}{k!}\,z^k \tag{B.13}$$

This solution is known as the confluent hypergeometric function of the first kind. Equation (B.13) has the following Eulerian integral representation,

$${}_1F_1\left(\begin{matrix}a\\b\end{matrix}\,\middle|\,z\right) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)}\int_0^1 \mathrm{d}t\, t^{a-1}(1-t)^{b-a-1}\,e^{zt} \tag{B.14}$$

Since Kummer's differential equation is of second-order, there exists another linearly independent solution, which is $z^{1-b}\,{}_1F_1(a+1-b;\,2-b\,|\,z)$ function. Tricomi introduced a superposition of these two solutions in 1927 (Bateman et al. 1953, p. 257), defined as

$$U\big(a,\,b\,\big|\,z\big) = \frac{\Gamma(1-b)}{\Gamma(a+1-b)}\,{}_1F_1\left(\begin{matrix}a\\b\end{matrix}\,\middle|\,z\right) + \frac{\Gamma(b-1)}{\Gamma(a)}\,{}_1F_1\left(\begin{matrix}a+1-b\\2-b\end{matrix}\,\middle|\,z\right) \tag{B.15}$$

This is also a solution to Kummer's differential equation and it is known as the confluent hypergeometric function of the second kind. Equation (B.15) has the following integral representation,

$$U\big(a,\,b\,\big|\,z\big) = \frac{1}{\Gamma(a)}\int_0^\infty \mathrm{d}t\, t^{a-1}(1+t)^{b-a-1}\,e^{-zt} \tag{B.16}$$

Confluent hypegeometric functions are known as such because they are a confluence of two singularities, given that they arise from limiting Gauss' hypergeometric function: where the singularity at $b$ tends to that of infinity (Andrews et al. 1999, p. 188; Bateman et al. 1953, p. 262). Thus, a limiting cases of Gauss' hypergeometric are given by,

$${}_1F_1\left(\begin{matrix}a\\c\end{matrix}\,\middle|\,z\right) = \lim_{b\to\infty}\,{}_2F_1\left(\begin{matrix}a,\,b\\c\end{matrix}\,\middle|\,\frac{z}{b}\right) \tag{B.17}$$

$$z^a U\big(a,\,a-b\,\big|\,z\big) = \lim_{c\to\infty}\,{}_2F_1\left(\begin{matrix}a,\,b\\c\end{matrix}\,\middle|\,\frac{1-c}{z}\right) \tag{B.18}$$

Regarding limiting cases, the ${}_0F_1$ function is also known as the confluent hypergeometric limiting function, since it is a limiting case of the ${}_1F_1$ function,

$${}_0F_1\left(\begin{matrix}\\b\end{matrix}\,\middle|\,z\right) = \lim_{a\to\infty}\,{}_1F_1\left(\begin{matrix}a\\b\end{matrix}\,\middle|\,\frac{z}{a}\right)$$

## B.4 Bivariate Hypergeometric Functions

This section primarily aims to bring the following generalisations to the reader's attention*, since these are natural extensions to the previously encountered functions.

There are two primary classes of bivariate hypergeometric functions, the first class generalises Gauss' hypergeometric function to two variables, while the second class generalises the confluent hypergeometric function of the first kind.

Since bivariate hypergeometric series are double series, it is important to note that we now have composite terms between the two series, such as $a^{\overline{m+n}}$, as well as singular terms $b_1^{\overline{m}}$ and $b_2^{\overline{n}}$. As a result, we need to adapt our notation to accomodate for this, therefore we consider

$$g\left(\begin{matrix}a\,:\,b_1\\c_2\end{matrix}\,\middle|\,x,\,y\right) = \sum_{m,n\geqslant 0}\frac{a^{\overline{m+n}}b_1^{\overline{m}}}{c_2^{\overline{n}}}\frac{1}{m!\,n!}\,x^m y^n$$

$$h\left(\begin{matrix}a\,:\,b_1,\,b_2\\c_1\end{matrix}\,\middle|\,x,\,y\right) = \sum_{m,n\geqslant 0}\frac{b_1^{\overline{m}}b_2^{\overline{n}}}{a^{\overline{m+n}}c_1^{\overline{m}}}\frac{1}{m!\,n!}\,x^m y^n$$

* For further reading, consult Bateman et al. (1953, §5.7).

Here $g$ and $h$ are any arbitrary bivariate hypergeometric functions. Our first adjustment to currently known convention is the addition of a subscript to each non-composite parameter, since it provides an indication to its associated series. Secondly, we separate the composite terms with a colon, with this parameter also being position sensitive, thus indicating whether it is a lower or upper parameter. The parameters to the right of the colon follow our previously established convention.

### B.4.1 Appell Functions

In 1880 Appell developed four series which are analogous to Gauss' hypergeometric (Bateman et al. 1953, p. 222). Here we consider two of the four Appell functions,

$$F_1\left(\begin{matrix}a\\b\end{matrix} : c_1,\, c_2 \;\middle|\; x,\, y\right) = \sum_{m,n\geqslant 0} \frac{a^{\overline{m+n}} c_1^{\overline{m}} c_2^{\overline{n}}}{b^{\overline{m+n}}} \frac{1}{m!\, n!} x^m y^n \tag{B.19}$$

$$F_2\left(\begin{matrix}a\\ \end{matrix} : \begin{matrix}b_1,\, b_2\\ c_1,\, c_2\end{matrix} \;\middle|\; x,\, y\right) = \sum_{m,n\geqslant 0} \frac{a^{\overline{m+n}} b_1^{\overline{m}} b_2^{\overline{n}}}{c_1^{\overline{m}} c_2^{\overline{n}}} \frac{1}{m!\, n!} x^m y^n \tag{B.20}$$

### B.4.2 Humbert Functions

Humbert introduced seven bivariate functions in 1926, which extend the confluent hypergeometric functions (Bateman et al. 1953, p. 225), two of which are given by

$$\Psi_1\left(\begin{matrix}a\\ \end{matrix} : \begin{matrix}b_1\\ c_1,\, c_2\end{matrix} \;\middle|\; x,\, y\right) = \sum_{m,\,n\geqslant 0} \frac{a^{\overline{m+n}} b_1^{\overline{m}}}{c_1^{\overline{m}} c_2^{\overline{n}}} \frac{1}{m!\, n!} x^m y^n \tag{B.21}$$

$$\Psi_2\left(\begin{matrix}a\\ \end{matrix} : \begin{matrix}\\ b_1,\, b_2\end{matrix} \;\middle|\; x,\, y\right) = \sum_{m,\,n\geqslant 0} \frac{a^{\overline{m+n}}}{b_1^{\overline{m}} b_2^{\overline{n}}} \frac{1}{m!\, n!} x^m y^n \tag{B.22}$$

The Appell functions are part of fourteen *complete* series, while the Humbert functions are part of twenty *confluent* series. The complete list is known as Horn's list of bivariate hypergeometric functions (Bateman et al. 1953, p. 224).

