An Objective Bayesian Approach for Discrete Scenarios

Cristiano Villa

School of Mathematics, Statistics and Actuarial Sciences
University of Kent, Canterbury

A Thesis Submitted for the Degree of
Doctor in Philosophy

September, 2013

Acknowledgement

First and foremost I would like to thank my supervisor, Prof. Stephen Walker. He has been a source of knowledge, experience and ideas during these three years. And a friend.

Thank you to the wonderful persons I met during this journey: Isadora, Xue and Claire. From a chat to a coffee, to a suggestion, you rendered the work somehow easier and more pleasant.

My thank goes to the School, for providing the funding (in conjunction with ESPRC) for the Ph.D.

To my wife, for coping with the change of course that these years meant to us. And for the ones to follow.

Abstract

Objective prior distributions represent a fundamental part of Bayesian inference. Although several approaches for continuous parameter spaces have been developed, Bayesian theory lacks of a general method that allows to obtain priors for the discrete case.

In the present work we propose a novel idea, based on losses, to derive objective priors for discrete parameter spaces. We objectively measure the *worth* of each parameter values, and link it to the prior probability by means of the *self-information* loss function. The *worth* is measured by taking into consideration the surroundings of each element of the parameter space. Bayes theorem is then re-interpreted, where prior and posterior beliefs are not expressed as probabilities, but as losses. The approach allows to retain meaning from the beginning to the end of the Bayesian updating process. The prior distribution obtained with the above approach is identified as the Villa–Walker prior.

We illustrate the approach by applying it to various scenarios. We derive objective priors for five specific models: a population size model, the Hypergeometric and multivariate Hypergeometric models, the Binomial-Beta model, and the Binomial model. We also derive the Villa–Walker prior for the number of degrees of freedom of a t distribution. An important result in this last case, is that the objective prior has to be truncated.

We finally apply the idea to discrete scenarios other that parameter spaces: model selection, and variable selection for linear regression models. We show how an objective model prior can be obtained, by applying our approach, on the basis of the importance that each model has with respect to the other ones. We illustrate various cases: nested and non-nested models, models with discrete and continuous supports, uniparameter and multiparameter models. For the variable selection scenario, the prior includes a loss component due to the complexity of each regression model.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Objective Bayes represents an important aspect of Bayesian analysis and, more in general, of statistical inference. The motivations behind objective Bayesian procedures can be different, but all originate from the same assumption: there is little or no prior knowledge about the quantity of interest; or, as it may be also the case, the knowledge is intentionally ignored. It is not our intention to contribute to the debate about the "legitimacy" of objective Bayes (debate far from being over). Detailed discussions on the matter can be found, for example, in Berger (2006). Our work focuses on Bayesian objective methods for discrete parameters where, according to the literature, there is a lack of a general approach for defining prior distributions. We believe that the void can be filled by solving a foundational gap affecting objective Bayes: probabilities cannot be directly obtained through objectivity. Instead, we claim that they have to be derived through the objective definition of loss functions.

The Bayesian framework can be formalised as follows. Let us consider the Bayesian model $M = \{f(x|\theta), \pi(\theta)\}$: $f(x|\theta)$ represents a family of probability distributions chosen to model exchangeable or independently and identically distributed (i.i.d.) outcomes; $\pi(\theta)$ is the prior distribution representing the initial *guess* with respect to the true value of the unknown parameter $\theta \in \Theta$. Bayesian inference is then performed on the basis of the posterior distribution

$$\pi(\theta|x) \propto f(x|\theta) \times \pi(\theta), \tag{1.1}$$

where the initial *guess* about θ is updated on the basis of the information gained from an experiment, expressed by the likelihood $f(x|\theta)$.

We assume that the densities exist, with respect to some measure on $\mathcal{X} \times \Theta$, where $\mathcal{X}$ is the support of $f(x|\theta)$ ($x \in \mathcal{X}$), and Θ is the parameter space. For simplicity in the notation, π indicates both the prior and the posterior; the context will give indications on which one is discussed. Furthermore, x represents both the random variable from which the observations are drawn in an experiment, and the vector of observations itself: $x = (x_1, \ldots, x_n)$; θ can be a scalar or a vector of parameters: $\theta = (\theta_1, \ldots, \theta_d)$.

Here $\pi(\theta)$ in (1.1) represents the initial uncertainty we have about the true value of θ, and can be defined in two ways: subjectively or objectively. The former presumes some knowledge about θ prior to the experiment. The method to subjectively obtain the prior distribution are beyond the scope of this thesis, and therefore not discussed; discussions about subjective Bayes can be found, among others, in Ramsey (1964), de Finetti (1937), Lindley (1972), French (1982) and Goldstein (2006). In the objective approach, the idea is to have a procedure that, free from personal considerations, allows one to define $\pi(\theta)$ once $f(x|\theta)$ has been chosen. This case constitutes the main topic of the thesis.

If prior distributions are the building blocks of the Bayesian approach, objective priors represent one of the cornerstones. Even though it is appealing (and advisable) to rationally take advantage of any suitable prior information that may be available, this is not always feasible. In some circumstances there is no such initial information; and in others, even though this knowledge is theoretically available, it might be prohibitive even to think about using it. As an example, consider complex and large models, where the number of parameters can be easily of the magnitude of hundreds or thousands. It would be unrealistic to think that a subjective definition of the prior for each one of these parameters can be performed.

The literature about objective priors is vast. Several general methods to objectively obtain $\pi(\theta)$ have been designed: Jeffreys' prior (Jeffreys, 1961), reference priors (Bernardo, 1979), Probability Matching Priors (Welch and Peers, 1963), among others.

When θ is discrete, solutions to find an objective prior tend to be problem specific. No effective general cases have been so far proposed. For this reason, we focus on discrete scenarios, and propose a general approach that may be applied to any model for which the parameter takes values on a discrete space. For example, the number of trials n in a Binomial model; or the number of populations units R that have a certain property in a Hypergeometric distribution. Furthermore, the approach is extended to other discrete problems: model selection and variable selection for linear regression models. In fact, procedures to assign prior mass to each model, in an objective way, can be defined.

A second aspect about objective priors is that, in general, they are improper. In practice, this does not constitute an issue, as long as the posterior is proper, thus suitable for inference. The marginalisation paradox that may rise from the use of improper priors (Dawid et al., 1973), has been overestimated (Berger, 2006). In fact, objective improper priors have been and still are widely used. However, a conceptual gap remains: the prior and posterior do not represent the same "thing", as the posterior represents probabilities while the prior does not. Therefore, if we regard at the Bayesian procedure as a process with an input (the prior) and an output (the posterior), there is no retention of meaning from one end to the other. Attempts to justify this incongruence, mainly from a probabilistic point of view, have been made. Our method gives a new view of the problem resolving the conceptual gap.

The idea we propose is simple and it is the following. Instead of representing initial beliefs by probabilities directly, we objectively represent them through losses and, by means of the *self-information* loss function, derive the prior mass. Recall that objectivity arises from the absence of knowledge, actual or alleged, about the true value of the parameter, we can see the justification of this approach, as we can still have an idea of the *worth* that each parameter value represents in the model. And by assigning the mass to each parameter value by a measure of its *worth*, we are not subject to the constraints of properness, intrinsic in a probability measure.

The *worth* of an element of the parameter space can be assigned by answering the following question: "What do we lose, if an element of the parameter space is removed and it is the true one?" More formally, let us consider the prior distribution $\pi(\theta)$ for the discrete parameter $\theta \in \Theta$. If a prior mass π has been assigned then we link this to a *worth* by means of the *self-information* loss function $-\log \pi(\theta)$ (Merhav and Feder, 1998). We can then find an objective way to associate a loss to each θ, representing its *worth* in the model line-up, and the prior distribution $\pi(\theta)$ then follows. Furthermore, we note that in this way the Bayesian approach is conceptually consistent, as we update an initial *worth* assigned to θ, through the application of Bayes' theorem, to obtain the resulting *worth* expressed by $-\log \pi(\theta|x)$. Indeed, there is an elegant procedure akin to Bayes which works from a loss point of view, namely that

$$-\log \pi(\theta|x) = K - \log f(x|\theta) - \log \pi(\theta),$$

which has the interpretation of

$$\text{Loss}(\theta|x, \pi) = K + \text{Loss}(\theta|x) + \text{Loss}(\theta|\pi).$$

This is a cumulative loss function for assessing the loss of θ in the presence of two pieces of mutual information x and π. Here K is a constant which does not depend on θ.

The next part that we have to clarify is how the *worth* is objectively assigned. The *worth* to be assigned to each model is equal to the Kullback–Leibler divergence (Kullback and Leibler, 1951) measured from the model to the nearest one. This is justified by the fact that, if the model is misspecified (which it would be if we remove θ and it turned out to be the true value), the posterior distribution accumulates asymptotically at the nearest model with respect to the Kullback–Leibler divergence (Berk, 1966); also, refer to Theorem 3.1. Thus, this divergence represents the loss incurred by removing the model, and is the true one, and this will be the quantification of the *worth* of that model. The objectivity of this measure is obvious, as it will depend on the available set of options (i.e. choice of the family of densities) solely. Thus, we have that the utility of keeping θ in Θ is $u(\theta) = D_{KL}(f(x|\theta)\|f(x|\theta'))$, where $D_{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence (refer to Section 2.1.1). We can therefore associate a loss to each parameter value as

$$l(\theta) = -D_{KL}(f(x|\theta)\|f(x|\theta')),$$

representing the loss in keeping θ in the space Θ. We link this measure of the *worth* of θ via the *self-information* loss function by setting $-\log \pi(\theta)$, and the resulting prior is

$$\pi(\theta) \propto \exp\left\{D(f(x|\theta)\|f(x|\theta'))\right\}.$$

Outline of the work

In Chapter 2 we present a review of the current objective approaches to derive prior distributions for parameter spaces. It has three sections. The first one discusses motivations for an objective Bayesian approach, criticisms, and a general discussion on improper priors. The second section discusses three approaches for continuous parameter spaces: Jeffreys' prior, reference priors and probability matching priors. The last section of the chapter refers to the challenges in defining objective priors for discrete parameter spaces, illustrating the current methodologies. Chapter 3 contains the main result of the thesis: the novel approach we propose that allows to design objective priors for discrete parameter spaces. After a section discussing our motivations, we briefly discuss loss functions in general and, in particular, the *self-information* loss function. The last section of the chapter presents the formal definition of our approach.

The following two chapters provide examples of the application of our approach to specific models. Chapter 4 focuses on five particular models discussed in Berger et al. (2012). In Chapter 5 we show how our approach can be applied in estimating the number of degrees of freedom of a t density. In the chapter we also provide the analysis of the posterior, both for i.i.d. samples and a regression model with t-distributed errors.

Chapter 6 is a first example of the application of our objective approach to model selection problems. We first briefly review major objective approaches, then present our method with some illustrations: nested and non-nested models, discrete and continuous supports, uniparameter and multiparameter models. Chapter 7 refers to a particular type of model selection scenario: variable selection in linear regression models. Our prior is derived and its use is illustrated on a real data situation. Comparison with other two objective priors is carried out on the basis of marginal posterior inclusion probabilities. The chapter includes also some interesting results that, although not directly relevant to the work discussed in this thesis, are noteworthy and should drive future investigation.

Finally, Chapter 8 provides a general discussion of the overall results of the thesis, including some ideas for future work. In particular, possible extension of our approach to continuous parameter spaces and other selection problems (polynomial regression and mixture models, for example). In each of the chapters from number 4 to number 7 we present a discussion of the specific results obtained therein.

Chapter 2

Background

In this Chapter, we review and discuss some of the objective procedures to assign prior probability to parameter spaces. To set the appropriate context for this work it is paramount not only to examine these process from a mathematical point of view, but to grasp what is the motivation behind their development.

We start by presenting some historic facts about the origins of Bayesian inference. We show that, in fact, this has been the first type of inference. We move then to discuss invariance property, which represents the trigger for the development of Jeffreys' priors. Reference analysis and probability matching priors are presented next, alongside with other less general objective approaches. Finally, as the core of this thesis is about discrete parameter spaces, we review some approaches designed specifically to deal with this type of scenarios.

2.1 Notation and initial considerations

The literature about objective Bayes is full of different terminology about prior distributions defined through *objective approaches*. Some words are non-informative prior, ignorant prior, vague prior. As these refer, at least in principle, to the same concept, we use throughout this work the adjective *objective* prior. By *objective* prior, we mean a prior distribution that is obtained through a procedure that does not involve subjective input after a model has been chosen.

We consider the model $M = \{f(x|\theta), \theta \in \Theta\}$, where $x \in \mathcal{X}$, with $\mathcal{X}$ being the support of $f(\cdot|\theta)$, and Θ the parameter space. The aim is to make inference about the value of the unknown parameter θ. In a Bayesian framework, this is achieved by obtaining a distribution of the parameter, *posterior* to the observation of a sample $x = (x_1, \ldots, x_n)$ drawn from $f(\cdot|\theta)$. The posterior is the result of the combination of the information contained in the sample, expressed by the likelihood function $f(x|\theta)$, and the prior $\pi(\theta)$ representing the uncertainty about θ

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\, d\theta}.$$

Throughout this thesis, we assume that all probability functions exist with respect to some reference measure on $\mathbb{X}^n \times \Theta$.

As it appears from the above, a key step is the definition of $\pi(\theta)$. In essence, there are two ways of doing this. If we possess sufficient and sensible information about θ, we can use it to elicit the prior. The information can come in various forms, such as expert knowledge, but in many circumstances it derives from historical data. A subjective prior is extremely powerful, if both the basis and the process of elicitation are robust and rational. However, the conditions for elicitation are not always possible, or realistic. Sometimes this prior information is not available, because there is no sufficient (or not at all) historical data; or because models are complex, in the sense that the number of parameters is too high to allow a sensible elicitation for each one of them.

There are also other more subtle motivations in deciding not to subjectively define $\pi(\theta)$, as detailed in Berger (2006). First, the idea of subjectivity in the non-scientific community creates the belief that the analysis does not bear the necessary scientific rigour. For example, there may be the concern that results not supported solely by experiments, which could be objectively replicated, are somehow the consequence of skilled "manoeuvres" intent to support biased outcomes. Objective Bayes can also be considered as a way of connecting with frequentist methods. In Bayarri and Berger (2004) there is an exhaustive review of literature aimed to support the argument that objective procedures are an interface between the Bayesian and the classical (i.e. frequentist) approach.

Besides motivations, in Berger (2006) it is also possible to find criticisms to objective Bayes. In addition to the alleged lack of scientific rigour mentioned above, doubts are driven by the presence of multiple objective methods which, in some cases, lead to different results. Thus, the necessity to choose between these methods is perceived as a weakness in the overall idea. The general approach for discrete parameter spaces (Berger et al., 2012) suffers from this issue.

2.1.1 Definitions

The objective approach we propose is based on an asymptotic property of the posterior distribution, when the model is misspecified, which involves the Kullback–Leibler divergence (Kullback and Leibler, 1951).

Definition 2.1 (Kullback–Leibler divergence). *The Kullback–Leibler divergence between probability mass functions $f(x|\theta)$ and $f(x|\phi)$ is given by*

$$D_{KL}(f(x|\theta)\|f(x|\phi)) = \sum_{\mathcal{X}} f(x|\theta) \log\left\{\frac{f(x|\theta)}{f(x|\phi)}\right\}.$$

If $f(x|\theta)$ and $f(x|\phi)$ are probability density functions, the Kullback–Leibler divergence has the form

$$D_{KL}(f(x|\theta)\|f(x|\phi)) = \int_{\mathcal{X}} f(x|\theta) \log\left\{\frac{f(x|\theta)}{f(x|\phi)}\right\}\, dx.$$

Objective prior distributions are, in general, improper.

Definition 2.2 (Improper distribution). *A probability mass function $f(x|\theta)$, with $x \in \mathcal{X}$ and $\theta \in \Theta$, is improper if*

$$\sum_{\mathcal{X}} f(x|\theta) = \infty.$$

If $f(x|\theta)$ is defined in the continuous, that is it is a probability density function, then it is improper if

$$\int_{\mathcal{X}} f(x|\theta)\, dx = \infty.$$

2.1.2 A few words on improper priors

Objective approaches lead in many circumstances to improper priors, in the sense that these distributions do not integrate (or sum, in the discrete case) to one. This happens because, as we want to represent as less knowledge as possible about the parameter value, the parametric space is often unbounded.

There are cases where objective priors are proper. For example, a commonly accepted objective prior for the parameter $\theta \in (0,1)$ of a binomial distribution, representing the probability of success, is $\pi(\theta) = Be(1/2, 1/2)$, where Be is the Beta density. However, a bounded parameter space is not *per se* a sufficient condition for having a proper objective prior. As an example, if we consider a Negative Binomial distribution with parameters (r, p), where $r > 0$ and $p \in (0,1)$, the usually recommended objective prior for p is $\pi(p) \propto p^{-1}(1-p)^{-1/2}$; this distribution, although the parameter space is bounded, turns out to be improper. Finally, there are scenarios where for an unbounded parameter space it is possible to have proper objective priors, as for the case of the ratio of two multinomial parameters, where the parameter space is $(0, \infty)$ (Bernardo, 1997).

Given that inference depends on the posterior, improper priors can be used in practice, as long as the posterior is proper. However, improper priors are not probability distributions, and they simply represent positive functions, that is a technical *device* to be used in Bayes theorem to obtain (proper) posterior distributions (Bernardo, 1997). But it is obvious that, conceptually, Bayes theorem no longer applies.

Berger et al. (2009) give a justification on the adoption of improper prior distributions. If an improper prior $\pi(\theta)$ is defined, then Bayes theorem does not apply and its use has to be justified. Berger et al. (2009) show that the posterior $\pi(\theta|x)$ is a suitable limit of posteriors obtained from proper priors. Consider the increasing sequence of compact sets of Θ, $\{\Theta_j\}_{j=1}^{\infty}$. The sequence of proper priors $\pi(\theta_j)$, defined on Θ_j, is called the approximating sequence of posteriors $\{\pi_j(\theta|x)\}_{j=1}^{\infty}$, approximating the formal posterior $\pi(\theta|x)$. Thus, the sequence of posteriors is said to be *expected logarithmically convergent* (to the formal posterior) if

$$\lim_{j\to\infty}\int_{\mathcal{X}} D_{KL}(\pi(\cdot|x)\|\pi_j(\cdot|x))f_j(x)\,dx = 0, \tag{2.1}$$

where $f_j(x) = \int_{\Theta_j} f(x|\theta)\pi_j(\theta)\,d\theta$. The conclusion is that a prior distribution satisfying the property in (2.1) yields a posterior that, in expectation, is an approximation of the formal posterior; in the sense that it approximates the posterior that would be obtained by restricting the sample space Θ to a large compact set.

2.2 Review for continuous parameter spaces

2.2.1 A brief discussion on the term *non-informative*

We do not wish to debate on etymological aspects, but we deem appropriate to spend a few words to clarify the meaning of *non-informative*, when it is referred to prior distributions. For an interesting discussion on the matters, refer to Bernardo (1997).

Bernardo and Smith (1994) pointed out that "there is no prior that represents ignorance". Every prior distribution carries some amount of information (although sometimes minimal), in the sense that it depends on the model that has been chosen. In fact, it is commonly agreed that *objectivity* is intended from the moment that the model has been selected to represent the quantity of interest. Therefore, when we refer to a prior distribution representing "ignorance", it has to be understood in the above sense.

Many terms have been used to label this type of distribution: *conventional*, *default*, *flat*, *formal*, *neutral*, *non-subjective* and *objective* (Bernardo, 1997). Independently on what expression we decide to adopt, there is some common agreement on what a prior which is not elicited should represents: derive posterior distributions, through Bayes theorem, where the contribution from the observations is as large as possible. In other words, it is the data that should dominate the scene.

2.2.2 Inverse Probability and the Uniform Prior

Even though in Bayes' essay (Bayes, 1763) the prior distribution is not mentioned, it is clear that, in his attempt to estimate an unknown probability, he used a continuous uniform distribution as the prior for the unknown parameter. A simplified illustration of the experiment described by Bayes is as follows. Consider a pool table of a length that conventionally we refer to as one. A ball is placed on it following a uniform distribution. What we have to guess is the distance of the ball from one of the ends of the table, say the left-end. To do this, we throw another ball on the table, and the number of times it is closer to the left-end than the other ball is counted. We use this information to make our guess.

The inference problem, in modern terms, is to estimate the parameter p of a Bernoulli distribution. To do this, even though not explicitly stated, Bayes puts a uniform prior on the parameter space: that is $\pi(p) \propto 1$, with $p \in [0,1]$. He then considers the likelihood of observing x successes, given p, which is the number of times the second ball gets closer to the left-end of the table than the first ball. Thus, he combines the prior information with the likelihood function $f(x|p)$ to obtain the posterior distribution of the parameter

$$\begin{aligned} \pi(p|x) &= \frac{f(x|p)\pi(p)}{\int f(x|p)\pi(p)dp} \\ &= \frac{p^x(1-p)^{n-x}}{\int_0^1 p^x(1-p)^{n-x}dp} \\ &= \frac{(n+1)!}{(n-x)!\,x!}p^x(1-p)^{n-x}, \end{aligned} \tag{2.2}$$

which is a Beta distribution with parameters $x+1$ and $n-x+1$. Note that, while the successes are independent when they are conditional on p, they are not when they are unconditional on p, i.e. marginalised.

Of course, the above result in (2.2) is the outcome of a reinterpretation process, where modern considerations have been made. Nevertheless, it seems that at least two key points can be noted. First, from this example it appears that Bayes intention was of starting with an initial guess and update this by consequent observations. This, as we know, is the core of the Bayesian framework. Second, his initial guess was made in a condition of total ignorance, and he has translated this ignorance in a probability distribution that treats equally each value in the parameter space; in other words, Bayes considered the uniform distribution as the distribution for ignorance.

This idea was developed, independently from Bayes, by Laplace a few years later; possibly in a more comprehensive and sophisticated manner (Laplace, 1774). We are not going to detail Laplace's contribution, as it would lead to the results we have outlined above. Besides a formal definition of Bayes theorem, as it is today known, he has clearly specified what a prior distribution representing ignorance should be. His idea that, if we know nothing about the value of a parameter there is no reason to assigning more mass to a value than another, took the name of *indifference principle* and it dominated the statistical inference scenery up to the birth of the frequentist approach. As mentioned in Fienberg (2006), Laplace started the statistical quest of finding prior distributions that reflect ignorance; a quest that it is still going on under the name of *objective* Bayes.

In his work, Laplace reinforced the concept that a prior distribution on the unknown parameter of a Bernoulli distribution p, which aims to represent ignorance, has to be uniform on the interval $[0, 1]$. He has also adopted the same approach for other cases, such as for location parameters. As reported in Fienberg (2006), Laplace has clearly expressed that the posterior distribution for a parameter θ, is proportional to the likelihood function, times the (uniform) prior distribution

$$\pi(\theta|x) \propto f(x|\theta).$$

The concept of uniform prior is implied.

The terminology *inverse probability* appearing in the title of this section, reflects the concept of inferring backwards from the data to the parameters or, as it can also be put, from the effect to the causes. However, the name came into use later and was used until the middle of the last century, when replaced by Bayesian inference (Fienberg, 2006).

2.2.3 Jeffreys

The main criticism about uniform priors is that they do not represent ignorance. Knowing nothing about θ and knowing that it can take any value with the same probability are two well distinct facts. The above criticism to uniform priors mainly came from the fact that, in general, they are not invariant under one-to-one reparameterisations. This property is by many seen as a must for an objective prior (Dawid, 1983; Jaynes, 1968; Bernardo, 1997). In particular, Jaynes asserts that the way a model is parametrised involves subjectivity; as such, a prior distribution that is influenced by this subjective choice, cannot be considered entirely objective. The state of knowledge about a model does not change by simply rearranging its parameters. Let us better understand the meaning of invariance under one-to-one reparameterisations. Consider a statistical model $f(\cdot|\theta)$ with the prior $\pi(\theta) \propto 1$, that is a uniform. If we do not have any knowledge about θ, we also do not have any knowledge about $1/\theta$. Therefore, by applying the change-of-variables formula for random variables on the one-to-one transformation $g(\phi) = 1/\theta$, we have

$$\pi(\phi) = 1 \cdot \left| \frac{d}{d(1/\theta)} g^{-1}(\phi) \right| = -\frac{1}{\phi^2},$$

which is not uniform.

In designing an objective approach to derive priors, Jeffreys stressed the importance that the resulting distributions were invariant under any one-to-one (differentiable) transformations. He then based his method on the considerations that Fisher information (Edgeworth, 1908), $I(\theta)$, is a quantification of the amount of information about the parameter θ that is expressed by the model, and that it is invariant under these type of transformations. Fisher information is defined as

$$I(\theta) = -\mathbb{E}_\theta \left\{ \frac{d^2}{d\theta^2} \log f(x|\theta) \right\},$$

where $\mathbb{E}_\theta$ is the expectation with respect to model $f(x|\theta)$, and $\log f(x|\theta)$ is the log-likelihood function. For example, Casella and Lehmann (1998) show that, if $\psi = h(\theta)$ and θ are two parameterisation of the same estimation or decision problem, and ψ is a continuously differentiable function of θ, then

$$I(\theta) = I(\psi)(h'(\theta))^2, \tag{2.3}$$

where $h'(\theta)$ represents the derivative of $h(\theta)$ with respect to ψ. Expression (2.3) links the Fisher information of the parameterisation θ and the one of the parameterisation ψ. Thus, by taking the square root of both parts in (2.3), we have

$$I(\theta)^{1/2} = I(\psi)^{1/2}|h'(\theta)|.$$

Therefore, the prior for θ will be linked to the prior for its reparameterisation ψ by

$$\pi(\theta) \propto I(\theta)^{1/2} = \pi(\psi)|h'(\theta)|, \tag{2.4}$$

which is the well known expression of Jeffreys prior, that is *the square root of the determinant of Fisher information.* On the right-hand-side of (2.4) it is possible to recognise the transformation formula, showing Jeffreys prior invariance property.

As an illustration, Jeffreys prior for the parameter $\theta \in (0,1)$ of a Binomial distribution with known n, is given by $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$; that is, a Beta with both parameters equal to $1/2$. Also, if we consider a Normal distribution with unknown mean μ and known variance, it can be shown that $\pi(\mu) \propto 1$; showing that the uniform prior can still be a valid objective prior, in the sense that is complies to the desiderata of being invariant under one-to-one reparameterisations.

An important limit of these type of priors, noticed by Jeffreys himself, is that in general it does not lead to acceptable results when applied to a vector of parameters. Let us consider a distribution function $f(x|\theta)$, where $\theta = [\theta_1, \dots, \theta_d]^T$ is a vector of d parameters. The Fisher information matrix for this vector of parameters is given by

$$(I(\theta))_{i,j} = -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x|\theta)\right\},$$

Thus, Jeffreys prior for the multiparameter case can be found by taking the square root of the determinant of the Fisher information matrix, that is

$$\pi(\theta) \propto det(I(\theta))^{1/2}.$$

The prior obtained according to Jeffreys' rule for the unknown parameters (μ, σ) of a normal distribution is $\pi(\mu, \sigma) \propto 1/\sigma$; this prior has poor convergence performance (Chopin et al., 2009).

To overcome this weakness, Jeffreys suggested to consider the two parameters as independent *a priori*: $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma) \propto 1/\sigma^2$ which has desirable properties. To distinguish between the two priors, we call the first one as *Jeffreys' rule* prior (as it has been obtained applying directly Jeffreys' method), whilst the second (assuming parameter independent a priori) is called *Jeffreys independent* prior.

2.2.4 Reference priors

We examine in detail reference priors because, as we will see in Section 2.3, they represent an important building block of what can be considered the more evolute general approach for deriving objective priors for discrete parameter spaces that can be currently found in the literature.

Approaches based on Jeffreys' method where used to deal with multiparameter problems until the early 70's (Bernardo, 1997), when marginalisation paradoxes began to emerge. Up to then, no particular issues were identified in using improper priors, such as Jeffreys', in Bayesian inference. It appeared that having to deal with a proper posterior was sufficient, independently of the prior used (i.e. proper or improper). These paradoxes, presented and discussed in a systematic way in Dawid et al. (1973), show that the use of improper priors in multiparameter problems may lead to marginal posterior distributions that do not posses Bayesian properties, as it would be the case if proper priors would be instead adopted. In Berger (2006) and Berger and Sun (2006), there are interesting discussions about the marginalisation paradoxes. In particular, it is agreed that the avoidance of these paradoxes, through the design of appropriate priors, may not be a fundamental task. It is in fact possible to find optimal posteriors even though they suffer from the paradox. Conversely, there are posteriors (also coming from subjective priors) far from being good which are free from the paradox. As such, we decided not to further pursue this topic.

Even though reference priors were not a direct answer to the above paradox (Bernardo, 1997), they allow one to deal with multiparameter problems avoiding the marginalisation paradoxes. Based on the work of Lindley (1956), who first thought about using information theory concepts to *measure* the difference in information between prior and posterior, Bernardo (1979) laid the groundwork for reference priors. The work was subsequently developed and structured and, finally, grouped under the name of *reference analysis*. Extensive reference on the subject can be found in Berger and Bernardo (1989, 1992a,b), Clarke and Sun (1997, 1999) and Berger et al. (2009).

The basic idea of the reference prior is as follows. The posterior distribution, as known, is the "combination" of the prior knowledge about the parameter and the likelihood. Therefore, if we measure the difference in information between the posterior and the prior, this difference can only be the information about the parameter (the unknown quantity of interest) that is contained in the data. We have already mentioned that the aim of an objective approach is to obtain posterior distributions where the contribution of the data is as large as possible. And this can be interpreted as defining a prior such that the difference in information between posterior and prior, called the *missing information*, is maximised (in expectation). An important contribution of reference priors is that they allow, through a stepwise procedure, to deal with multidimensional parameter spaces, where only a number of them are considered of interest, and the remaining are considered as *nuisance parameters*. If there are no nuisance parameters (plus certain regularity conditions are satisfied) and, in particular, in the one-dimensional case, reference priors coincide with Jeffreys' rule prior. When there are nuisance parameters, the reference priors will in general differ from Jeffreys' rule prior.

We start by presenting the case of a model with one parameter only, as for this situation reference priors are defined without heuristic components. We then extend to the multiparameter case where, as Berger et al. (2009) state, not all definitions and theorems are supported by non-heuristic arguments as for the uniparameter case. Reference priors in the presence of nuisance parameters are discussed in this extension.

Missing information

The notion of missing information introduced by Bernardo (1979) is based on the concept of gain in information provided by an experiment, discussed by Lindley (1956).

Let us consider a set of observations from a statistical distribution $f(x|\theta)$, with $\theta \in \Theta$ is an unknown parameter. A random sample of size n from $f(x|\theta)$ can be represented by the sequence of i.i.d. random variables $x = (x_1, \ldots, x_n)$. The gain in information provided by the experiment is based on information theory concepts developed by Shannon (1948), and it is given by the Kullback–Leibler divergence between the prior distribution for θ, $\pi(\theta)$, and the posterior given the data, $\pi(\theta|x)$. That is

$$\begin{aligned} K_n &= D_{KL}(\pi(\theta|x)\|\pi(\theta)) \\ &= \int_\Theta \pi(\theta|x) \log \frac{\pi(\theta|x)}{\pi(\theta)}\, d\theta. \end{aligned} \tag{2.5}$$

The expected gain in information K_n^π is given by the expectation of (2.5)

$$K_n^\pi = \mathbb{E}_X \left\{ D_{KL}(\pi(\theta|x)||\pi(\theta)) \right\},$$

where the expectation is taken with respect to the marginal $m(x) = \int f(x|\theta)\pi(\theta)d\theta$. The missing information is the value of K_n^π for large values of n, and the prior that maximises this missing information is the reference prior. That is, the distribution $\pi(\theta)$ maximizing $K_\infty^\pi = \lim_{n\to\infty} K_n^\pi$.