## B.5 Differentiation of Hypergeometric Functions

**Identities B.2**

$$x^{\overline{m+n}} = x^{\overline{m}}(x+m)^{\overline{n}} = x^{\overline{n}}(x+n)^{\overline{m}} \tag{B.23}$$

Differentiation is probably best digested via an example. As such, when we differentiate equation (B.3), we may consider the following example.

**Example B.2 Differentiating the Generalised Hypergeometric**

Note that when differentiating a series, we need to assure ourselves that the first term is

still consistent with what is expected,

$$\frac{\mathrm{d}}{\mathrm{d}z}\,{}_pF_q\left(\begin{matrix}a_1,\,\ldots,\,a_p\\ b_1,\,\ldots,\,b_q\end{matrix}\,\middle|\,z\right)=\sum_{k\geqslant 1}\frac{a_1^{\overline{k}}\,a_2^{\overline{k}}\cdots a_p^{\overline{k}}}{b_1^{\overline{k}}\,b_2^{\overline{k}}\cdots b_q^{\overline{k}}}\frac{1}{(k-1)!}\,z^{k-1}$$

Here we are required to adjust the index to maintain consistency, therefore the series has to start at $k=1$, however we introduce a translation accross the index to produce a zero-index starting point.

$$\begin{aligned}
&=\sum_{k+1\geqslant 1}\frac{a_1^{\overline{k+1}}\,a_2^{\overline{k+1}}\cdots a_p^{\overline{k+1}}}{b_1^{\overline{k+1}}\,b_2^{\overline{k+1}}\cdots b_q^{\overline{k+1}}}\frac{1}{k!}\,z^k\\
&=\sum_{k\geqslant 0}\frac{a_1(a_1+1)^{\overline{k}}\,a_2(a_2+1)^{\overline{k}}\cdots a_p(a_p+1)^{\overline{k}}}{b_1(b_1+1)^{\overline{k}}\,b_2(b_2+1)^{\overline{k}}\cdots b_q(b_q+1)^{\overline{k}}}\frac{1}{k!}\,z^k\\
&=\frac{a_1\cdots a_p}{b_1\cdots b_q}\,{}_pF_q\left(\begin{matrix}a_1+1,\,\ldots,\,a_p+1\\ b_1+1,\,\ldots,\,b_q+1\end{matrix}\,\middle|\,z\right)
\end{aligned}$$

If we take the second derivative, then we may note that the same would happen for the translated parameters, i.e. the current term would appear as a coefficient and the parameters of the function would translate once more. As such, we may generalise this as,

$$\frac{\mathrm{d}^n}{\mathrm{d}z^n}\,{}_pF_q\left(\begin{matrix}a_1,\,\ldots,\,a_p\\ b_1,\,\ldots,\,b_q\end{matrix}\,\middle|\,z\right)=\frac{a_1^{\overline{n}}\cdots a_p^{\overline{n}}}{b_1^{\overline{n}}\cdots b_q^{\overline{n}}}\,{}_pF_q\left(\begin{matrix}a_1+n,\,\ldots,\,a_p+n\\ b_1+n,\,\ldots,\,b_q+n\end{matrix}\,\middle|\,z\right)\tag{B.24}$$

We observe that differentiation acts upon all parameters, however, it is possible to translate only a single parameter, meaning that the operation acts upon one particular parameter, in that case we may consider this operator, $(\vartheta+a_i)$ with $\vartheta=z\frac{\mathrm{d}}{\mathrm{d}z}$ (Graham et al. 1989, p. 219). This operator is known as the homogeneity operator, and allows us to manipulate the function as follows,

$$\begin{aligned}
(\vartheta+a_1)\,{}_pF_q\left(\begin{matrix}a_1,\,\ldots,\,a_p\\ b_1,\,\ldots,\,b_q\end{matrix}\,\middle|\,z\right)&=\sum_{k\geqslant 1}(k+a_1)\frac{a_1^{\overline{k}}\,a_2^{\overline{k}}\cdots a_p^{\overline{k}}}{b_1^{\overline{k}}\,b_2^{\overline{k}}\cdots b_q^{\overline{k}}}\frac{1}{k!}\,z^k\\
&=\sum_{k\geqslant 1}\frac{a_1(a_1+1)^{\overline{k}}\,a_2^{\overline{k}}\cdots a_p^{\overline{k}}}{b_1^{\overline{k}}\,b_2^{\overline{k}}\cdots b_q^{\overline{k}}}\frac{1}{k!}\,z^k\\
&=a_1\,{}_pF_q\left(\begin{matrix}a_1+1,\,\ldots,\,a_p\\ b_1,\,\ldots,\,b_q\end{matrix}\,\middle|\,z\right)
\end{aligned}$$

Higher derivatives follow in the same way as before, except acting only on the particular parameter. In order to generalise this to higher order derivatives affecting any upper parameter, we update the previous operator: $(\vartheta+s_i)_j$, here $j$ is the order of the differential operator, $i$ is the index of the upper parameter and $s_i=a_i+(j-1)$. The $n^{\text{th}}$ successive derivative may then be realised as

$$(\vartheta+s_i)_n\,{}_pF_q\left(\begin{matrix}a_1,\,\ldots,\,a_p\\ b_1,\,\ldots,\,b_q\end{matrix}\,\middle|\,z\right)=a_i^{\overline{n}}\,{}_pF_q\left(\begin{matrix}a_1,\,\ldots,\,a_i+n,\,\ldots,\,a_p\\ b_1,\,\ldots,\,b_q\end{matrix}\,\middle|\,z\right)\tag{B.25}$$

Note that this acts only upon the upper parameters. In order to apply this to the lower parameters, we first need to take care in noting that after the operator acts upon the function, the terms

translate in the numerator, while our lower parameters appear in the denominator. As such, we consider equation (B.23) given in identities B.2, by taking the ratio of the two equivalent quantities and setting $m = 1$, we have $\frac{x}{x^{\overline{n}}} = \frac{x+n}{(x-1)^{\overline{n}}}$.