Definition of reference priors

Let us now discuss the derivation of reference priors for the case where θ is a scalar.

We have introduced the definition of expected logarithmic convergence condition in Section 2.1.2, anticipating that reference priors satisfy this property. In particular, Berger et al. (2009) define as *permissible prior* any $\pi(\theta)$ for which this is true. Consider model $M = \{f(x|\theta), \theta \in \Theta\}$. Note that x represents the entire vector of observations, and the model M represents, in the context of reference priors, the probability model for the actual vector of observations. In fact, the theory of reference priors requires the theoretical possibility to replicate the experiment.

Definition 2.3 (Permissible prior). *A strictly increasing probability function $\pi(\theta)$ is a permissible prior for model M, if*

1. $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)\,d\theta < \infty$ *for all* $x \in \mathcal{X}$*; and*

2. *for some increasing sequence* $\{\Theta_j\}_{j=1}^{\infty}$ *of subsets of the parameter space, such that* $\lim_{j\to\infty}\Theta_j = \Theta$*, and* $\int \pi(\theta)\,d\theta < \infty$*,*

$$\lim_{j\to\infty}\int_{\mathcal{X}} f_j(x)\delta\{\pi_j(\theta|x), \pi(\theta|x)\}\,dx = 0,$$

where $\pi_j(\theta)$ *is the renormalised restriction of* $\pi(\theta)$ *to* Θ_j*,* $\pi_j(\theta|x)$ *is the corresponding posterior,* $f_j(x) = \int f(x|\theta)\pi_j(\theta)\,d\theta$ *the corresponding predictive distribution, and* $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$*.* δ *is the intrinsic discrepancy between the distributions:* $\delta\{p,q\} = \min\{D_{KL}(p\|q), D_{KL}(q\|p)\}$*.*

To measure the difference in information between prior and posterior, which is at the basis of reference priors, Berger et al. (2009) suggest Shannon's expected information (Shannon, 1948; Lindley, 1956).

Definition 2.4 (Expected information). *For model M, the information expected from one observation, with prior $\pi(\theta)$, is*

$$\begin{aligned} I(\pi|M) &= \int_{\mathcal{X}}\left\{\int_{\Theta}\pi(\theta|x)\log\frac{\pi(\theta|x)}{\pi(\theta)}\,d\theta\right\}m(x)\,dx \\ &= \int_{\mathcal{X}} D_{KL}(\pi(\theta|x)\|\pi(\theta))m(x)\,dx, \end{aligned}$$

where $m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)\,d\theta$ *is the marginal for observation* x*, and* $\pi(\theta|x) = f(x|\theta)\pi(\theta)/m(x)$*.*

The expected information $I(\pi|M)$ represents what it is gained in observing x from model M, given that the prior on θ is π. If we extend the information to a sequence of k vectors of observations, $x^{(k)} = (x_1, \ldots, x_k)$, intuitively, the gain will be higher, as we would learn more and more from the data as k becomes bigger. Thus, if we indicate by $I(\pi|M^k)$ this information, we would expect that, for $k \to \infty$, the result would be a quantification of the missing information about θ, given the initial one represented by the prior $\pi(\theta)$. So, if we define by $\mathcal{P}$ the set of priors that can represent the initial information we have about θ, the sought distribution will be the one in this set that maximises the missing information.

Two issues raises when the parameter set is continuous. The first one is that the $\lim_{k\to\infty} I(\pi|M^k)$ is not finite, in general; second, on unbounded sets, the expected information is not defined. To solve these problems, Berger et al. (2009) consider the following

Definition 2.5 (Maximising Missing Information property). *Let M be a model with continuous parameter $\theta \in \Theta \in \mathbb{R}$, and let $\mathcal{P}$ be the class of proper prior distributions for θ. The function $\pi(\theta)$ is said to have the Maximising Missing Information (MMI) property for model M, given $\mathcal{P}$, if for any compact set $\Theta_0 \in \Theta$ and any $p \in \mathcal{P}$*

$$\lim_{k\to\infty} \left\{ I(\pi_0|M^k) - I(p_0|M^k) \right\} \geq 0,$$

with π_0 and p_0 the (renormalised) restrictions to Θ_0 of, respectively, π and p.

The definition of the MMI ensures that the missing information exists for any k. This is a consequence of the restriction to a compact set. This "device", as labelled by Berger et al. (2009), allows one to handle the fact that the missing information diverges when k tends to infinity; and this is done by noting that the reference prior will always provide more (missing) information than any other potential prior in the set $\mathcal{P}$.

Now, Berger et al. (2009) are ready to give the formal definition of reference prior for model M.

Definition 2.6 (Reference prior). *A function $\pi(\theta) = \pi(\theta|M, \mathcal{P})$ is a reference prior for the model M, given $\mathcal{P}$, if it is permissible and has the MMI property.*

It is worth to mention that, prior to this formal definition, the justification of working with the maximisation of missing information was somehow heuristic. For example, refer to Bernardo (1979) and Berger and Bernardo (1989).

It is key, for the existence of the reference prior, that both $I(\pi_0|M^k)$ and $I(p_0|M^k)$ are finite, when we consider the (artificial) replications k of the experiment (Berger et al., 2009).

Properties of the reference prior

Reference priors hold three desirable properties (Berger et al., 2009) for an objective prior distribution. These are independence from the sample size, compatibility with sufficient statistics and consistent under reparameterisation. The first property says that, if $x = (x_1, \ldots, x_n)$ is a random sample from model $M = \{f(x|\theta), \theta \in \Theta\}$, with reference prior $\pi(\theta)$, then $\pi(\theta|M^n) = \pi(\theta|M)$ for any fixed n. This property is satisfied for i.i.d. observations. In cases where observations are not i.i.d., such as time series, then the reference prior may depend on n.

Consider model M as above, with sufficient statistic $t = t(x) \in \mathcal{T}$. Let $M_t = \{f(t|\theta), t \in \mathcal{T}, \theta \in \Theta\}$ be the transformed model in terms of t. Then, because expected information is invariant under this type of transformation, we have $\pi(\theta|M) = \pi(\theta|M_t)$, where $\pi(\cdot)$ is the reference prior for θ.

For model M above, consider the one-to-one transformation of θ given by $\phi(\theta)$. Let M_ϕ indicate the model reparametrised according to this transformation. Then, for the invariant property under one-to-one transformation of the expected information, $\pi(\phi|M_\phi)$ will be the reference prior induced from $\pi(\theta|M)$ by the appropriate probability transformation.

Practical computation of reference priors

Definition 2.6 has no practical use. To compute reference priors Berger et al. (2009) give the following theorem. Consider vector $x^{(k)} = (x_1, \ldots, x_k)$ of (artificial) independent replicates of a vector of observation, and let $t_k = t_k(x_1, \ldots, x_k) \in \mathcal{T}$ be any sufficient statistics of $x^{(k)}$.

Theorem 2.1. *Assume a standard model* $M = \{f(x|\theta), \theta \in \Theta \subset \mathbb{R}\}$ *and the standard class* $\mathcal{P}_s$ *of candidate priors. Let* $\pi^*(\theta)$ *be a continuous strictly positive function such that the corresponding formal posterior*

$$\pi^*(\theta|t_k) = \frac{f(t_k|\theta)\pi^*(\theta)}{\int_\Theta f(t_k|\theta)\pi^*(\theta)\, d\theta},$$

is proper and asymptotically consistent, and define, for any interior point θ_0 *of* Θ,

$$\pi_k(\theta) = \exp\left\{\int_{\mathcal{T}_k} \pi(t_k|\theta) \log \pi^*(\theta|t_k)\, dt_k\right\} \quad \textit{and}$$

$$\pi(\theta) = \lim_{k\to\infty} \frac{\pi_k(\theta)}{\pi_k(\theta_0)}.$$

If (i) each $\pi_k(\theta)$ *is continuous and* $\pi_k^0(\theta)/\pi_k^0(\theta_0)$ *is either monotonic in* k *or is bounded above by some* $h(\theta)$ *which is integrable on any compact set, for any fixed* θ *and sufficiently large* k, *and (ii)* $\pi(\theta)$ *is a permissible prior function, then* $\pi(\theta|M, \mathcal{P}_s) = \pi(\theta)$ *is a reference prior for model* M *and prior class* $\mathcal{P}_s$.

Generalisation to the multiparameter case

Reference priors for the case where the model has more than one parameter are computed by simply generalising the case of one parameter. However, Berger et al. (2009) assert that not all definitions and theorems can be extended. In particular, Theorem 2.1 does not have an analogous explicit representation for the multiparameter case. However, it is in principle possible to simply consider a model $M = \{f(x|\theta), \theta \in \Theta\}$, and Bernardo (1979) showed, somehow heuristically, that the distribution maximising the expected missing information,

$$K_n^\pi = \mathbb{E}_X\left\{D_{KL}(\pi(\theta|x)\|\pi(\theta))\right\},$$

which is computed for large n (i.e. $n \to \infty$) is Jeffreys' prior. That is

$$\pi(\theta) \propto \det(I(\theta))^{1/2}. \tag{2.6}$$

This can be seen as the distribution maximising the expected missing information on each compact set Θ_j of Θ, given by $\pi(\theta_j) \propto \det(I(\theta))$, were (weakly) converging to (2.6), as shown in Bernardo (1979).

The result in (2.6) is an important aspect of reference priors, as anticipated. If there are no nuisance parameters, that is if all the parameters are of interest (or, in the case of uniparameter models, the parameter), then Jeffreys' prior is the *reference* prior, in the sense that it is the prior distribution which maximises the expected missing information. The result has been formally shown by Clarke and Barron (1990, 1994). Under regularity conditions, which assure asymptotic posterior normality, by repeated sampling from model $M = \{f(x|\theta), \theta \in \Theta\}$ the above result is attained by the prior

$$\pi(\theta) = I(\theta)^{1/2},$$

where

$$I(\theta) = -\int_{\mathcal{X}} f(x|\theta) \left\{ \frac{\partial^2}{\partial\theta^2} \log[f(x|\theta)] \right\} dx.$$

The result holds for uniparameter and multiparameter models.

Nuisance parameters

Bernardo (1979) proposed to apply reference priors to the case of nuisance parameters. The procedure is, in short, as follows. Consider a model M as above, where the parameter of interest is θ and the nuisance parameter is λ. To deal with this problem, Bernardo (1979, 2005) proposes the following three-steps algorithm.

1. The prior for the parameters (θ, λ) can be written as $\pi(\theta, \lambda) = \pi(\lambda|\theta)\pi(\theta)$. Model $f(x|\theta, \lambda)$, conditional on θ, depends on λ only. Therefore, the one-parameter reference prior $\pi(\lambda|\theta)$ can be found.

2. The nuisance parameter can then be integrated out to derive the marginal model

$$f(x|\theta) = \int_{\Lambda} f(x|\theta, \lambda)\pi(\lambda|\theta)\, d\lambda. \tag{2.7}$$

3. The one-parameter reference procedure can then be applied to (2.7) to obtain $\pi(\theta)$ and, therefore, $\pi(\theta, \lambda)$. The reference posterior for the parameter θ will then be

$$\pi(\theta|x) \propto \int_{\Lambda} f(x|\theta, \lambda)\pi(\theta, \lambda)\, d\lambda = f(x|\theta)\pi(\theta).$$

However, the prior $\pi(\lambda|\theta)$ is improper, so that (2.7) is not a valid statistical model. In this case, the proposed solution is to restrict the integral to a sequence of compact sets. In particular, it is defined the increasing sequence of subsets of Λ, $\{\Lambda_j\}_{j=1}^{\infty}$ converging to Λ. By restricting $\pi(\lambda|\theta)$ on each Λ_j, we obtain the marginals

$$f_j(x|\theta) = \int_{\Lambda_j} f(x|\theta, \lambda)\pi_j(\lambda|\theta)\, d\lambda,$$

from which we can derive the reference posteriors $\pi_j(\theta|x)$. Note that $\pi_j(\lambda|\theta)$ represents the renormalised proper restriction of $\pi(\lambda|\theta)$ to Λ_j. In other words, from the sequence $\{\Lambda_j\}_{j=1}^{\infty}$ we obtain the sequence of posteriors $\{\pi_j(\theta|x)\}_{j=1}^{\infty}$, and the required reference posterior for the parameter of interest is given by $\pi(\theta|x) = \lim_{j\to\infty} \pi_j(\theta|x)$.

The reference prior does not depend from the nuisance parameter, but it may depend on the choice of the parameter of interested. In the former case, for any $\psi = \psi(\theta, \lambda)$ such that (θ, ψ) is a one-to-one function of (θ, λ), we have

$$\pi(\theta, |\psi) = \pi(\theta, \lambda) \left| \frac{\partial(\theta, \lambda)}{\partial(\theta, \psi)} \right|,$$

which is the probability transformation of the reference prior. In the latter case, the reference prior for θ is not the same as the reference prior for $\phi = \phi(\theta, \lambda)$, unless ϕ is a one-to-one transformation of θ, or it does not (asymptotically) depend from it. The reason is that the prior maximising the expected missing information about θ is not (in general) the same that maximises the expected missing information about ϕ (Bernardo, 2005).

In this case as well, under regularity conditions (i.e. asymptotic normality) the prior for the parameter of interest coincide with Jeffreys' prior.

The approach for dealing with nuisance parameters can be extended to models where the number of parameters is greater than two. Consider the d-parameters model $M = \{f(x|\theta), \theta = \{\theta_1, \ldots, \theta_d\}, \theta \in \Theta\}$. If the parameter of interest is, for example, θ_1, and therefore the remaining $d-1$ parameters are nuisance parameters, and under normality hypotheses for the conditional posterior of θ_1 given $\{\theta_2, \ldots, \theta_d\}$ and for the marginal of θ_1, the algorithm for one nuisance parameter described above can be extended to obtain each element of the

$$\pi(\theta) = \pi(\theta_d|\theta_1, \ldots, \theta_{d-1}) \ldots \pi(\theta_2|\theta_1)\pi(\theta_1) \tag{2.8}$$

which corresponds to the reference prior distribution for that particular ordering of the parameters. Intuitively, (2.8) represents the distribution maximising the missing information about parameter θ_1, but also the one which maximises the missing information about θ_2 given θ_1 and so on so forth. Practically, the reference priors are obtained "backwards", that is, before finding $\pi(\theta_1)$, one has to find $\pi(\theta_2|\theta_1)$, and so on. Bernardo (2005) shows that the prior is sensible to the ordering of the parameters. In fact, this ordering should reflect the prior knowledge in terms of *inferential* importance for the parameters, being θ_1 the most important one and θ_d the less important one. The formal procedure to deal with reference priors for multiparameter models can be found in Bernardo and Smith (1994).

It is important to see that in the multivariate case, unlike for the univariate case, the reference prior does not yield Jeffreys' prior, as the following example shows. The reference prior for a location-scale model (Fernández and Steel, 1999a), is $\pi_R(\mu, \sigma) = \sigma^{-1}$, both for ordering (μ, σ) and (σ, μ). Whilst Jeffreys' is $\pi_J(\mu, \sigma) = \sigma^{-2}$, which we know already to be inappropriate as it produces both marginalisation paradoxes (Dawid et al., 1973) and strong inconsistencies (Eaton and Freedman, 2004).

The problem of eliminating nuisance parameters in the Bayesian framework, especially for practical purposes, is important. It is true that, because in the framework we can compute the marginal posterior of the parameter(s) of interest, this may appear as a false problem. However, Liseo (2005) shows that there are practical consequences in eliminating nuisance parameters; several approaches within the Bayesian framework are considered. In particular, for objective Bayes, the integrated likelihood approach (Berger et al., 1999), and the reference prior approach (Liseo, 1993).

2.2.5 Other priors based on maximising missing information

In the literature it is possible to find other approaches in finding objective priors based on the concept of maximising the missing information. On the line of reference priors, the idea is to find a prior distribution for which the expected "difference" in the prior and posterior distributions is maximum. As we have seen, in reference prior the "difference" between the distributions is measured by means of the Kullback–Leibler divergence.

Clarke and Sun (1997, 1999) use the Chi-squared distance to maximise the expected missing information. If we consider density $p(x)$ and density $q(x)$, the Chi-square distance between the two is given by

$$D_{\chi^2}(p(x)\|q(x)) = \int \frac{(q(x) - p(x))^2}{p(x)}\, dx.$$

Therefore, for model $M = \{f(x|\theta), \theta \in \Theta\}$, the prior $\pi(\theta)$ is the one which maximises

$$\chi^2(\pi(\theta)) = \int_{\mathcal{X}} \left\{ \int_{\Theta} \frac{(\pi(\theta|x) - \pi(\theta))^2}{\pi(\theta)}\, d\theta \right\} m(x)\, dx,$$

where $\pi(\theta|x)$ is the posterior and $m(x)$ the marginal of x. The main result by Clarke and Sun (1997, 1999) is that, in the case of the uniparameter exponential family of distributions, where the canonical parameter is the parameter of interest, the prior obtained by maximising the expected Chi-squared distance between prior and posterior is different from Jeffreys' (including the case of nuisance parameters). In particular, it turns out to be the fourth root of the Fisher information.

Ghosh et al. (2011) consider a more general divergence

$$R^\beta(\pi) = \frac{1 - \int \left\{ \int \pi^\beta(\theta)\pi^{1-\beta}(\theta|x)\, d\theta \right\} m(x)\mu(dx)}{\beta(1-\beta)} \qquad \beta < 1, \tag{2.9}$$

where $\mu(dx)$ is a dominating measure. Expression (2.9) represents a family of divergences, indexed by β; when $\beta \to 0$, for example, (2.9) becomes the Kullback–Leibler divergence. Other interesting cases are when $\beta = -1$, for which is the Chi-square distance, and $\beta = -1/2$, for which it represents the Bhattacharrya-Hellinger distance (Hellinger, 1909; Bhattacharyya, 1943). The main result of Ghosh et al. (2011) is that Jeffreys' prior is the prior distribution which maximises the expected missing information in (2.9), with the exception when the Chi-square distance is considered (i.e. $\beta = -1$)

2.2.6 Probability matching priors

A different approach in obtaining objective prior was firstly proposed by Welch and Peers (1963) and Peers (1965). The idea is to obtain prior distributions such that, exactly or as an approximation of a certain order, the posterior probability of the Bayesian credible set coincide with the corresponding frequentist coverage probability. To grasp the idea, we see the following example from Datta and Sweeting (2005).

Example 2.1. *Consider a random variable x normally distributed with unknown mean θ and variance equal to 1. The prior we put on θ is the uniform, which is known to be improper: $\pi(\theta) \propto 1$. Thus, the distribution of x and the posterior for θ are the same, that is $f(x|\theta) = \pi(\theta|x)$. If we consider the posterior distribution of $Z = \theta - x$, we have $P_\pi \left\{\theta \leq \theta_{\alpha(x)|x}\right\} = P_\theta \left\{\theta \leq \theta_\alpha(x)\right\} = \alpha$, with $\theta_\alpha(x) = x + z - \alpha$ and z_α is the α-quantile of the standard normal distribution. We can then see that a credible interval for θ with posterior probability equal to α, is also a confidence interval with confidence level equal to α.*

We then say that the uniform distribution is Probability Matching Prior (PMP).

The main reason PMP method has been developed lies in its frequentist properties. In particular, objective priors can be seen as those prior distributions that "let the data speak" (Kass and Wasserman, 1996), in the sense that the major contribution to the posterior should come from the likelihood (as discussed above). One then may argue that the posterior should lead to inference results that are close to the one coming from classical inference: if the posterior probabilities agree with the sampling ones, we would have obtained the desired result of letting the data "speak". It is in fact for this that probability matching is seen as a property that objective priors could have. An interesting conclusion in this direction can be found in Datta and Sweeting (2005), where they argue that PMP cannot be seen as a general approach in defining objective priors but, rather, as a nice property that priors can have, alongside the invariance property, for example. The main reasons behind this apparently not-favourable argument with respect to PMP, can be sought in the fact that there are many matching criteria and that, in the multiparameter case, there may be infinite possible priors.

The probability matching property can be obtained either exactly or asymptotically. The former is difficult to attain, making the latter of more frequent application. An example on exact PMP can be found in Lindley (1958), Datta et al. (2000a) and Datta and Mukerjee (2004). The authors developed in succession proofs more and more general for a transformation $\tau = g(\theta)$ resulting in a location model with a location parameter. In this case, by assigning a uniform prior on the location parameter, exact matching holds.

We will discuss asymptotic PMP in the reminder of the section.

PMP for one parameter models

The asymptotic matching can be reached in different orders of approximation. There is not a unique terminology on what is classified as first-order, second-order and so on. We use the same approach as in Datta and Sweeting (2005), where an approximation of the coverage probability differs from the credible interval by terms of order n^{-1} is defined as second-order, and one that differs by terms of $n^{-3/2}$ is defined as third-order.

Let us consider the random sample $x = (x_1, \dots, x_n)$ from the density $f(x|\theta)$, with $\theta \in \Theta \in \mathbb{R}$. Under regularity conditions, we choose the α-quantile of the posterior distribution $\theta_{\pi,\alpha}$ such that for the posterior probability we have

$$P\{\theta \leq \theta_{\pi,\alpha}|x\} = \alpha + O(n^{-\gamma}),$$

for some strictly positive γ; and, at the same time, for the coverage probability we have

$$P\{\theta \leq \theta_{\pi,\alpha}\} = \alpha + O(n^{-\gamma}). \tag{2.10}$$

Then we say that some order of matching has been achieved. In particular, if $\gamma = 1$, the second-order probability matching has been achieved. If $\gamma = 3/2$, then the third-order probability matching has been achieved. Welch and Peers (1963) has showed that equation (2.10) holds if and only if $\pi(\theta) \propto \{I(\theta)\}^{1/2}$. Therefore, Jeffreys' prior is second-order probability matching. This particular method for obtaining PMP, is called the *quantile* matching method, as it is based on finding an appropriate quantile of the posterior distribution. And it can be extended to two sided intervals by finding quantiles $\theta_{\pi,\alpha}$ and $\theta'_{\pi,\alpha}$ such that

$$P\left\{\theta_{\pi,\alpha} \leq \theta \leq \theta'_{\pi,\alpha}|x\right\} = P\left\{\theta_{\pi,\alpha} \leq \theta \leq \theta'_{\pi,\alpha}\right\} = \alpha,$$

for some order of precision $O(n^{\gamma})$.

PMP for multiparameter models

Let us first consider the case where a model $f(x|\theta_1, \ldots, \theta_d)$ with $d > 1$, has one parameter of interest and $d-1$ nuisance parameters. It is known (Datta and Sweeting, 2005) that the approximation to normality, both from a Bayesian and a frequentist point of view, holds at the first-order level. As such, similarly to the uniparameter case, for multiparameter models there is always a PMP of order $O(n^{1/2})$. Consider the scalar parameter θ_1 and the vector of nuisance parameters given by $(\theta_2, \ldots, \theta_d)$. Let $\theta_{\pi,\alpha}$ be the α-quantile of the marginal posterior distribution of θ_1 satisfying

$$P\{\theta_1 \leq \theta_{\pi,\alpha}|x\} = \alpha,$$

where x is a random sample from $f(x|\theta_1, \ldots, \theta_d)$. Then, the prior $\pi(\cdot)$ is of second-order probability matching prior with respect to the parameter of interest θ_1 if

$$P\{\theta_1 \leq \theta_{\pi,\alpha}\} = \alpha + O(n^{-1}),$$

for every $\alpha \in (0, 1)$. Whilst it is possible to find cases where the PMP is the same independently of the parameter of interest, in general, the prior changes when the parameter of interest changes.

There are two types of PMP in the multiparameter case: the *simultaneous marginal* PMP, and the *joint* PMP. For the first type, the priors are simultaneously PMP for each parameter of interest and, in general, second order PMP of this kind do not exist (Peers, 1965; Datta, 1996).

The second type of PMP for multiparameter models are the *joint* probability matching priors. These are obtained by matching the joint posterior and frequentist cumulative density functions. These have been discussed by Mukerjee and Ghosh (1997).

Other types of matching priors include, matching priors for highest posterior density regions, moment matching priors and predictive PMP. Highest posterior density (HPD) regions are, either in uniparameter or multiparameter models, d-dimensional intervals with associated the highest volume, for a given credible interval. When these regions have also frequentist validity, in the sense that they match the corresponding confidence region (or interval) we have the HPD matching priors. For models where θ is a scalar parameter, Peers (1968) and Severini (1991) have shown that for location and for scale models, Jeffreys prior is HPD matching.

A particular type of matching priors has been proposed by Ghosh and Liu (2011), and it goes under the name of *moment matching priors*. The basic idea is to define prior distributions such that the posterior mean matches, up to a certain order of approximation, the maximum likelihood estimator (MLE).

A first motivation for these type of priors is that, for obvious reasons, they share the same optimal asymptotic properties held by MLE's. A second motivation is that credible regions for the parameters of interest, can be found only on the basis of posterior mean and variance. And these regions, approximatively, match the confidence intervals based on maximum likelihood.

Some interesting remarks (Ghosh and Liu, 2011). Moment matching priors, conversely to probability matching priors, are not invariant under one-to-one reparameterisations. In the multiparameter case, the approaches are similar (Ghosh and Liu, 2011).

Another way of looking at the matching approach is to consider predictive distributions. We consider a future observation y from the model $f(x|\theta)$, with $\theta \in \Theta$ a real-valued parameter. On the basis of a random sample $x = (x_1, \ldots, x_n)$, the α-quantile $\theta_{\pi,\alpha}$ of the predictive distribution, based on the prior $\pi(\cdot)$, is such that

$$P\{Y > \theta_{\pi,\alpha}|x\} = \alpha.$$

If it is also the case that

$$P\{Y > \theta_{\pi,\alpha}\} = \alpha + O(n^{-\gamma}),$$

then $\pi(\theta)$ is predictive probability matching (Datta et al., 2000b; Sweeting, 2008), with γ typically equal to 2. Similarly to PMP, the matching can be achieved at quantile level, as discussed above, but also in terms of the highest predictive density region (Sweeting, 2008). These type of priors have some interesting properties (Sweeting, 2011), such as avoiding the problems related to improper priors, as the only requirement is that $\pi(\theta) > \infty$. Furthermore, it seems a more appropriate approach when the interest is predicting data yet to be observed. However, unlike PMP, the prior can depend on the value of α.

2.3 Review of objective approaches - Discrete parameter spaces

We now examine objective approaches to define prior probabilities on discrete parameter spaces, which represent the main topic of this work. The literature on the matter tends to be model-specific, in the sense that there are not many approaches designed to be applied to (virtually) any discrete parameter space; rather, specific discrete priors are defined for each particular model. Appropriate model-specific priors will be examined in Chapters 4 and 5. It can be said that objective Bayesian estimation, when the parameter of interest has a discrete space (either finite or infinite), has always been challenging, and none of the methods discussed in Section 2.2 can be directly applied to such type of parameter spaces. Tools such as Fisher information are not defined in discrete scenarios.

If we first consider (for historical reasons) Jeffreys' prior, we know that for a positive unbounded real-valued parameter θ, the prior would have the form $1/\theta$. Therefore, Jeffreys (1961) proposes the prior $\pi(N) \propto 1/N$ for the unrestricted integer parameter $N = 0, 1, 2, \ldots$.

Rissanen (1983) proposes a method to derive objective priors for discrete parameter spaces based on information theory concepts. On signal decoding, to be precise. We briefly discuss this approach later in the section.

For what it concerns reference priors, Bernardo and Smith (1994) show that, in the case of a finite parameter space, the resulting prior is the uniform distribution, as the following proposition explains.

Proposition 2.1. *Let x be the observation from distribution $f(x|\theta)$, where θ is a discrete parameter defined over a finite space: $\theta \in \Theta = \{\theta_1, \ldots, \theta_N\}$. The reference prior for the parameter θ is then the discrete Uniform: $\pi(\theta_j) = c$, for $j = 1, \ldots, N$, with $c > 0$.*

This result is a consequence of the fact that, if the parameter space is finite, than the expected missing information is finite as well, and it is equal to the entropy

$$H[\pi(\theta)] = -\sum_{j=1}^{N} \pi(\theta_j) \log \pi(\theta_j),$$

which is maximised if and only if $\pi(\theta)$ is a discrete Uniform.

This result is not satisfactory (Berger et al., 2012), as it is not always advisable to have a uniform prior for discrete parameters with a structure. For example, suppose we wish to estimate the number of elements in a finite population having a certain characteristic. This problem can be represented by a Hypergeometric probability distribution, which has a specific structure. Therefore, Berger et al. (2012) present a method, based on four embedding approaches, to derive prior distributions for discrete parameter spaces. This approach represents the most general and recent one that allows to derive objective priors distributions for discrete parameter spaces; hence, it will be discussed in detail.

We also discuss the principle behind the approach proposed by Barger and Bunge (2008), based on the *linear difference score* function (Lindsay and Roeder, 1987), which allows to obtain a discrete version of the Fisher information matrix for some specific models.

A universal prior for integers

The prior proposed in Rissanen (1983) applies to the set of natural numbers: $N = \{1, 2, \dots\}$. This prior has the form

$$\pi(N) = \frac{1}{N}\frac{1}{\log_2 N}\cdots\frac{1}{\log_2\cdots\log_2 N}\frac{1}{c}, \qquad N = 1, 2, \dots \tag{2.11}$$

where $c = \sum 2^{-\log_2^* N} \simeq 2.865064$, with $\log_2^* N = \log_2 N + \log_2\log_2 N + \cdots$, which is the sum (finite) of all the terms that are non-negative. The prior in (2.11) derives from estimation problems related to information theory. Here the aim is, given a message $x = (x_1, \dots, x_n)$ generated by some probability model $f(x|\theta)$, to identify the shortest code that allows to describe x and the unknown parameter θ. This is achieved by minimising on θ the following

$$L(x, \theta) = -\log_2 f(x|\theta) + L(\theta), \tag{2.12}$$

where $f(x|\theta)$ is the likelihood, and $L(\theta)$ represents the total number of bits required to encode the parameter. Rissanen (1983) bases his result by optimising the worst case code performance; where the performance of a code is measured by the inverse of the ratio of the entropy and the mean code length

$$\min_L \sup_f \lim_{N\to\infty} \sum_{i=1}^N \Big\{f(i)L(i)\Big\} \Big/ \Big\{-\sum_{i=1}^N f(i)\log_2 f(i)\Big\}.$$

By setting $L(N) = \log_2^*(N) + \log_2 c$, a rewriting (2.12) in terms of powers of two, the expression is minimised for $\pi(N) = 2^{-L(N)}$, which represents (2.11).

The first term in (2.11) corresponds to the objective prior proposed by Jeffreys for discrete numbers; the remaining terms have the function to make $\pi(N)$ proper.

By setting $\pi(0) = 1/2$, and replacing c with $2c$, the prior in (2.11) becomes suitable for any parameter defined on the non-negative integers.

Prior based on the linear difference score function

It is well know that Fisher information is not defined for discrete parameter spaces, and it can only be found for likelihood functions which are differentiable with respect to the parameters. As such, Jeffreys' prior, and consequently, reference priors are not defined. However, Barger and Bunge (2008) have derived objective priors for discrete parameters of some specific model, on the basis of the *linear difference score* (LDS).

Definition 2.7. *Let $f(x|N)$ be a distribution with unknown discrete parameter N and let $L(N)$ be the likelihood function. Then, the difference score function in N is given by*

$$U(N) = \frac{L(N) - L(N-1)}{L(N)} = \frac{\nabla L}{L},$$

where ∇ is the backward difference operator.

The difference score can be seen, in discrete parameter settings, as the equivalent of the score function for continuous parameter settings. If the difference score for N can be expressed as $U(N) = (x - \mu_N)/c_N$, where μ_N and c_N are function of N, then the variance of the difference score is the information in N. And this information is interpretable as the Fisher information in the discrete case. Therefore, recalling the connection between Jeffreys' prior (and reference prior as well) and Fisher information, we have $\pi(N) \propto \{Var(U(N))\}^{1/2}$.