We are therefore able to extend this operation to the lower parameters, with a minor adjustment to the differential operator: $(\vartheta + t_i)_j$, where $i$ and $j$ are the same as before and $t_i = b_i - j$. From which it follows,

$$(\vartheta + t_i)_n \, {}_pF_q\left(\begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix} \,\middle|\, z\right) = (b_i - 1)^{\underline{n}} \, {}_pF_q\left(\begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_i - n, \ldots, b_q \end{matrix} \,\middle|\, z\right) \tag{B.26}$$

For bivariate functions, partial differentiation behaves in a similar (albeit expected) manner.

$$\frac{\partial^n}{\partial x^n} \Psi_1\left(\begin{matrix} a : b_1 \\ c_1, c_2 \end{matrix} \,\middle|\, x, y\right) = \frac{a^{\overline{n}} b_1^{\overline{n}}}{c_1^{\overline{n}}} \Psi_1\left(\begin{matrix} a+n : b_1+n \\ c_1+n, c_2 \end{matrix} \,\middle|\, x, y\right)$$

$$\frac{\partial^n}{\partial y^n} \Psi_1\left(\begin{matrix} a : b_1 \\ c_1, c_2 \end{matrix} \,\middle|\, x, y\right) = \frac{a^{\overline{n}}}{c_2^{\overline{n}}} \Psi_1\left(\begin{matrix} a+n : b_1 \\ c_1, c_2+n \end{matrix} \,\middle|\, x, y\right)$$

## B.6 Hypergeometric Summands

The generalised hypergeometric series has no upper bound, however, in practise, we may not necessarily encounter such unbounded summations. Therefore, it is of interest to consider possibilities where the series may have a definite upper bound. We start by defining hypergeometric summands, or the partial hypergeometric function as

$${}_pF_q^{(K)}\left(\begin{matrix} a_1, a_2, \ldots, a_p \\ b_1, b_2, \ldots, b_q \end{matrix} \,\middle|\, z\right) = \sum_{k=0}^{K} \frac{a_1^{\overline{k}} \cdots a_p^{\overline{k}}}{b_1^{\overline{k}} \cdots b_q^{\overline{k}}} z^k \tag{B.27}$$

Let the $K^{\text{th}}$ term of such a series, $c_K$, be represented by

$$\left[{}_pF_q^{(K)}\left(\begin{matrix} a_1, a_2, \ldots, a_p \\ b_1, b_2, \ldots, b_q \end{matrix} \,\middle|\, z\right)\right]_K = \frac{a_1^{\overline{K}} \cdots a_p^{\overline{K}}}{b_1^{\overline{K}} \cdots b_q^{\overline{K}}} z^K$$

The series given by equation (B.27) is said to be summable in hypergeometric terms, with constants $w$ and $W$, if

$${}_pF_q^{(K)}\left(\begin{matrix} a_1, a_2, \ldots, a_p \\ b_1, b_2, \ldots, b_q \end{matrix} \,\middle|\, z\right) = w\left[{}_mF_n^{(K)}\left(\begin{matrix} A_1, A_2, \ldots, A_m \\ B_1, B_2, \ldots, B_n \end{matrix} \,\middle|\, Z\right)\right]_K + W \tag{B.28}$$

such that $A_1, \ldots, A_m; B_1, \ldots, B_n;$ and $Z$ exist (Graham et al. 1989, p. 224). Let the term of the ${}_mF_n{}^K$ partial hypergeometric function containing these parameters be $C_k$.

The partial hypergeometric series may be evaluated as such if the ratio of successive terms may be expressed in terms of polynomials in $k$,

$$\frac{c_{k+1}}{c_k} = \frac{p(k+1)}{p(k)} \frac{q(k)}{r(k+1)} \tag{B.29}$$

With $q$ and $r$ subject to

$$(k+\alpha) \mid q(k) \qquad \text{and} \qquad (k+\beta) \mid r(k) \tag{B.30}$$

Where $\alpha - \beta < 0 \in \mathbb{Z}$. Divisibility means that the quotient are still a polynomials (Graham et al. 1989, p. 225). The conditions specified in equation (B.30) may be achieved by setting $p(k) = 1$, for the time being, then $q(k) = (k + a_1) \cdots (k + a_p)z$ and $r(k) = (k + b_1 - 1) \cdots (k + b_q - 1)k$. If $q$ and $r$ have factors $(k + \alpha)$ and $(k + \beta)$ such that $\alpha - \beta = N > 0$, then they may be divided out and reconstruct $p(k)$ such that the condition holds:

$$p(k)(k + \alpha - 1)^{\underline{N}} = p(k)(k + \alpha - 1)(k + \alpha - 2) \cdots (k + \beta + 1)$$

By evaluating the definite sum given by equation equation (B.28), we are interested in finding a hypergeometric term $C_k$, such that

$$c_k = wC_{k+1} - wC_k \tag{B.31}$$

The hypergeometric term of interest may *also* be represented in terms of polynomials, in a similar fashion to equation (B.29) and in terms of $c_k$,

$$wC_{k+1} = \frac{r(k)s(k)c_k}{p(k)} \tag{B.32}$$

Here $s$ is conditioned from equations (B.29) and (B.31), leading to

$$p(k) = q(k)s(k + 1) - r(k)s(k) \tag{B.33}$$

For the degree $t$ polynomial $s(k) = \alpha_t k^t + \alpha_{t-1}k^{t-1} + \cdots + \alpha_0$, with $\alpha_t \neq 0$, let $Q(k) = q(k) - r(k) = \beta k^{v-1} + \cdots$ and $R(k) = q(k) + r(k) = \gamma k^v + \cdots$, with $\gamma \neq 0$, then the recurrence in equation (B.33) may be rewritten as

$$\begin{aligned} 2p(k) &= Q(k)\big(s(k + 1) + s(k)\big) + R(k)\big(s(k + 1) - s(k)\big) \\ &= (2\beta\alpha_t + \gamma t\alpha_t)k^{t+v-1} + \cdots \end{aligned} \tag{B.34}$$

This enables the following two constraints: either $2\beta + \gamma t \neq 0$ and $t = \deg(p) - \deg(R) + 1$, or $2\beta + \gamma t = 0$ and $t > \deg(p) - \deg(R) + 1$ (Graham et al. 1989, p. 226). This procedure is known as *Gosper's method*.