Example 2.2. *Let x be a random variable with a binomial distribution, $x \sim Bin(n,p)$, where the parameter n, representing the number of trials, is unknown, and the parameter p, representing the probability of success at each independent trial, is known. As the likelihood is $L(n|x) = \binom{n}{x}p^x(1-p)^{n-x}$, the LDS is given by*

$$\begin{aligned} U(n) &= \frac{\binom{n}{x}p^x(1-p)^{n-x} - \binom{n-1}{x}p^x(1-p)^{n-1-x}}{\binom{n}{x}p^x(1-p)^{n-x}} \\ &= 1 - \frac{n-x}{n(1-p)} \end{aligned}$$

$$= \frac{x - np}{n(1-p)}.$$

Therefore, the information about n is

$$\begin{aligned} Var(U(n)) &= Var(x)\frac{1}{n^2(1-p)^2} \\ &= \frac{p}{n(1-p)}, \end{aligned}$$

leading to Jeffreys' prior $\pi(n) \propto (1/n)^{1/2}$.

In addition to the binomial case in Example 2.2, Barger and Bunge (2008) derive a prior with the same principle for two Poisson-based models, with applications to estimation of the number of species.

It is worth mentioning that the class of models with the LDS property, that is models for which the LDS can be factorised as $U(N) = (x - \mu_N)/c_N$, is substantial, as shown in Lindsay and Roeder (1987). However, it appears that no research has been performed to generalise the results in Barger and Bunge (2008) to this wide class. Perhaps, it could be interesting to explore this possibility and find, if any, possible general results connected to Jeffreys rule (or reference analysis).

2.3.1 "Reference" priors for discrete parameter spaces

What is probably the most recent, and most comprehensive, approach to define objective priors for discrete parameter spaces, has been introduced by Berger et al. (2012). As we have mentioned, when reference analysis is applied to finite parameter spaces, the result is a uniform prior, and this is not always a desirable result when the problem has certain types of structure.

The general idea in Berger et al. (2012) is to *embed* the discrete problem into a continuous one, such that the structure is preserved, and then apply the standard reference analysis (as seen in Section 2.2.4) to derive the objective prior of interest. The particularity of this method, which we believe representing also one of its major limits, is that it does not exist a unique way to embed the discrete problem into a continuous one; therefore, there is not a "universal" method that can be applied indistinctly to any model with discrete parameters. In addition, when more than one method can be applied for the same model, the priors obtained are in general different, and it is necessary to adopt some comparison procedures to identify the most appropriate.

It has to be noted that, even though we refer to this priors as *reference*, they are not strictly as such, in the sense that they do not arise by the asymptotic maximisation of the missing information of the original (discrete) problem. In fact, Berger et al. (2012) do not call these prior as *reference*; however, for simplicity and consistency with the work here presented, we prefer to label them as "reference" priors.

Let us consider model $f(x|\theta)$, with $\theta \in \Theta$, where the set Θ is discrete. The embedding approaches identified by Berger et al. (2012) are four, and we will discuss some applications in Chapter 4. The approaches are the following.

Approach 1: assuming parameters are continuous The first approach and, possibly, the most simple, is to treat the discrete parameter θ as continuous. It can be applied, for example, in the estimation process for a Hypergeometric model.

There are some limitations to this approach, however. It is quite likely that the "new" probability model will not integrate to 1, and a normalising constant has to be introduced. Therefore, the actual model is going to be $K(\theta)^{-1}f(x|\theta)$, where $K(\theta) = \int f(x|\theta)dx$. As such, it may be possible that the new continuous structure is no longer the same as the discrete one; therefore, Berger et al. (2012) do not recommend this approach when a new normalisation constant is introduced.

A way to overcome this problem, when feasible, is to treat as continuous the data x as well. It may be possible that no additional normalising constant is added, and the approach can be applied. An example is when x is a uniform random variable on the discrete set of integers $\{1, 2, \ldots, \theta\}$. By considering both x and θ as continuous, we obtain the new problem $x \sim U(0, \theta)$, and no additional normalising constant is added. In this case, the reference prior for θ (Bernardo and Smith, 1994) is $\pi(\theta) \propto 1/\theta$.

Approach 2: continuous hierarchical hyperparameter This approach consists in adding a hierarchical level of modelling, with the aim of having a continuous parameter (i.e. hyperparameter), for which standard reference analysis can be applied. In general, we will have the model $f^*(\theta|\theta^*)$, with θ^* continuous, representing the probability distribution of the discrete parameter θ. The problem is solved by finding the objective prior $\pi(\theta) = \int f^*(\theta|\theta^*)\pi^*(\theta^*)\, d\theta^*$, where $\pi^*(\theta^*)$ is the reference prior for the continuous hyperparameter θ^*.

Although this is an appealing approach, it is rare to have scenarios where it can be applied. In addition, even when applicable, it is possible that there is more than one way of adding a hierarchical level, leading in general to different objective priors. It appears then that the objectivity of this approach, even in the limited number of circumstances under which is feasible, may be severely impaired.

Approach 3: consistent estimator To understand this approach, we recall that reference priors are based on considering the asymptotic behaviour, for $k \to \infty$, of a set of k (imaginary) independent replications of the data observed from the model, that is $x^{(k)} = (x_1, \ldots, x_k)$, where each element is a vector of observations in turn (refer to Section 2.2.4). Analogously, this approach first considers a consistent linear estimator $\hat{\theta}_k$ of θ (which is continuous for $k \to \infty$); then finds its asymptotic sampling distribution, and pretends that the parameter θ is continuous in this distribution. Finally, the reference prior is derived with the usual procedure. For example, if $c_k(\hat{\theta}_k - \theta)$ is normally distributed with zero mean and variance $\sigma^2(\theta)$, for some constants c_k, the prior will be given by $\pi(\theta) \propto \sigma(\theta)^{-1}$ (Bernardo, 2005).

There are two important issues with this approach. First, the estimators used can only be inefficient (Berger et al., 2012), leading to conceptual (i.e. philosophical) problems. Second, different estimators may lead to different priors which, as discussed for Approach 2, raises some conceptual concerns about the objectivity of the method. Berger et al. (2012) suggest that this approach, more than resulting in objective priors, simply gives prior distributions which have to be validated by other criteria (e.g. frequentist coverage properties).

Approach 4: parameter-based asymptotics The fourth approach defined by Berger et al. (2012) consists in letting the discrete parameter go to infinity and, in the limiting asymptotic distribution of x, let θ be continuous. Thus, standard reference analysis is applied to obtain a prior for θ. In other words, a formal limiting operation in θ is used to make the parameter of interest continuous.

Recalling the example introduced in Approach 1, we note that x/θ has a uniform distribution on the discrete set $\{0, 1/\theta, \dots, (\theta-1)/\theta, 1\}$. As θ tends to infinity, we can replace the elements of the set by the continuous interval $(0,1)$. Therefore, we can consider the distribution of $x/\theta \sim U(0,1)$, and pretending both x and θ as continuous, we have $\pi(\theta) \propto 1/\theta$.

The limit of this approach is that it defines a prior for large values of θ, but it may not represent a suitable solution for relatively small values of the parameter. As such, similarly to Approach 3, this can be seen more as a method to suggest objective priors which will require validation by other means.

Chapter 3

An Objective Prior Based on Loss Functions

In this chapter we present a new method to derive objective prior probabilities for discrete parameter spaces. In Chapter 1 we have mentioned that our approach aims to obtain the prior mass for a parameter value by objectively measuring a loss. Therefore, in order to have an appropriate understanding of our method, we briefly introduce loss functions and, in the specific, we discuss the *self-information* loss function. We then give the formal definition of our method. The generalisation of our approach to other discrete scenarios, such as model selection and variable selection, will be outlined in Chapter 6 and Chapter 7, respectively.

3.1 Criterion

The essence of an objective approach is (or should be) to provide a result that does not involve subjective input. We understand that the above statement can be somehow too strong, and it is therefore necessary to put it into the appropriate context.

In Bayesian parametric inference, a prior has to be assigned to the parameters of the model. There is now common agreement (Bernardo, 2005) that the objectivity of a Bayesian procedure is considered from the moment the model has been chosen. In other words, whilst it is possible to define a sort of automated process that derives a prior distribution for the parameter of a model, the choice of the statistical model is subjective. We can then conceptually represent an objective Bayesian criterion as a sealed black box, containing principles and procedures, where we input the chosen model and the prior for the parameters is returned.

We believe that the choice of the model necessarily includes its parameterisation. The idea of prior distributions that are invariant under one-to-one transformations, as discussed in Chapter 2, it is at the basis of Jeffreys' prior, for example. A thorough discussion on the invariance property for the most common objective priors is carried out by Datta and Ghosh (1996). The message there is that, although invariance is a nice property to have, it does not constitute a necessity. After all, it is plausible to assume that a choice of a particular model would include the choice of its parameterisation as well. Furthermore, for discrete parameter spaces, the concept of invariance under one-to-one transformation looses meaning, given that assumes differentiability.

An important criticism to objective Bayes, as seen in Section 2.1, relates to the existence of several approaches which may lead to different priors for the same problem. It is in fact legitimate to expect that, if an objective prior distribution on a parameter space (for a given model) exists, this should be unique and independent from the procedure applied to obtain it. For continuous parameter spaces this is true in some specific circumstances: when both reference prior and probability matching prior lead to Jeffreys' prior, for example. But it is not true in general. For discrete parameter spaces the picture is more complicated. From one side, different approaches lead to different priors (Berger et al., 2012), forcing to select the best option on the basis of some criterion (mostly subjective). In addition, there are no general Bayesian procedures that allow to objectively determine a prior distribution for any discrete parameter. The lack of such a procedure represents the main motivation of our work. It is in fact on this ground that we have developed the criterion presented in the thesis.

The criterion proposed, as anticipated, deals with losses instead of probabilities directly. For this reason, our idea allows a different interpretation of the Bayesian framework. Prior probability represents the uncertainty about the true value of the parameter; therefore, if we assume to have no knowledge about the parameter value, it makes perfectly sense that this cannot be encapsulated in a proper prior distribution as, in other words, our uncertainty would be "infinite". This concept has been expressed in Bernardo and Smith (1994), where they claim that is not possible to objectively define a prior representing the absence of knowledge (i.e. ignorance). We can then notice an incongruence in applying Bayes theorem: we begin the process by defining a prior distribution which does not represent probabilities, update it through the likelihood function, i.e. expressing the information contained in the observed data, and we obtain a posterior distribution which is proper. There is then a conceptual deficiency brought by the fact that we start the Bayesian procedure with an entity of a certain nature (the prior) and we end it with another entity of different nature (the posterior). And, as the Bayesian paradigm is based on updating initial beliefs through observation, the fact that the meaning (nature) is not retained throughout the process is a conceptual incongruence. We will see later in Section 3.3 that Bayes theorem can be represented in terms of loss functions, allowing for the retention of meaning throughout the process.

The next section discusses loss functions and their properties, in general, as they represent an important component of our idea.

3.2 Loss functions

Loss functions are used to measure the loss $l(\cdot)$ that one would incur if an event e occurs, and the loss is quantified by $l(e)$. For general considerations on loss functions see, for example, Hirshleifer and Riley (1992). In some cases, loss functions are used to associate a loss to a pair of events, say e and a; in this case, the loss of any combination of the events will be expressed by $l(a, e)$. In particular, if event a is under our control whilst event e is not, the first one is identified as *action*, and the loss function represents the cost deriving from a specific action we take (or decision we make), when the event that is out of our control arises.

Another way of looking at loss functions is by considering the fact that each action we take leads to certain consequences which do not depend on the action only, but also on external circumstances which can only be predicted up to a a certain level of certainty. In this light, the cost of the consequence is measured by a loss function.

Regardless of the way we look at the problem, which is undoubtedly a decision problem, it is licit to assume that a rational behaviour would be the one that aims to minimise the loss. In particular, given that the cost of our actions depends on something uncertain, the aim is to minimise the expected loss, where the weights of this expectation are represented by a probability distribution describing our uncertainty around the random event.

In statistics, for example, loss functions can be applied in estimation or prediction (Berger, 1985). In inference the unknown quantity of interest is the parameter of a model; in prediction, the quantity of interest is the future value of a random variable. To illustrate this, let us assume that we are interest in estimating a parameter $\theta \in \Theta$ of a given family of densities $f(\cdot|\theta)$.

Definition 3.1. *Let $\mathcal{X}$ be the set of all possible outcomes of random variable x. A* decision rule *δ is a function that maps these outcomes (or a subset of $\mathcal{X}$ in the continuous case) into space $\mathcal{A}$, representing all the possible actions that can be taken: $\delta : \mathcal{X} \to \mathcal{A}$.*

In other words, action $a = \delta(x)$ represents the estimate of θ, based on the observations. The loss function can be then re-interpreted as a real-valued function, upper bounded by zero, which measures the cost of estimate a of the true (and unknown) value of the parameter θ. That is, $l(a = \delta(x), \theta)$. $\delta(x)$ represents the estimator of θ, and a possible loss function employed is $l(\delta(x), \theta) = (\delta(x) - \theta)^2$, which expectation with respect to the probability distribution $f(x|\theta)$ represents the *risk function*

$$R(\delta(x), \theta) = \int_{\mathcal{X}} (\delta(x) - \theta)^2 f(x|\theta) d\theta.$$

The above loss function is called the *squared-error* loss function. Other possible loss functions are the *absolute-error* loss function, $l(\delta(x), \theta) = |\delta(x) - \theta|$, and the

0-1 loss function

$$l(\delta(x), \theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_i \\ 1 & \text{if } \theta \in \Theta_j \qquad (j \neq i) \end{cases}$$

where $\Theta_i \cup \Theta_j = \Theta$. More information on loss functions and how to select the most appropriate one can be found, for example in French and Insua (2000) and Berger (1985).

3.2.1 Self-information loss function

An important type of loss function that we consider is the *self-information* loss function. In order to understand it properly, we need to introduce the information theory concept on which it is based upon: *self-information.*

Uncertainty and information are highly related. In fact, we can say that they represent two sides of the same coin. For if there is no uncertainty, then there will be no information as well. To understand this, let us first assume that we have complete information with respect to an event, say e. In this case there would be no uncertainty around it as we would be one hundred percent sure about its realisation. At the other extreme, if we have no information about event e, we immediately see that we are in the maximum possible level of uncertainty. From the point of view of uncertainty, it is clear that the more uncertainty we have about e the more information we will gain if e occurs; similarly, if the level of uncertainty is zero, its realisation will no add any information to the existing one.

Remark 3.1. *Information is a* measure *of the* decrease of uncertainty *from the* receiver *point of view, for if we are highly certain about an event, its occurrence would not significantly decrease our uncertainty. If we are highly uncertain, its occurrence would considerably decrease the level of our uncertainty.*

The measure of the information content associated with the level of uncertainty of a probabilistic event e is called *self-information*, and it is based on the following three axioms.

Axiom 3.1. *When event e has an associated probability of occurrence equal to one, the self-information it carries is zero.*

Axiom 3.2. *Self-information is a decreasing function of the probability associated with event e. In other words, the higher the probability that event e occurs, the lower the level of self-information the event carries, and vice versa.*

From Axioms 3.1 and 3.2, we see that the self-information is a non-negative and unbounded function.

Axiom 3.3. *If events e and e' are independent, the self-information of the joint event representing the simultaneous occurrence of both e and e', is equal to the sum of the self-information associated with each event.*

The logarithmic function simultaneously satisfies the above three Axioms.

Definition 3.2. *Let e be an event with probability of occurrence equal to $P(e)$, for some probability function P. The self information associated with (the occurrence of) e is given by*

$$I(e) = \log(1/P(e)) = -\log P(e).$$

Thus, from Axiom 3.1, we have that if $P(e) = 1$, then $I(e) = 0$. From Axiom 3.2, we have that, if $P(e) < P(e')$, then $I(e) > I(e')$. Finally, from Axiom 3.3, if $P(ee') = P(e)P(e')$ (i.e. the two events are stochastically independent), then $I(ee') = I(e) + I(e')$.