**Example B.3** **Sum-thing Old and Sum-thing New**

Students occasionally encounter popular series such as the sum of $n$ integers or the sum of $n$ square integers. Most commonly, a closed form for these (finite) series is determined via the technique of sum of summations (finding the series $2S_n$).

$$\sum_{k=0}^{K} k^2$$

We will, as is expected, employ our new knowledge in order to find a closed form for this series. Starting with the ratio of successive terms,

$$\frac{p(k + 1)}{p(k)} \frac{q(k)}{r(k + 1)} = \frac{(k + 1)^2}{k^2}$$

Here we assert that $p(k) = k^2$ and $r(k) = q(k) = 1$. The condition on $s$ given by equation (B.34) suggests that $s$ is of degree 3: $s(k) = \alpha_3 k^3 + \alpha_2 k^2 + \alpha_1 k^1$. In addition to this, we also have that $p$ is the finite difference of $s$:

$$p(k) = s(k+1) - s(k) = \Delta_+ s(k)$$

Meaning that the calculation reduces to the determination of coefficients,

$$\begin{aligned} \Delta_+ s(k) &= 3\alpha_3 k^2 + (3\alpha_3 + 2\alpha_2)k + (\alpha_3 + \alpha_2 + \alpha_1) \\ \implies \quad \alpha_3 = \tfrac{1}{3} \quad &, \quad \alpha_2 = -\tfrac{1}{2} \quad , \quad \alpha_1 = \tfrac{1}{6} \end{aligned}$$

For which we have the closed form for the $K^{\text{th}}$ term,

$$\begin{aligned} wC_k = s(k) &= \tfrac{1}{3}k^3 - \tfrac{1}{2}k^2 + \tfrac{1}{6}k \\ &= \tfrac{1}{6}k(2k-1)(k-1) \end{aligned}$$

In certain cases, the partial hypergeometric series may be solved and expressed in the same way as general a hypergeometric function (without loss of generality).

APPENDIX C

# Hyperspherical Variables and Sampling

## C.1 The Surface Area of the Hypersphere

**Identities C.1**

$$\Gamma(2n) = (\pi)^{-1/2}\, 2^{2n-1}\, \Gamma(n)\Gamma\left(n+\tfrac{1}{2}\right) \tag{C.1}$$

$$\Gamma\left(-n+\tfrac{1}{2}\right) = (-1)^n\, \pi^{1/2}\, 2^{2n}\, \frac{\Gamma(n+1)}{\Gamma(2n+1)} \tag{C.2}$$

$${}_2F_1\left(\begin{matrix} a,\ b \\ c \end{matrix}\,\middle|\, 1\right) = \frac{\Gamma(c-a-b)\Gamma(c)}{\Gamma(c-a)\Gamma(c-b)} \tag{C.3}$$

The Jacobian's determinant serves as preserving area between coordinate transformations, and it is equivalent to the *surface element* of a parametrised surface. This equivalence can be asserted through the concept of the surface integral for any parametric function, $f(\boldsymbol{x})$. Consider the case with $f(x, y, z)$, then the surface integral is given by

$$\iint_S \mathrm{d}S\, f(x,y,z) = \iint_D \mathrm{d}A\, f\big(\boldsymbol{r}(u,v)\big)\, \|\boldsymbol{r}_u \times \boldsymbol{r}_v\|$$

Where $\boldsymbol{r}(u,v) = \big(x(u,v),\ y(u,v),\ z(u,v)\big)$ and $\boldsymbol{r}_u = (\partial_u x, \partial_u y, \partial_u z)$ and $\boldsymbol{r}_v = (\partial_v x, \partial_v y, \partial_v z)$. The norm of the cross product of these vectors, $\|\boldsymbol{r}_u \times \boldsymbol{r}_v\|$, is known as the surface element (Stewart 2009, p. 1112–1118). The *surface area* is then given by the unit surface integral,

$$A_S = \iint_S \mathrm{d}S = \iint_D \mathrm{d}A\, \|\boldsymbol{r}_u \times \boldsymbol{r}_v\| \tag{C.4}$$

And we have a representation for the surface area of an arbitrarily parametrised quantity. Consider the above, but in terms of the wedge product and differential forms,

$$\begin{aligned} \mathrm{d}x \wedge \mathrm{d}y &= \left(\frac{\partial u}{\partial x}\mathrm{d}u + \frac{\partial u}{\partial x}\mathrm{d}v\right) \wedge \left(\frac{\partial u}{\partial y}\mathrm{d}u + \frac{\partial v}{\partial y}\mathrm{d}v\right) \\ &= \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial v}{\partial x}\frac{\partial u}{\partial y}\right)(\mathrm{d}u \wedge \mathrm{d}v) \\ &= \|\boldsymbol{r}_u \times \boldsymbol{r}_v\|\, (\mathrm{d}u \wedge \mathrm{d}v) \end{aligned} \tag{C.5}$$

The surface element arises naturally. Here we observe that the Jacobian and the surface element are one and the same! This extends without loss of generality to arbitrary coordinates,

$$\mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_K = |J_K|\, \mathrm{d}u_1 \wedge \cdots \wedge \mathrm{d}u_K \tag{C.6}$$

For this we conclude that the surface area of a parametric function, $f$, of $K$-dimensions is the $(K-1)$-dimensional integral over the region $D$ where the $K-1$ coordinates $(u_1, u_2, \ldots, u_{K-1})$ span the surface of $f$.

We may thus proceed to determine the surface area of the hypersphere,

$$A_K(r) = \int_0^\pi \mathrm{d}\theta_{K-1} \int_0^\pi \mathrm{d}\theta_{K-2} \cdots \int_0^\pi \mathrm{d}\theta_2 \int_0^\pi \mathrm{d}\theta_1\, |J_K|$$

$$= 2\pi r^{K-1} \int_0^\pi \mathrm{d}\theta_{K-1} \, \sin^{K-2}\theta_{K-1} \int_0^\pi \mathrm{d}\theta_{K-2} \, \sin^{K-3}\theta_{K-2} \cdots \int_0^\pi \mathrm{d}\theta_2 \, \sin\theta_2$$
$$= 2\pi r^{K-1} \prod_{k=2}^{K-1} \int_0^\pi \mathrm{d}\theta_k \, \sin^{k-1}\theta_k \tag{C.7}$$

For the integral in equation (C.7), let $m$ represent arbitrary $k$, then we only need to consider solving the 1-dimensional integral, since the product essentially consists of the same, identical 1-dimensional integrals for various $k$.

$$\int_0^\pi \mathrm{d}\theta \, \sin^m\theta = \frac{1}{(2i)^m} \sum_{\ell=0}^{m} (-1)^\ell \binom{m}{\ell} \int_0^\pi \mathrm{d}\theta \, e^{(m-2\ell)i\theta} \tag{C.8}$$
$$= \frac{1}{i(2i)^m} \sum_{\ell=0}^{m} (-1)^\ell \binom{m}{\ell} \frac{1}{(m-2\ell)} \left[(-1)^m - 1\right] \tag{C.9}$$
$$= \frac{(-1)^n}{(2)^{2n+1}} \sum_{\ell=0}^{2n+1} (-1)^\ell \binom{2n+1}{\ell} \frac{1}{(n+\frac{1}{2}-\ell)} \tag{C.10}$$

Observe that $\left((-1)^m - 1\right)$ is zero for all even values of $m$ in equation (C.9), and $-2$ for all odd values, therefore we condition $m$ to be of the form $m = 2n+1$. Let $t_\ell$ represent the summand in equation (C.10), for which we may consider the ratio of successive summands as,

$$\frac{t_{\ell+1}}{t_\ell} = (-1)^{\ell+1} \binom{2n+1}{\ell+1} \frac{1}{(n+\frac{1}{2}-\ell-1)} \cdot \left[(-1)^\ell \binom{2n+1}{\ell} \frac{1}{n+\frac{1}{2}-\ell}\right]^{-1}$$
$$= \frac{1}{\ell+1} \frac{(\ell-2n-1)(\ell-n-\frac{1}{2})}{\ell-n+\frac{1}{2}} \tag{C.11}$$

This allows us to represent the sum in equation (C.10) as a hypergeometric function:

$$\sum_{\ell=0}^{2n+1} t_\ell = t_0 \, {}_2F_1\left(\begin{matrix}-2n-1, & -n-\frac{1}{2} \\ & -n+\frac{1}{2}\end{matrix} \,\middle|\, 1\right)$$
$$= \frac{1}{n+\frac{1}{2}} \, {}_2F_1\left(\begin{matrix}-2n-1, & -n-\frac{1}{2} \\ & -n+\frac{1}{2}\end{matrix} \,\middle|\, 1\right)$$

For which we now have the solution of the integral to be given by,

$$\int_0^\pi \mathrm{d}\theta \, \sin^m\theta = \frac{(-1)^n}{(2)^{2n+1}} \frac{1}{n+\frac{1}{2}} \, {}_2F_1\left(\begin{matrix}-2n-1, & -n-\frac{1}{2} \\ & -n+\frac{1}{2}\end{matrix} \,\middle|\, 1\right). \tag{C.12}$$

The hypergeometric function may be simplified via identities C.1, which yields

$${}_2F_1\left(\begin{matrix}-2n-1, & -n-\frac{1}{2} \\ & -n+\frac{1}{2}\end{matrix} \,\middle|\, 1\right) = \frac{\Gamma\left(-n+\frac{1}{2}+2n+1+n+\frac{1}{2}\right)\Gamma\left(-n+\frac{1}{2}\right)}{\Gamma\left(-n+\frac{1}{2}+2n+1\right)\Gamma\left(-n+\frac{1}{2}+n+\frac{1}{2}\right)}$$
$$= (\pi)^{-1/2}(2)^{2(n+1)-1} \frac{\Gamma(n+1)\,\Gamma\left(n+1+\frac{1}{3}\right)\Gamma\left(-n+\frac{1}{2}\right)}{\Gamma\left(n+\frac{3}{2}\right)}$$
$$= (\pi)^{-1/2}(2)^{2n+1} \left[(-1)^n(\pi)^{1/2}(2)^{2n} \frac{\Gamma(n+1)}{\Gamma(2n+1)} \Gamma(n+1)\right]$$
$$= (-1)^n(2)^{2n+1}(\pi)^{1/2} \frac{\Gamma(n+1)}{\Gamma\left(n+\frac{1}{2}\right)} \tag{C.13}$$

Resulting in the solution of the integral in equation (C.7), with $m = 2n + 1$, to be

$$\begin{aligned}\int_0^\pi \mathrm{d}\theta\, \sin^m\theta &= \frac{(-1)^n}{(2)^{2n+1}}\frac{1}{n+\frac{1}{2}}(-1)^n(2)^{2n+1}(\pi)^{1/2}\frac{\Gamma(n+1)}{\Gamma(n+\frac{1}{2})} \\ &= (\pi)^{1/2}\frac{\Gamma(n+1)}{\Gamma(n+\frac{3}{2})}\end{aligned} \tag{C.14}$$

Now we may fully determine the surface area of a $K$-dimensional hypersphere,

$$\begin{aligned}A_K(r) &= 2\pi r^{K-1}\prod_{k=2}^{K-1}\int_0^\pi \mathrm{d}\theta_k\, \sin^{k-1}\theta_k \\ &= \frac{2(\pi)^{K/2}\, r^{K-1}}{\Gamma(\frac{K}{2})}\end{aligned} \tag{C.15}$$

## C.2 Random Sampling in Spherical Space

Suppose we have any vector in Cartesian space, $\boldsymbol{x}$, that points to the surface of a (hyper)cube. For any such cube, we can constrain vectors such as $\boldsymbol{x}$ to the surface of a sphere of radius $r$ at the same origin. In performing such a projection, we are constraining the norm of $\boldsymbol{x}$ to be equal to the radius at all times, and as such we consider the following,

$$\boldsymbol{y} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}r \tag{C.16}$$

Here the variables $\boldsymbol{y}$ are the result of constraining the norm of $\boldsymbol{x}$ to the radius, so the new variables result in $\|\boldsymbol{y}\| = r$. This is essentially the projection of $\boldsymbol{x}$ onto $\boldsymbol{y}$, $\mathrm{proj}_{\boldsymbol{y}}\,\boldsymbol{x}$, and given that they are parallel, it merely reduces to a constraint on the norm of $\boldsymbol{x}$, which constrains it to the surface of a sphere of radius $r$. In essence, this is a *projection onto a spherical surface*.