Before formalising the *self-information* loss function, as discussed in detail in Merhav and Feder (1998), we give an intuitive definition of it. To do this, we take into consideration the familiar statistical task of estimating a parameter characterising a probability distribution. This task, as usual, is to make a sensible guess on the parameter θ of $f(x|\theta)$, with $x \sim f(x|\theta)$. Given an observed sample, the inference is made by means of a loss function. Let us first consider the simple case where the sample size is $n = 1$. Thus, the *self-information* loss function for the estimation of θ, on the basis of observation $x_1 \sim f(x|\theta)$, has the form

$$l(\theta, x_1) = -\log f(x_1|\theta). \tag{3.1}$$

If we observe x_1, and θ represents our sensible guess about the true value of the parameter, loss function (3.1) measures the self-information of our choice by

considering the probability associated with the distribution of x_1, when θ is the parameter value. In particular, if what we have observed (i.e. x_1) is very likely to be generated by the distribution $f(x|\theta)$, then the loss associated to the estimate would be relatively low; in fact, the probability $f(x_1|\theta)$ would be relatively large. On the other hand, if it is unlikely that x_1 comes form the distribution $f(x|\theta)$, the loss associated to the sensible guess would be relatively large.
It can be noted that the *self-information* loss function is nothing more than the likelihood function for the observed value x_1. For a sample of size n, that is $x = (x_1, \ldots, x_n)$, the likelihood of having observed this sample is given by $\prod_{i=1}^{n} f(x_i|\theta)$. Therefore, by extending equation (3.1) to the general case of n observations from x, we have

$$\begin{aligned} l(\theta, x) &= -\log\left(\prod_{i=1}^{n} f(x_i|\theta)\right) \\ &= -\sum_{i=1}^{n} \log f(x_i|\theta), \end{aligned} \tag{3.2}$$

which is called *cumulative self-information* loss function, and it can also be derived by Axiom 3.3, considering that $(x_1, \ldots, x_n)$ is a random sample. In other words, the *self-information* loss for a sample of size n, on the basis of the choice (i.e. estimate) θ, is given by the sum of the *self-information* loss in choosing θ for each element of the sample. An interesting aspect is that equation (3.2) can be rearranged as

$$v(\theta, x) = \exp\{l(\theta, x)\} = \prod_{i=1}^{n} f(x_i|\theta), \tag{3.3}$$

where is clear that, in order to minimise the (self-information) loss, we need to maximise the likelihood function, on the right-hand-side of (3.3). Our best guess of θ is nothing more than the Maximum Likelihood Estimate (MLE) of the parameter.

The above example of deriving the *self-information* loss function in a scenario of parameter estimation, can be generalised, and it goes under the subject of *universal prediction* (Merhav and Feder, 1998).

Let us assume that we have observed outcomes $(x_1, \cdots, x_n)$ from a certain phenomenon with support $\mathcal{X}$. The idea is to predict the next outcome, x_n, on the basis of the first $n-1$ observations. That is, we make a decision b_n, and we measure the quality of this decision by means of the loss $l(b_n, x_n)$. As seen, this loss can be measured in different ways. Another way of proceeding is by considering the level of confidence we may have about each possible next outcome x_n. That is, define a probability function on x_n given observations $x^{n-1} = (x_1, \dots, x_{n-1})$. This probability distribution is indicated by the function $b_n(\cdot|x^{n-1})$. Once we have observed x_n, we can evaluate the "goodness" of b_n by considering its value for x_n: $b_n(x_n|x^{n-1})$. The loss function representing this evaluation should give relatively low values for relatively high values of b_n; vice versa, the function would give relatively high values for low probabilities. The *self-information* loss function represents an appealing candidate for this role. Thus, we have that, for every probability distribution $b = \{b(x), x \in \mathcal{X}\}$, for every $x \in \mathcal{X}$, the *self-information* loss function is defined as

$$l(b, x) = -\log b(x),$$

where we consider the logarithm to be the natural logarithm.

There are several reasons why this particular loss function is appealing. As discussed in Merhav and Feder (1998):

- It satisfies the condition that it has to decrease monotonically with the probability assigned to an event;
- As it is based on logarithms, which transform products into sums, this loss function is one of the easiest to work when dealing with joint probabilities;
- In estimation problems, shows that the best guess is the MLE.

3.3 The formal definition of our approach

This thesis proposes a procedure to define objective prior distributions for discrete scenarios. These include objective prior distributions for discrete parameter spaces and objective prior masses on model spaces in model selection and variable selection problems.

Let us assume that a specific probability distribution $f(x|\theta)$ has been chosen to model a certain quantity of interest. This probability distribution has the form of either a probability mass function, if the quantity of interest takes values in a discrete set, or of a probability density function. In this case, the quantity of interest is defined over a subset of the set of real numbers $\mathbb{R}$, or a subset of it. The parameter θ, which can be a vector of parameters, takes values in the discrete space Θ.

The aim is to assign a prior probability $\pi(\theta)$ representing the initial uncertainty around the true value of the parameter.

If $\pi(\theta)$ is determined through objective Bayesian methods, this distribution will often be improper. This fact raises some important concerns about defining objective probabilities *directly*. Contrary to the subjective approach, whereby the prior and the posterior retain the same meaning, the same can not be said of an objective prior. For the posterior derived from it must, at some point, represent beliefs in order to be used. We believe that a solution to this difficulty is not to be objective by assigning a mass to every element of the discrete parameter space Θ, but by assigning a *worth* to every one of them. In other words, to "work" with losses instead of probabilities. Recalling that objectivity arises from the absence of knowledge, actual or alleged, about the true value of the parameter of interest, we can see the justification of the proposed approach, as we can still have an idea on the *worth* that each parameter value represents in the model.

The *worth* of an element of the parameter space can be assessed by describing and evaluating what is lost if this value is removed from the space. And by assigning a mass to each parameter value as a function of its *worth*, we are not subject to the constraint of properness, intrinsic in a probability measure.

By looking at the problem of assigning a prior mass from a differ perspective, we can connect it with the *self-information* loss function discussed in Section 3.2.1. For the prior probabilities represent a probability assignment on the elements of θ: $\pi = \{\pi(\theta), \theta \in \Theta\}$. We can then assign a loss to each element of the parameter space by setting $l(\pi, \theta) = -\log \pi(\theta)$. This loss will be expressed simply as $l(\theta)$. Thus, if a prior π has been assigned, we can then link this to a *worth* of each element by means of this particular loss function. Therefore, we can identify an appropriate objective way to associate a loss to each $\theta \in \Theta$, representing its *worth* in the model, and the prior distribution $\pi(\theta)$ then follows.

Before discussing how the *worth* can be objectively determined, let us examine the impact of our approach on the Bayesian paradigm. We note that, by considering the *worth* as expressed by the *self-information* loss, the Bayesian approach is conceptually consistent, as we update the initial *worth* assigned to θ, through the application of Bayes theorem, to obtain the resulting final *worth* expressed by the *self-information* loss $-\log \pi(\theta|x)$. Indeed, there is an elegant procedure akin to Bayes which works from a loss point of view, namely that

$$-\log \pi(\theta|x) = K - \log f(x|\theta) - \log \pi(\theta), \tag{3.4}$$

where K is a constant which does not depend on θ. Equation (3.4) can be read in the following way: the *initial* information (expressed by the *self-information* loss) contained in the probability statement about θ, is updated on the basis of the information contained in the sample (expressed by the log-likelihood function) in order to obtain the posterior information about the parameter (again, expressed by the *self-information* loss, which in this case is contained in the probability statement of the posterior distribution). In terms of losses, equation (3.4) can be interpreted as a cumulative loss function for assessing the loss of θ in the presence of two pieces of mutual information, x and π. That is, the information coming from the data and the information coming from the prior. We can then rewrite the equation as

$$\text{Loss}(\theta|x, \pi) = K + \text{Loss}(\theta|x) + \text{Loss}(\theta|\pi).$$

To better understand how the *worth* can be objectively measured, we recall that our approach assign a level of "importance" to each element θ of the discrete parameter space Θ by considering what do we lose if we remove from the space that parameter value, and it is the true value. The following theorem (Berk, 1966, 1970) is at the foundations of the quantification of the *worth*. Note that is a simplified version of the actual theorem.

Theorem 3.1. *Consider model $M = \{f(x|\theta), \theta \in \Theta\}$. Let us assume that the* *true value of the parameter is* $\theta_0 \notin \Theta$. *The posterior distribution for the parameter* $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$, *for some prior distribution* $\pi(\theta)$, *asymptotically accumulates on the value* $\theta' \in \Theta$, *such that* $D_{KL}(f(x|\theta_0)\|f(x|\theta'))$ *attains its minimum.*

The result of Theorem 3.1 states that, if a parameter value is removed from the parameter space, and it is the true parameter value, the posterior tends to accumulate on a specific value of Θ: the value such that the distance from the true model, with respect to the Kullback–Leibler divergence, is minimised. In other words, the divergence above represents the utility of keeping θ in the parameter space.

The objectivity of this measure of utility is obvious, as it will depend on the available set of options solely, which is determined by the choice of the model. Once we have selected the model we want to use to represent the quantity of interest, the *worth* of each element of the parameter space would be determined by considering the relative "position" of the possible models.

To better illustrate how an objective criterion to assign a worth to each element of the parameter space can be derived, the following example may be helpful. Let us assume we have a scenario where the possible models are three: f_1, f_2 and f_3, that is $f_1 = f(x|\theta_1)$, $f_2 = f(x|\theta_2)$ and $f_3 = f(x|\theta_3)$, with parameter space $\Theta = \{\theta_1, \theta_2, \theta_3\}$. Let us also assume that f_1 and f_2 are very similar, whilst f_3 is significantly different from the other two. We do not question the rational behind this choice of model options, we just assume that there is one. If we remove from the scenario either f_1 or f_2, as they are relatively close, there is no appreciable change in the whole structure of options, as we still have the remaining model (either f_2 or f_1) to support that specific position. On the other hand, if we remove f_3, the structure of options is considerably different from the original, as only two very similar models are left. We then see that f_3 is more "valuable" than f_1 or f_2, because, if it is removed, the scenario is significantly altered; or, alternatively, we can say that the loss in removing f_3 is higher than the loss in removing either f_1 or f_2. An important aspect is that the loss associated to each model takes into consideration the surrounding models: the more "isolated" θ is, the more its *worth*, the higher its "prior probability".

The formal derivation of the prior distribution for θ on the basis of our idea, can be expressed as follows. The *worth* associates to a particular value of θ is represented by the Kullback–Leibler divergence between the model with θ and the nearest one. That is, $u(\theta) = \min_{\theta' \neq \theta} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta'))$. Therefore, $-\min_{\theta' \neq \theta} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta'))$ has to represent the loss in keeping θ in the parameter space. We link the loss to $\pi(\theta)$ via the *self-information* loss function as follows

$$-\log \pi(\theta) = -\min_{\theta' \neq \theta \in \Theta} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta')), \tag{3.5}$$

for both sides of (3.5) represent the loss measured at point θ of the parameter space. As in fact we have already seen, the *self-information* loss function represents the loss at θ when the probability assignment π has been defined; that is, $l(\theta, \pi) = -\log \pi(\theta)$. Therefore, the prior can be obtained by computing the exponential on both sides of (3.5), with the result

$$\pi(\theta) \propto \exp\left\{ \min_{\theta' \neq \theta \in \Theta} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta')) \right\}. \tag{3.6}$$

The prior distribution in (3.6) represents the core of our approach. It shows some important aspects and properties. First, the objective criterion, on the basis of which we define a prior for a discrete parameter space, consider both the value in the parameter space on which the mass is going to be put on, as well as the surrounding elements. It is in fact the relative proximity to other models that dictates the importance of the value of θ.

It is now necessary to make a fundamental consideration about loss functions in relation to our approach. Loss functions, in general, depend on a constant; in fact, if the objective is to minimise the loss, multiplying by a real constant does not affect the result. In our case, however, we have that the objective approach aims to equate two particular types of loss, as explicited in (3.5): the loss in information represented by the *self-information* loss, and the loss in information in selecting the wrong model, represented by the Kullback–Leibler divergence to the nearest model. Therefore, given that we equate two losses in information (i.e. the same "thing"), there is no need to introduce a scalar constant.

In general, when the *worth* of an element of Θ is zero, then $\pi(\theta) \propto 1$. In other words, if the loss associated to a value of the parameter is zero, it is not *worthy* to keep it in the set, then the prior distribution expresses this by assigning a mass proportional to one. However, sometimes it is desirable that the prior behaves in a more logical way. That is, if the *worth* is zero, then it has to be that $\pi(\theta) = 0$.

To obtain this result, we proceed as follows. The *worth* associated to a particular value of θ is represented by a function $g(\cdot)$ of the minimum Kullback–Leibler divergence, where this divergence represents the utility $u(\theta) \geq 0$ of that particular value of the parameter. To identify the appropriate form of function $g(\cdot)$, we make the following considerations. We map the *worth* of θ to its prior mass by means of the *self-information* loss function, $-\log \pi(\theta) = -g(u(\theta))$, and therefore

$$\pi(\theta) \propto \exp\left\{g(u(\theta))\right\}. \tag{3.7}$$

Given relation (3.7), function g should take value $-\infty$ when the *worth* of θ is zero, and approach $+\infty$ as the *worth* increases. A natural way of defining g, so that it will have the appropriate behaviour, is to put

$$g(u) = \log(e^u - 1). \tag{3.8}$$

While $g(u) = \log u$ would appear more obvious, to map $(0, +\infty)$ to $(-\infty, 0)$, we believe it is more appropriate to remain as close as possible to the original scale - i.e. u, rather than the log-scale (the Kullback–Leibler divergence is already on a log-scale). Hence equation (3.8), which is close to the u scale while mapping 0 to $-\infty$. By setting the functional form of g in (3.7) as it is defined in (3.8), we derive the objective prior for the discrete parameter θ

$$\pi(\theta) \propto \exp\left\{\min_{\theta' \neq \theta \in \Theta} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta'))\right\} - 1, \tag{3.9}$$

which has the sought after property of assigning null mass to element of the parameter space that have no *worth*, in the sense here discussed. The Kullback–Leibler divergence between nearest models tends to be very small. As such, we have that, in general, $\log(e^u - 1) \approx \log u$. Hence, the difference in using $\log(e^u - 1)$ rather than the more direct $\log u$, is going to be minimal.

3.4 Discussion

In this chapter, we present our novel idea to derive objective prior distributions for discrete parameter spaces. We show that, by considering the *worth* of each element of the parameter space, with respect to the surroundings elements, a prior distribution can be obtained by considering losses. We measure the *worth* as the distance from $f(\cdot|\theta)$ to the nearest model, with respect to the Kullback–Leibler divergence. The prior is then derived by linking the *worth*, interpreted as a loss, to $\pi(\theta)$ by means of the *self-information* loss function

$$\pi(\theta) \propto \exp\left\{\min_{\theta \neq \theta' \in \Theta} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta'))\right\}.$$

The prior is objective given that, once the model $f(\cdot|\theta)$ is chosen, the nearest neighbour depends solely on the structure of the model itself.

The prior we obtain is, in general, improper, as the illustrations in Chapter 4 show. The application of Bayes theorem to improper priors is problematic. However, we show in (3.3) that working with losses allows a reinterpretation of the Bayesian procedure, where prior and posterior retain the same meaning. In particular, the beliefs about θ are represented by losses instead of probabilities; and this occur both at the beginning of the procedure, when prior beliefs are defined, and at the end, when posterior beliefs are obtained.

The idea can be extended to other discrete spaces, such as model spaces. Chapter 6 shows how a model prior can be obtained through the same criterion. Chapter 7 shows a further extension to a particular case of model selection: variable selection.

For simplicity and convenience in the exposition, the objective prior obtained by applying our approach will be indicated as the Villa–Walker prior.

Chapter 4

Discrete Parameter Spaces

The content of this chapter constitutes the body of Villa and Walker (2013a).