We may conceptualise this in terms of sampling. Supposing that $\boldsymbol{x}$ is a random vector in Cartesian space, then we may consider samples which are sampled on a spherical surface to be constrained via equation (C.16) (Von der Linden et al. 2014, p. 519). In that case, we project $\boldsymbol{x}$ onto the sphere of radius $r$,

$$\begin{aligned}p(\boldsymbol{y}\,|\,\kappa) &= \int \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{y}\,|\,\boldsymbol{x})\, p(\boldsymbol{x}) \\ &\overset{\text{LI}}{=} \int \mathrm{d}\boldsymbol{x}\left[\prod_{k=1}^{K}\delta\big(y_k - \tfrac{r}{\|\boldsymbol{x}\|}x_k\big)\right]p(\boldsymbol{x}) \\ &= \int_0^\infty \mathrm{d}s \prod_{k=1}^{K}\left[\int \mathrm{d}x_k\, \delta\big(y_k - x_k\tfrac{1}{s}\big)\, p(x_k)\right]\delta\big(s - \tfrac{\|\boldsymbol{x}\|}{r}\big) \\ &= \int_0^\infty \mathrm{d}s\, r s^K p(s\boldsymbol{y})\, \delta\big(s(r - \|\boldsymbol{y}\|)\big) \qquad \text{(C.17)} \\ &= r\,\delta\big(r - \|\boldsymbol{y}\|\big)\int_0^\infty \mathrm{d}s\, s^{K-1}\, p(s\boldsymbol{y}) \qquad \text{(C.18)}\end{aligned}$$

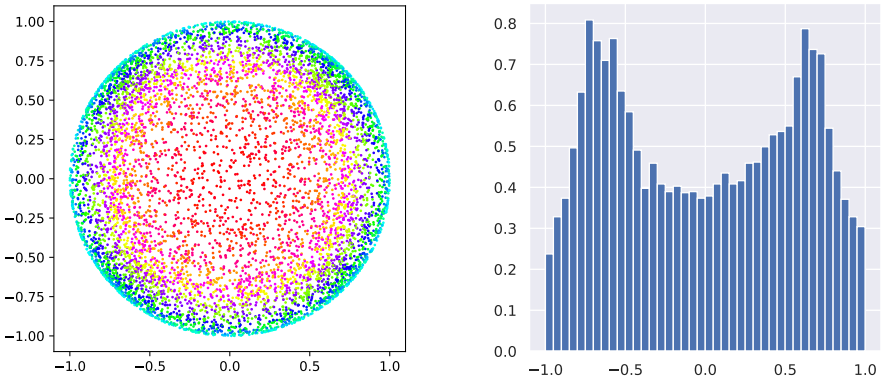
The PDF $p(s\boldsymbol{y})$ may then be chosen for whichever sample of random variables may be desired, which will be sampled in spherical space and on the surface of a sphere.

### C.2.1 Uniform Variables on a Spherical Surface

A natural assumption may be to consider sampling from a uniform distribution and constraining it to the surface of a sphere. In that case, we consider $p(s\boldsymbol{y}) = U(sy_k, -1, 1)$, resulting in

$$\begin{aligned} p(\boldsymbol{y} \mid \kappa) &\overset{\text{LI}}{=} r\,\delta(r - \|\boldsymbol{y}\|) \int_0^\infty \mathrm{d}s\, s^{K-1} \prod_{k=1}^{K} U(sy_k, -1, 1) \\ &= r\,\delta(r - \|\boldsymbol{y}\|) \int_0^\infty \mathrm{d}s\, s^{K-1}\, 2^{-K} \prod_{k=1}^{K} \tfrac{1}{2}\big(\Theta(s + \tfrac{1}{y_k}) - \Theta(s - \tfrac{1}{y_k})\big) \\ &= 2^{-K}\, r\,\delta(r - \|\boldsymbol{y}\|) \int_0^{M^{-1}} \mathrm{d}s\, s^{K-1} \\ &= \frac{r}{K2^K} \delta(r - \|\boldsymbol{y}\|) M^{-K} \end{aligned} \tag{C.19}$$

Here $M = \max_k(x_k)$. This is the result when sampling uniform random variables on the surface of a hypercube and constraining it to the surface of a sphere — simply standard uniform random variables projected onto the surface of the $K$-dimensional hypersphere. The result is such that, if considering the unit sphere, then if $\boldsymbol{x}$ points along a coordinate axis we have that $M^{-K} = 1$ and if it has equal contributions in all coordinates we have that $M^{-K} = K^{K/2}$, which implies that samples will be sparse if the vector points perpendicular to one of the faces of the hypercube, while more densely populated if the vector points towards the corners of the hypercube (Von der Linden et al. 2014, p. 521)




(i) Samples projected onto the $xy$-plane. (ii) Histogram of samples along the $x$-axis.

Figure C.1: 7500 samples from the uniform distribution on the 3-dimensional unit sphere. Colour coding indicates different contours of latitude on the sphere.

We can clearly observe that the samples are *not* uniformly distributed on the sphere, meaning that a uniform distribution does not maintain spherical symmetry and is thus not rotationally symmetric in spherical space.

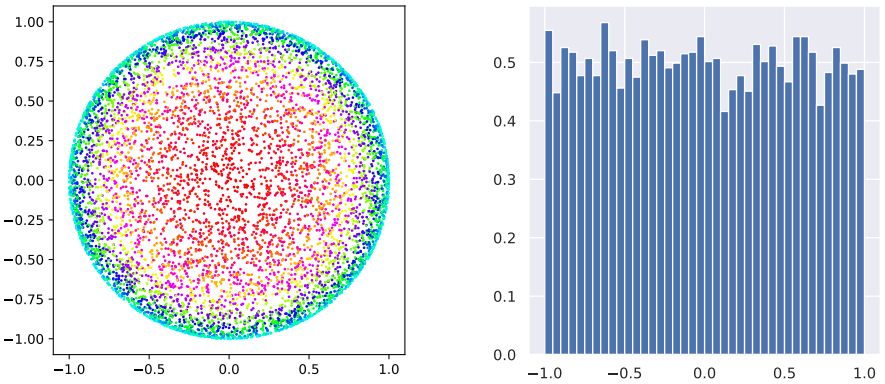
### C.2.2 Gaussian Variables on a Spherical Surface