In this chapter we discuss objective priors for specific discrete parameter spaces. We apply the approach discussed in Chapter 3, in particular in Section 3.3, to some common discrete problems. Along with our proposed prior, for each of the specific problems treated, we discuss alternative results as found in the literature.

The models we discuss are: a population size model, the Hypergeometric model, the multivariate Hypergeometric model, the binomial-beta model, and the binomial model. For all these scenarios the approach we adopt is to have a prior distribution that assigns zero mass when the associated loss is zero, as discussed in Section 3.3.

4.1 A population size model

The first case considered is the estimation of the size of a population by means of a Type II censoring. In this experiment, we have a sample of N units with a lifetime that is modelled by an Exponential distribution with rate parameter λ. Both N and λ are unknown. The experiment is ended when a predetermined number of failures R is reached. The times associated with each failure are indicated by $t_1 \leq \ldots \leq t_R$. For example (Berger et al., 2012), if we were interested in assessing the reliability of a specific software, N would represent the unknown number of bugs in the application, and $t_1, \ldots, t_R$ represent the exponentially distributed length of time of each of the first R reported bugs.

Failure times $t_1 \leq \ldots \leq t_R$, which are assumed to be independent, have the following joint probability density function

$$f(t_1, \ldots, t_r | N, \lambda) = \frac{N!}{(N-R)!} \lambda^R \exp\{\lambda[t_1 + \ldots + (N-R)t_R]\}, \qquad N \geq R. \quad (4.1)$$

As shown in Goudie and Glodie (1981), the variables $V = (t_1 + \ldots + t_R)/t_R$ and $W = t_R$ are minimal sufficient for N and λ. Given that the transformation from $(t_1, \ldots, t_R)$ to $(ct_1, \ldots, ct_R)$, for $c > 0$, induces the transformations (N, λ) to $(N, c\lambda)$ and (V, W) to (V, cW), a maximal invariant statistics is V. Therefore, the joint density for V and W is

$$f(V, W | N, \lambda) = \frac{R}{(R-2)!} \binom{N}{R} \lambda^R W^{R-1} \exp\{-\lambda(V + N - R)W\} \, g_R(V), \quad (4.2)$$

with $1 < V < R$, $W > 0$ and

$$g_R(V) = \sum_{i=1}^{[V]} (-1)^{i-1} \binom{R-1}{i-1} (V - i)^{R-2}.$$

Marginalising, the density for V is

$$f(V|N) = \frac{1}{(R-2)!} \frac{N!}{(N-R)!} \frac{1}{(V+N-R)^R} g_R(V), \qquad 1 < V < R, \qquad (4.3)$$

which depends on N only. In other words, inference about N can be carried out with (4.3). This can be obtained by considering Jeffreys' prior (which is also the reference prior) for λ given N, that is $\pi(\lambda|N) = \lambda^{-1}$ in (4.2) and integrating out λ. As $\int_0^\infty \lambda^R \exp\{-\lambda(V + N - R)W\} \, d\lambda \propto \Gamma(R)[W(V + N - R)]^{-R}$, up to a proportionality constant, from (4.2) we obtain (4.3).

Common objective priors for N would be a constant prior, that is $\pi(N|R) \propto 1$, or the prior $\pi(N|R) \propto 1/N$. Note that the latter, for general discrete parameters, has been suggested by Jeffreys (1961), as discussed in Section 2.3. However, the likelihood in (4.3) tends to one as $N \to \infty$, and neither the constant nor the Jeffreys' priors would be suitable, as the posterior would be improper.

In Berger et al. (2012), the objective prior for parameter N is obtained by applying Approach 1 introduced in Section 2.3.1. The parameter space of N, $\{R, R+1, \ldots\}$, is embedded in the interval $(R-0.5, \infty)$, considering the fact that $f(V|N)$ remains a density function with the same normalisation for each $N \geq R - 0.5$. Thus, the reference prior would coincide with Jeffreys' rule prior, that is $\pi(N|R) \propto \sqrt{I_R(V)}$, where $I_R(N)$ is the Fisher information derived in Lemma 2.1 of Berger et al. (2012)

$$I_R(N) = \sum_{j=0}^{R-1} \left\{ \frac{1}{(N-j)^2} \right\} - \frac{RN!}{(R-2)!\,(N-R)!} J_{R,N},$$

with

$$J_{R,N} = \frac{2}{R^3 - R} \sum_{i=0}^{R-1} (-1)^i \binom{R-1}{i} \frac{1}{(N-R+1+i)^3}.$$

The prior is then computed for parameter $\theta = N - R + 1$, as to have the re-parametrised parameter space $\mathcal{N} = \{1, 2, 3, \ldots\}$. Thus, the prior for N proposed by Berger et al. (2012) is

$$\pi^*(\theta|R) \propto \sqrt{I_R(\theta + R - 1)}, \qquad \theta \in \mathcal{N}. \tag{4.4}$$

For some special cases with $R = 2, 3, 4$, the prior in (4.4) has the form

$$\pi(\theta|R) = \begin{cases} \dfrac{1}{\theta(\theta+1)} & \text{if } R = 2, \\ 1.3036 \dfrac{\sqrt{(\theta+2)\theta + 4/3}}{\theta(\theta+1)(\theta+2)} & \text{if } R = 3, \\ 1.6017 \dfrac{\sqrt{[(\theta+3)\theta + 22/5]\,(\theta+3)\theta + 27/5}}{\theta(\theta+1)(\theta+2)(\theta+3)} & \text{if } R = 4. \end{cases} \tag{4.5}$$

The priors in (4.5) are proper, therefore the normalising constant is included. This has been numerically verified (Berger et al., 2012) up to $R = 100$, by showing that the tail is of order $1/\theta^2$; we recall that this is a necessary condition for having a proper posterior. Furthermore, the prior distributions are all very similar, except for $\theta = 1$. Therefore, Berger et al. (2012) recommendation is to use the prior for $R = 2$ for any value of R, as it is considered a good approximation. Figure 4.1 shows a plot of the priors in (4.5).

4.1.1 The Villa–Walker prior for the population size model

The prior distribution for N obtained on the basis of the objective approach we propose, has the form of (3.9). That is

$$\pi(N|R) \propto \exp\left\{\min_{N' \neq N \in \mathcal{N}} D_{KL}(f(N|R) \| f(N'|R))\right\} - 1, \tag{4.6}$$

which, as seen in Section 3.3, has the property of assigning zero mass to values of N associated with zero loss.

The Kullback–Leibler between two densities of the from (4.3), which differ for the value of N only, is given by

$$\begin{aligned} D_{KL}(f(V|N) \| f(V|N+c)) &= \int_1^R f(V|N) \log\left\{\frac{f(V|N)}{f(V|N+c)}\right\} dV \\ = \int_1^R f(V|N) \log &\left\{\frac{\frac{1}{(R-2)!}\frac{N!}{(N-R)!}\frac{1}{(V+N-R)^R} g_R(V)}{\frac{1}{(R-2)!}\frac{(N+c)!}{(N+c-R)!}\frac{1}{(V+N+c-R)^R} g_R(V)}\right\} dV \\ = \log\left\{\frac{N!}{(N-R)!}\frac{(N+c-R)!}{(N+c)!}\right\} &+ R\,\mathbb{E}\left[\log\left\{\frac{V+N+c-R}{V+N-R}\right\}\right], \end{aligned}$$

where c is an integer, and the expectation is taken with respect to $f(V|N)$. The above expression is an increasing function in c, meaning that the nearest model to $f(V|N)$, in terms of Kullback–Leibler divergence, is either $f(V|N-1)$ or $f(V|N+1)$. Computationally, we have verified that the nearest model to $f(V|N)$ is for $c = +1$, that is $f(V|N+1)$. The computation has been carried out for $R = 2$, $R = 3$, $R = 4$ and $R = 5$, and for values of N up to 100. It has to be noted that for values of $R > 5$ and/or values of $N > 20$, the Kullback–Leibler divergence becomes very small, and can be considered zero. As it seems reasonable, the divergence between contiguous models decreases for $N \to \infty$; in fact, considering the original density in (4.1), we note that, for fixed R and λ, the influence of R with respect to N becomes less prominent for large values of the sample units. As such, contiguous models are more and ore similar to each other.

In the light of this result, the prior for the parameter N, given R, is determined by (4.6) and is given by

$$\pi(N|R) \propto \frac{N+1-R}{N+1} \exp\left\{R\,\mathbb{E}\left[\log\left(\frac{V+N+1-R}{V+N-R}\right)\right]\right\} - 1. \tag{4.7}$$

The prior distribution (4.7) is proper, as the following theorem shows.

Theorem 4.1. *Consider the density $f(V|N,R)$ defined in (4.3), with $N \geq R = \{1, 2, 3, \ldots\}$ and $V \in (1, R)$, and where R is known. The prior for the unknown discrete parameter N, representing the population size of interest, is proper.*

Proof. By applying Jensen's inequality, we have

$$\begin{aligned}\frac{N+1-R}{N+1} \exp\left\{R\,\mathbb{E}\left[\log\left(\frac{V+N+1-R}{V+N-R}\right)\right]\right\} - 1 \quad &\leq \\ \left(1-\frac{R}{N+1}\right)\mathbb{E}\left[\left(1+\frac{1}{V+N-R}\right)^{R}\right] - 1 &\leq \\ \left(1-\frac{R}{N+1}\right)\left(1+\frac{1}{N-R}\right)^{R} - 1&,\end{aligned} \tag{4.8}$$

as V is positive. The last expression on the right-hand-side of (4.8) can be approximated by

$$\begin{aligned}\left(1-\frac{R}{N+1}\right)\left(1+\frac{1}{N-R}\right)^{R} - 1 \quad &\approx \quad \left(1-\frac{R}{N+1}\right)\left(1+\frac{R}{N-R}\right) - 1 \\ &= \frac{R}{N-R} - \frac{R}{N+1} - \frac{R^2}{(N+1)(N-R)} \\ &= \frac{R(N+1) - R(N-R) - R^2}{(N+1)(N-R)} \\ &= \frac{R}{(N+1)(N-R)}.\end{aligned} \tag{4.9}$$

The last term in (4.9) behaves like $1/N^2$, therefore the theorem statement is proved. □

Theorem 4.1 is necessary given that, as pointed out above, the likelihood function converges to a constant for $N \to \infty$.

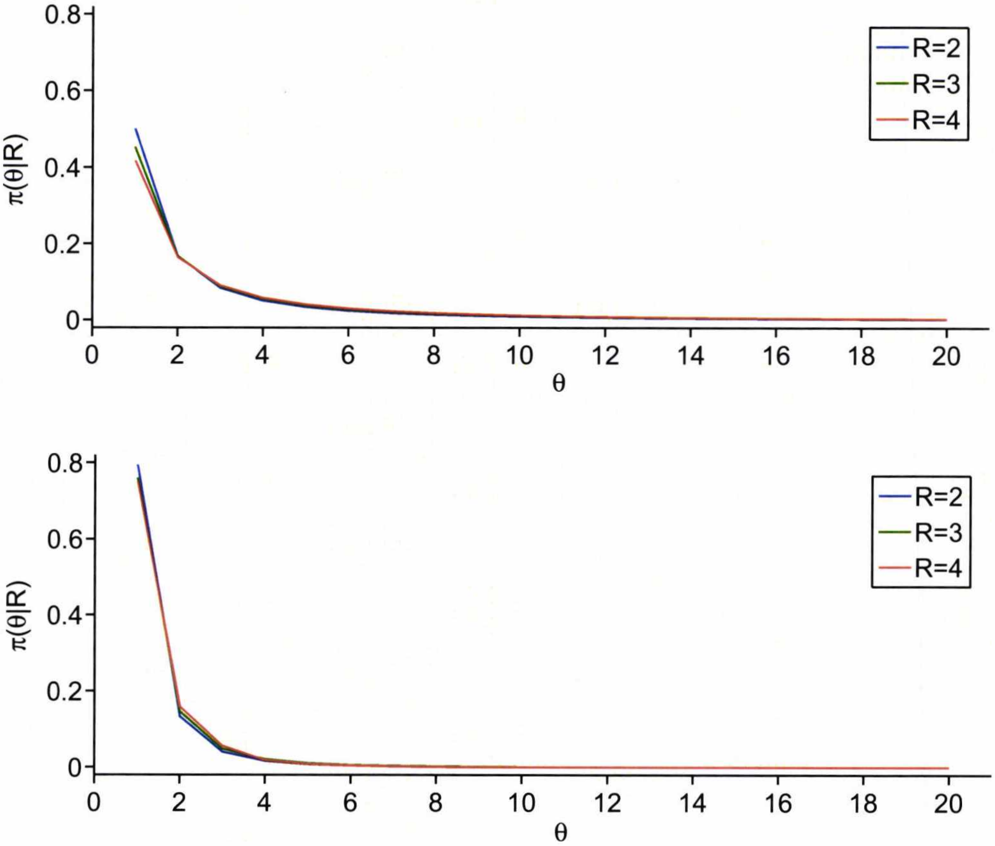

Figure 4.1: Objective prior for the transformed parameter θ of the population size model, given $R = 2$, $R = 3$ and $R = 4$ (top to bottom at $\theta = 1$ in both graphs). The top graph shows the prior in Berger et al. (2012); the bottom graph the Villa–Walker prior.

The prior distribution for N assigns large mass for small values of the parameter space, as expected from the behaviour of the Kullback–Leibler divergence, and rapidly decreases to zero. This is graphically verifiable in Figure 4.1, where the priors for $R = 2$, $R = 3$ and $R = 4$ have been plotted. Note that, in order to be able to compare the obtained prior with the one proposed by Berger et al. (2012), we have transformed the sample space of N in $\theta = N - R + 1$. From Figure 4.1, we can also note that the distributions for different values of R are very similar, and have a behaviour that traces Berger et al. (2012) priors.

4.2 Hypergeometric model

Let us consider now a hypergeometric distribution with probability mass function given by

$$f(r|N, R, n) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}, \qquad r \in \mathcal{R} = \{\max(0, n-(N-R)), \min(n, R)\}, \tag{4.10}$$

with the population size N and the sample size n both known, and $R = 0, 1, \ldots, N$, representing the units in the population which satisfy a certain criterion (or with a certain property). The parameter R is unknown, and the aim is to objectively define the prior $\pi(R|N, n)$.

At first glance, it may seem appropriate to assign a uniform prior to R: $\pi(R|N, n) = 1/(N+1)$. This prior (Briggs and Zaretzki, 2009) assumes that any value of R is as likely to be the true one as any other value. Although a uniform prior appears to be a sort of "natural" choice when the parameter space is discrete, this might not always be the most sensible approach. As pointed out in Berger et al. (2012), a Hypergeometric model shows a well defined structure. In fact, when the population size N grows, the ratio R/N can be seen as the probability of success in a Binomial model. Therefore, the prior for R, or possibly its reparametrisation $p = R/N$, should reflect this structure and resemble the commonly used objective prior $\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$ (Jeffreys, 1961).

It is in fact on the basis of the above considerations, that Berger et al. (2012) obtain the prior distribution for R by applying the embedding Approach 2, as discussed in Section 2.3.1. The idea is to assume that the unknown parameter has a Binomial hierarchical model $Bin(R|N, p)$, where p is an unknown continuous parameter. Therefore, the problem reduces in finding the objective prior for p. As discussed in Bernardo and Smith (1994), the reference prior for a hierarchical model is found by marginalising out the lower level parameters (R in this case), and then applying reference analysis (Bernardo, 2005). Thus, the first step gives

$$\begin{aligned} f(r|n,N,p) &= \sum_{R=0}^{N} f(r|n,R,N) f(R|N,p) \\ &= \binom{n}{r} p^r (1-p)^{n-r}, \end{aligned} \qquad (4.11)$$

where $f(r|n,R,N)$ is the distribution in (4.10) and $f(R|N,p) = \binom{N}{R} p^R (1-p)^{N-R}$. We then note that (4.11) is a Binomial model with parameters n and p. Given that the reference prior for the parameter p of a Binomial distribution, when n is known, is the Jeffreys' rule prior, that is a Beta distribution with both parameters equal to $1/2$, we have

$$\begin{aligned} \pi(R|N) &= \int_0^1 Bin(R|N,p) Be(p|1/2,1/2)\, dp \\ &= \frac{1}{\pi} \frac{\Gamma(R+1/2)\Gamma(N-R+1/2)}{\Gamma(R+1)\Gamma(N-R+1)}, \qquad R = 0,1,\dots,N. \end{aligned} \qquad (4.12)$$

For how it is defined, the distribution in (4.12) is proper. It is worth to mention that this objective prior was initially proposed by Jeffreys (1961).