We now want to obtain a spherically symmetric solution to equation (C.18), in doing so we would maintain the principle of indifference in spherical space which in turn would lead to an

uninformative prior, we therefore opt to sample from a Gaussian distribution, $p(s\boldsymbol{y}) = \mathcal{N}\big(s\|\boldsymbol{y}\| \,\big|\, 0, 1\big)$. In projection onto the sphere we have

$$\begin{aligned} p(\boldsymbol{y} \,|\, \kappa) &= r\,\delta(r - \|\boldsymbol{y}\|) \int_0^\infty \mathrm{d}s\, s^{K-1} (2\pi)^{-K/2} \exp\Big(-\tfrac{1}{2}s^2 \sum_{k=1}^{K} y_k^2\Big) \\ &= r\,\delta(r - \|\boldsymbol{y}\|) \int_0^\infty \mathrm{d}s\, s^{K-1} (2\pi)^{-K/2} e^{-\frac{1}{2}(sr)^2} \\ &= (2\pi)^{-K/2}\, r\,\delta(r - \|\boldsymbol{y}\|) \int_0^\infty \mathrm{d}v\, \tfrac{\sqrt{2}}{r} \Big(\tfrac{\sqrt{2}}{r} v\Big)^{K-1} e^{-v} \\ &= \frac{\delta(r - \|\boldsymbol{y}\|)}{A_K(r)} \end{aligned} \tag{C.20}$$

Here we see that the normalisation is given by the surface area of the hypersphere, moreover we note that symmetry can be observed,




(i) Samples projected onto the $xy$-plane. (ii) Histogram of samples along the $x$-axis.

Figure C.2: 7500 Gaussian distributed samples on the 3-dimensional unit sphere. Colour coding indicates different contours of latitude on the sphere.

The Gaussian samples are indeed unform on the surface of a hypersphere. For this we conclude that the Gaussian distribution can be chosen as an unbiased prior in spherical space.

### C.2.3 Symmetric Samples Within a Sphere

We may now take this one step further moving from the surface of the sphere to within the sphere. Traversing the volume of the sphere implies moving along the radius and maintaining spherical symmetry while considering the spherical volume means that the distribution has to be uniform at all points along the radius. This means to account for the variation of volume near the equator in comparison to variation near the poles.

The Volume of the $K$-dimensional hypersphere of radius $r$ is given by

$$\begin{aligned} V_K(r) &= \int_0^r \mathrm{d}R\, A_K(R) \\ &= \frac{2(\pi)^{K/2}\, r^K}{K\Gamma\big(\frac{K}{2}\big)} \end{aligned} \tag{C.21}$$

And we can note that the volume grows as $r^K$ with the radius. Note that the volume is more concentrated closer to the surface compared to the centre and longer radii essentially require more samples*.

In traversing the volume of the sphere, we can *project onto the radius* and in doing so we have that

$$\begin{aligned} p(\boldsymbol{y} \mid \kappa) &= \int \mathrm{d}\boldsymbol{x} \int_0^\infty \mathrm{d}R\, p(\boldsymbol{y} \mid \boldsymbol{x}, R, \kappa)\, p(\boldsymbol{x} \mid R, \kappa)\, p(R \mid \kappa) \\ &= (2\pi)^{-K/2} \int \mathrm{d}\boldsymbol{x}\, e^{-\frac{1}{2}\|\boldsymbol{x}\|^2} \int_0^\infty \mathrm{d}R\, p(R) \prod_{k=1}^{K} \delta\big(y_k - x_k \tfrac{R}{\|\boldsymbol{x}\|}\big) \\ &= (2\pi)^{-K/2} \int \mathrm{d}\boldsymbol{x}\, e^{-\frac{1}{2}\|\boldsymbol{x}\|^2} \int_0^\infty \mathrm{d}s\, \|\boldsymbol{x}\|\, p(s\|\boldsymbol{x}\|) \prod_{k=1}^{K} \delta(y_k - x_k s) \\ &= \frac{p(\|\boldsymbol{y}\|)}{A_K(\|\boldsymbol{y}\|)} \\ &= \frac{p(r)}{A_K(r)} \end{aligned} \tag{C.22}$$

Now, in order to maintain spherical symmetry in the hyperspherical volume, i.e. uniform sampling in terms of radii, we can think in terms of layers of spheres with respect to their surfaces. If a random variable is uniform on the surface of the $K$-dimensional sphere, then it will be uniform within the volume of the $(K-1)$-dimensional sphere†, and to sample uniformly in terms of radii, we are thus required to have a prior on $r$ which scales as $p(r) \propto r^{K-1}$ (Von der Linden et al. 2014, p. 522).

The results that we have obtained in equation (C.22) are more meaningful than they currently appear to be, and these results are explored in terms of modelling in Chapter 5 Spherical Symmetry (page 41).

* This can be thought of as considering the amount of points needed to fill the spherical shell beneath the surface of a sphere in contrast to the amount needed to fill a shell around its centre † This is easier to conceptualise in three dimensions: if a sphere is uniformly distributed and a disc is extracted from that sphere, then that disc will also be uniformly distributed.

# Bibliography


Aczel, J. (2006). *Lectures on Functional Equations and Their Applications*. Second edition. Dover Publications. ISBN: 978-0486445236.

Ahmed, N., T. Natarajan, and K.R. Rao (1974). "Discrete Cosine Transform". In: *IEEE Transactions on Computers* C-23 (1). URL: https://ieeexplore.ieee.org/document/1672377.

Aho, A.V. and J.D. Ullman (1994). *Foundations of Computer Science: C edition*. First edition. W.H. Freeman. ISBN: 978-0716782841.

Akaike, H. (1974). "A New Look at Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19 (6). URL: https://ieeexplore.ieee.org/abstract/document/1100705.

Andrews, G.E., R. Askey, and R. Roy (1999). *Special Functions (Encyclopedia of Mathematics and its Applications)*. First edition. Cambridge University Press. ISBN: 978-0521623216.

Anton, H. and C. Rorres (2005). *Elementary Linear Algebra*. Ninth edition. John Wiley & Sons. ISBN: 978-0471669593.

Bateman, H. et al. (1953). *Higher Transcendental Functions*. First edition. Vol. 1. McGraw-Hill. ISBN: 978-0486446141.

Burnham, K.P. and D.R. Anderson (2002). *Model Selection and Multimodel Inference*. Second edition. Springer. ISBN: 978-0387953649.

Cox, R.T. (1946). "Probability, Frequency and Reasonable Expectation". In: *American Journal of Physics* 14 (1).

— (1961). *Algebra of Probable Inference*. Second edition. The Johns Hopkins University Press. ISBN: 978-0801869822.

De Kock, M.B. (2014). "From Stable Priors to Maximum Bayesian Evidence via a Generalised Rule of Succession". PhD thesis. Stellenbosch University.

De Kock, M.B. and H.C. Eggers (2017a). *Bayesian Model Selection for Misspecified Models in Linear Regression*. Version 2. ArXiv: 1706.03343v2. URL: https://arxiv.org/pdf/1706.03343.pdf.

— (2017b). "Bayesian Variable Selection with Spherically Symmetric Priors". In: *Communications in Statistics - Theory and Methods* 46 (9). URL: https://www.tandfonline.com/doi/abs/10.1080/03610926.2015.1081945.

Dutka, J. (1984). "The Early History of the Hypergeometric Function". In: *Archive from History of Exact Sciences* 31 (1).

Graham, R.L., D.E. Knuth, and O. Patashnik (1989). *Concrete Mathematics. A Foundation for Computer Science*. Fourth edition. Addison-Wesley. ISBN: 978-0201558029.

Jaynes, E.T. (2003). *Probability Theory. The Logic of Science*. First edition. Cambridge University Press. ISBN: 978-0521592710.

Kekre, H.B. and J.K. Solanki (1978). "Comparative Performance of Various Trigonemetric Unitary Transforms and Transform Image Coding". In: *International Journal of Electronics* 43 (3).

Kolmogorov, A.N. (1956). *Foundations of the Theory of Probability*. Second edition. Original: *Grundbegriffe der Wahrscheinlichkeitrechnung* (1933). Chelsea Publishing Co. ISBN: 978-0828400237.

Liang, F. et al. (2007). "Mixtures of $g$-priors for Bayesian Variable Selection". In: *Journal of American Statistical Association* 103 (481).

Mackay, D.J.C. (2003). *Information Theory, Inference, and Learning Algorithms*. First edition. Cambridge University Press. ISBN: 978-0521642989.

McElreath, R. (2015). *Statistical Rethinking. A Bayesian Course with Examples in R and Stan*. First edition. Chapman and Hall/CRC. ISBN: 978-1482253443.

Paris, J. (1995). *The Uncertain Reasoner's Companion. A Mathematical Perspective*. First edition. Cambridge University Press. ISBN: 978-0521032728.

Pólya, G. (1941). "Heuristic Reasoning and the Theory of Probability". In: *The American Mathematical Monthly* 48 (7).

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *Annals of Statistics* 6 (2).

Skilling, J. (2005). "Bayesics". In: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering. 25$^{th}$ International Worskhop*. Ed. by K.H. Knuth et al. American Institute of Physics. ISBN: 978-0735402928.

Stewart, J. (2009). *Calculus*. Sixth edition. Brooks/Cole Cengage Learning. ISBN: 978-0495011606.

Terenin, A. and D. Draper (2017). *Cox's Theorems and the Jaynesian Interpretation of Probability*. Version 1. ArXiv: 1507.06597. URL: https://arxiv.org/pdf/1507.06597.pdf.

Von der Linden, W., V. Dose, and U. von Toussaint (2014). *Bayesian Probability Theory. Applications in the Physical Sciences*. First edition. Cambridge University Press. ISBN: 978-1107035904.

Watson, G.N. (1922). *Theory of Probability*. Second. Cambridge University Press. ISBN: 978-9333324083.

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with $g$-prior Distributions". In: *Bayesian Inference and Decision Techniques. Essays in Honour of Bruno de Finetti*. Ed. by P.K. Goel and A. Zellner. North Holland/Elsevier, pp. 233–243. ISBN: 978-0444877123.

# Glossary

**Akaike's Information Criterion**

A criterion for model selection, defined as $NQ(\widehat{\boldsymbol{\alpha}}) + 2K$.

**Bayesian Information Criterion**

A criterion for model selection, defined as $NQ(\widehat{\boldsymbol{\alpha}}) + N \log K$.

**Chi-squared criterion**

The chi-squared criterion is a loss function which measures the discrepency between a candidate model and the data, $NQ(\boldsymbol{\alpha}) = (\boldsymbol{z} - \boldsymbol{\zeta})^{\mathsf{T}}(\boldsymbol{z} - \boldsymbol{\zeta})$

**Construct**

In context of simulation, the construct is the underlying function involved in the generation of data.

**Evidence**

The evidence is the average over all possible parameter assignments for a given likelihood and prior, and provides a measure as to how a well the model represents the data.

**Identically and Independently Distributed**

Samples or data which are independent and distributed according to the same distribution

**Likelihood**

The likelihood, $p(\boldsymbol{z} \mid \boldsymbol{\alpha})$, is considered to be our model of data; it is a representation of the data if it were to be parametrised according to $\boldsymbol{\alpha}$.

**Maximum Likelihood Estimate (MLE)**

Point which maximises the likelihood/minimises chi-squared.

**Mean Squared Error (MSE)**

Mean of the squared error over several trials.

**Oracle estimate**

The Oracle estimate is the theoretical best prediction when evaluating information criteria. The Oracle evaluates criteria with the knowledge of the construct order against the model order, i.e. C = K.

**Order (model order)**

The number of parameters in a model. Denoted by $K$ for models and in the context of simulation, by $C$ for the construct.

**Overarching model**

The model of order $\overline{K}$, which is prescribed to the overaching radius

**Overarching radius**

The radius spanning between all models in spherically symmetric space.

**Posterior**

The posterior represents one's inference about the parameters (or the model) and it is one's current state of knowledge in light of the data.

**Prior**

The prior, $(\boldsymbol{\alpha})$, The *prior* is our assignment of the parameters of our (likelihood) model. It represents our conviction in the parameters without any consideration of the data.

**Probability Density Function (PDF)**

Probability distribution defined in continuous space.

**Probability Mass Function (PMF)**

Probability distribution defined in discrete space.

**Radial scale**

The scaling relation between the radii $R_K$ of respective models and the overarching radius $r$, represented by the parameter $\ell$.

**State of uncertainty**

The measure of the conviction in or the plausibility of a proposition, denoted $\psi(x)$.

# Index

"*There is nothing noble in being superior to your fellow man; true nobility is being superior to your former self.*"

E. Hemingway