4.2.1 The Villa–Walker prior for the Hypergeometric model

Let us consider two Hypergeometric models which differ for the value of parameter R only, say $f_R = f(r|N,R,n)$ and $f_{R'} = f(r|N,R',n)$. The Kullback–Leibler divergence between the two models is given by

$$\begin{aligned} D_{KL}(f_R \| f_{R'}) &= \sum_{r\in\mathcal{R}} f(r|N,R,n) \log\left\{ \frac{f(r|N,R,n)}{f(r|N,R',n)} \right\} \\ &= \sum_{r\in\mathcal{R}} \left[f(r|N,R,n) \log\left\{ \frac{\binom{R}{r}\binom{N-R}{n-r} \Big/ \binom{N}{n}}{\binom{R'}{r}\binom{N-R'}{n-r} \Big/ \binom{N}{n}} \right\} \right] \\ &= \log\left\{ \frac{R!\,(N-R)!}{R'!\,(N-R')!} \right\} \end{aligned}$$

$$
\begin{aligned}
&+\sum_{r\in\mathcal{R}}\left[f(r|N,R,n)\log\left\{\frac{(R'-r)!\,(N-R'-n+r)!}{(R-r)!\,(N-R-n+r)!}\right\}\right] \\
&=\log\left\{\frac{R!}{R'!}\right\}+\log\left\{\frac{(N-R)!}{(N-R')!}\right\}+\mathbb{E}\left[\log\left\{\frac{(R'-r)!}{(R-r)!}\right\}\right] \\
&+\mathbb{E}\left[\log\left\{\frac{(N-R'-n+r)!}{(N-R-n+r)!}\right\}\right], \qquad (4.13)
\end{aligned}
$$

where the expectation is taken with respect to the distribution $f(r|N,R,n)$.

For the construction of the objective prior for the parameter R, it is important to keep under consideration some symmetry properties of the Hypergeometric model. First, we note that the random process modelled through an Hypergeometric distribution is symmetrical around $R = N/2$. In fact, by swapping the role of the units which satisfy the criterion, we have

$$f(r|N,R,n) = f(n-r|N,N-R,n),$$

where $r \in \mathcal{R}$ and $n-r \in \{\max(0,n-(N-R)),\min(n,R)\}$. To prove the above result, it is sufficient to rearrange the terms of equation (4.10). In other words, the model with parameter R is equal to the model with parameter $N-R$, for the same values of N and n. Another symmetry property of the Hypergeometric models is obtained when we swap the role of the drawn units with the not-drawn units. In this case we have

$$f(r|N,R,n) = f(R-r|N,R,N-n).$$

A probabilistic proof of this property can be found, for example, in Davidson and Johson (1993).

In order to obtain the prior distribution for R by applying our approach, we need to identify for which value of N' the divergence in (4.13) is minimised. The following Lemma 4.1,which has proof in Appendix A, identifies the appropriate Kullback–Leibler divergence.

Lemma 4.1. *Consider the Hypergeometric distribution f_{R_0}, with parameters R_0, N and n, where N and n are assumed to be known. If we indicate by f_R the*

Hypergeometric distribution that differs from f_{R_0} only by the number of units in the population N which satisfy a certain criterion (i.e. R_0), then the Kullback–Leibler divergence from f_{R_0} to f_R is minimum when $R = R_0 + 1$, if $R_0 < N/2$, and $R = R_0 - 1$ if $R_0 > N/2$. If $R_0 = N/2$, then $D_{KL}(f_{R_0} \| f_{R_0+1}) = D_{KL}(f_{R_0} \| f_{R_0-1})$.

The result in Lemma 4.1 highlights some important aspects of the Hypergeometric distributions. For fixed values of parameters N and n, when we let R vary in its space, models become nearer and nearer when R tends to the middle point (i.e. $N/2$). Furthermore, the behaviour of the Kullback–Leibler divergence, considered in both directions, is symmetrical, property which, as we see below, will result in a prior distribution symmetrical as well.

In deriving the objective prior for the parameter R of the Hypergeometric distribution, we make first the following considerations. We assume that parameters N (population size) and n (sample size) are known. Given the result of Lemma 4.1, for $R \leq N/2$ the minimum Kullback–Leibler divergence is obtained from (4.13) by setting $R' = R + 1$

$$D_{KL}(f_R \| f_{R+1}) = \sum_{r \in \mathcal{R}} \left[\frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \log \right.$$
$$\left. \left\{ \frac{\binom{R}{r}\binom{N-R}{n-r} / \binom{N}{n}}{\binom{R+1}{r}\binom{N-(R+1)}{n-r} / \binom{N}{n}} \right\} \right]$$
$$= \log\left\{ \frac{N-R}{R+1} \right\} + \mathbb{E}\left[\log\left\{ \frac{R+1-r}{N-R-n+r} \right\} \right].$$

Therefore, for $R \leq N/2$, the prior is obtained by applying (4.6), and is given by

$$\pi(R|N,n) \propto \left(\frac{N-R}{R+1} \right) \exp\left\{ \mathbb{E}\left[\log\left(\frac{R+1-r}{N-R-n+r} \right) \right] \right\} - 1. \qquad (4.14)$$

By symmetry, the prior mass for R, when $R > N/2$, is given by

$$\pi(R|N,n) = \pi(N-R|N,n). \qquad (4.15)$$

To illustrate the behaviour of the prior distribution obtained by applying our approach, based on loss functions, we have plotted $\pi(R|N, n)$ (normalised) in

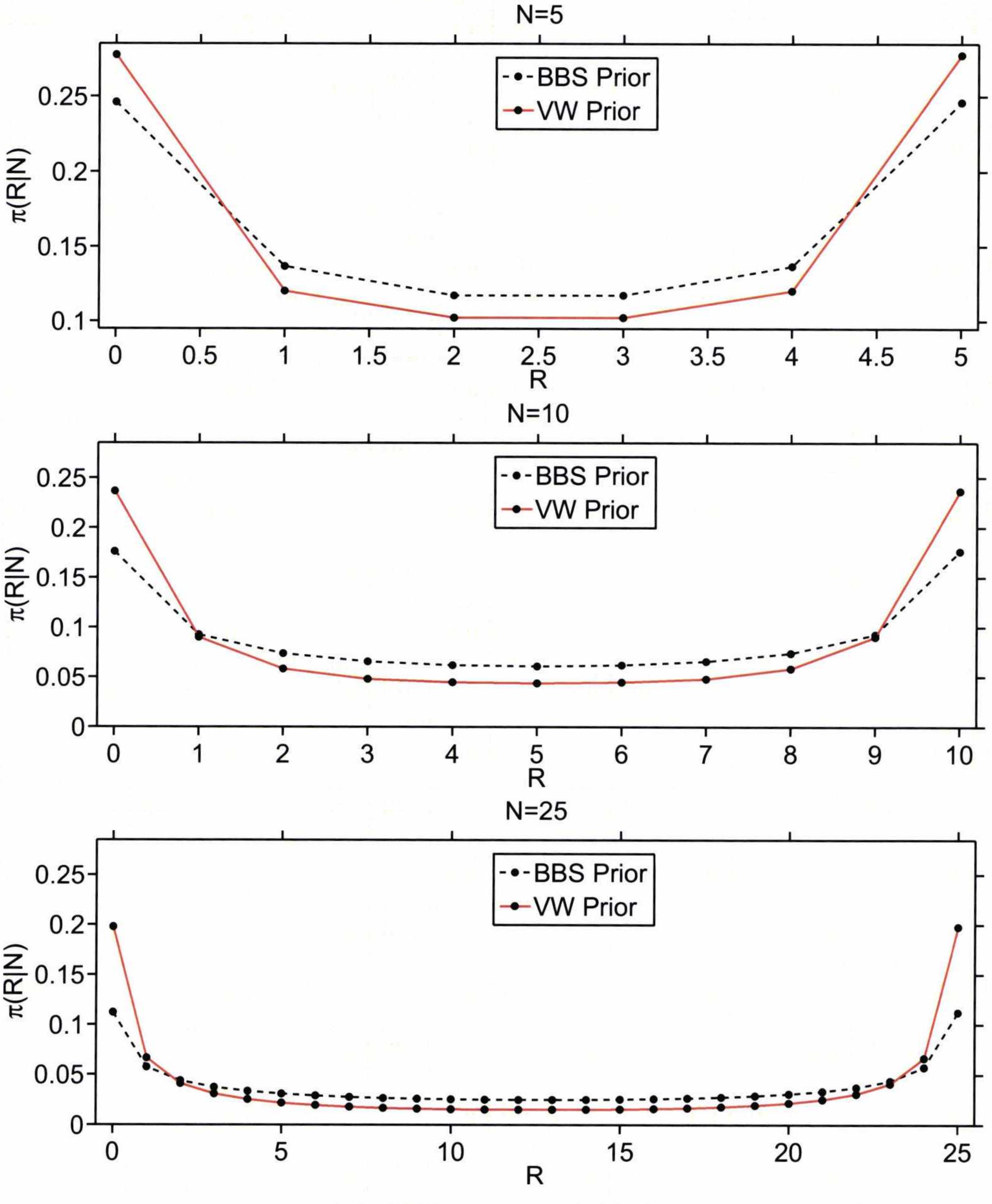

Figure 4.2: Comparison of the objective prior obtained by (Berger et al., 2012) (dashed black line) and the Villa–Walker prior (continuous red line). The prior has been computed for $n = 3$ and for $N = 5$ (top graph), $N = 10$ (middle graph) and $N = 25$ (bottom graph).

Figure 4.2 for three different values of N, given $n = 3$. In particular, $N = 5$ (top graph), $N = 10$ (middle graph) and $N = 25$ (bottom graph). For each prior (continuous red line) we have plotted the respective objective prior computed by Berger et al. (2012) (dashed black line).

By examining the prior, we note that higher mass is assigned to values of the parameter R at the extremes of the parameter space. Then, this mass rapidly decreases for values of $R \to N/2$. This behaviour is common to the three values of N. We have computed the prior for other values of N (not reported here) and noted that the shape of the distribution is similar.

If we compare the Villa–Walker prior with the one defined in Berger et al. (2012), although they are quite similar, it seems that our prior tends to assign more mass to the extreme values (i.e. $R = 0$ and $R = N$) and less mass toward the center of the parameter space.

It can also be noted that, as mentioned in Berger et al. (2012), for N increasing, the prior distribution approximates the reference prior for the parameter p of the Binomial distribution. Where, as seen above, p is approximated by the ratio R/N.

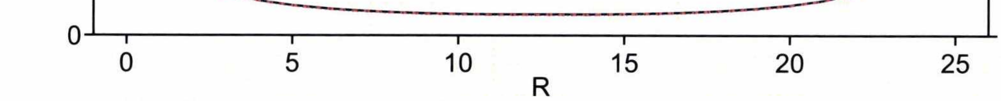

Figure 4.3: Objective prior for the parameter R of the Hypergeometric model, given $N = 25$ and $n = 1$ (continuous black line), $n = 5$ (dashed red line) and $n = 10$ (dotted blue line). For convenience in the comparison, the priors have been plotted as curves only, without highlighting the value in correspondence of each discrete R.

To conclude, as the prior in (4.14) and (4.15) depends on the parameter n, unlike the one specified in Berger et al. (2012), it seems appropriate to analyse the behaviour of the distribution as n changes. We have computed the prior $\pi(R|N,n)$ for a fixed $N = 25$ and three different values of the sample size n, see Figure 4.3. In particular, the graph shows the prior distribution for $n = 1$, $n = 5$ and $n = 10$. We note that the differences in the mass are minimal and they are limited to certain parts of the parameter space. In particular, to values close to $R = 0$ and $R = N$. This is a comforting result, as it means that the choice of the sample size does not have significant impact on the prior distribution.

4.3 Multivariate Hypergeometric model

Consider the multivariate Hypergeometric distribution $MH_d(N, R, n)$ of dimension d, with probability mass function

$$f(r|N,R,n) = \frac{\binom{R_1}{r_1} \cdots \binom{R_d}{r_d}}{\binom{N}{n}} \qquad r \in \mathbb{N}^d, \tag{4.16}$$

where $\mathbb{N}^d$ is the d-dimensional space of non-negative integers, and with $n \in \{0, 1, \ldots, N\}$, $\sum_{j=1}^d R_j = N$, $\sum_{j=1}^d r_j = n$, and $r_j \leq \min(n, R_j)$ for $j = 1, \ldots, d$. For $d = 2$ we obtain the univariate Hypergeometric distribution, discussed in Section 4.2. We assume that parameters N and n are known, and $R = (R_1, \ldots, R_d)$ represents the vector of unknown parameters.

The most commonly used objective prior for this scenario is, in essence, Jeffreys' prior (Jeffreys, 1961). This is derived by first transforming the problem into a continuous one, as it is done in Berger et al. (2012) by applying Approach 2. In this case, the hierarchical model for vector R is a Multinomial distribution with parameters N (i.e. the population size) and p, which is a vector of size d, where each element $p_i = R_i/N$. Thus, the probability mass function for R has the following form

$$Mu_d(R_d|N,p_d) = \frac{N!}{\prod_{j=1}^{d+1} R_j!}\prod_{j=1}^{d+1} p_j^{R_j}.$$

And the objective prior for p_d, by applying Jeffreys' rule, will have the form

$$\pi_J(p_d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\prod_{j=1}^{d+1} p_j^{\frac{1}{2}-1}. \tag{4.17}$$

Even though we do not specifically discuss nuisance parameters in this work, it is important to mention that Berger et al. (2012) have proposed a reference prior for the multivariate Hypergeometric model in the case the vector of parameters could be identified as the union of two subsets: the *parameters of interest* and the *nuisance parameters*. The prior obtained is given by

$$\pi_R(p_d) = \frac{1}{\pi^d}\prod_{j=1}^{d}\frac{1}{\sqrt{p_j(1-\delta_j)}}, \qquad \delta_j = \sum_{i=1}^{j} p_i.$$

4.3.1 The Villa–Walker prior for the multivariate Hypergeometric model

The derivation of the objective prior by applying our approach, is a generalisation of what discussed for the univariate case in Section 4.2. The identification of the minimum Kullback–Leibler divergence, which forms the basis of the approach, becomes rapidly challenging for dimensions of $d > 3$. As such, we will provide a formal procedure for $d = 3$ only; however, due to the symmetry properties of the Hypergeometric distribution, generalisable to any dimension, allow for an intuitive extension of the $d = 3$ result to, virtually, any dimension.

The Kullback–Leibler divergence between the multivariate Hypergeometric distribution with parameters N, R and n, indicated by $f_{N,R,n}$, and the multivariate Hypergeometric distributions with parameters $N, R + a$ and n, indicated by $f_{N,R+a,n}$, where $a \in \mathbb{Z}^d$, is given by

$$D_{KL}(f_{N,R,n}\|f_{N,R+a,n}) \;=\; \sum_r \left[f_{N,R,n}\log\left\{\frac{f_{N,R,n}}{f_{N,R+a,n}}\right\}\right]$$

$$
\begin{aligned}
&= \sum_{r} \left[p_{N,R,n} \log \left\{ \frac{\prod_{j=1}^{d} \binom{R_j}{r_j} \Big/ \binom{N}{n}}{\prod_{j=1}^{d} \binom{R_j + a_j}{r_j} \Big/ \binom{N}{n}} \right\} \right] \\
&= \mathbb{E} \left[\log \left\{ \prod_{j=1}^{d} \left(\frac{R_j!}{(R_j + a_j)!} \frac{(R_j + a_j - r_j)!}{(R_j - r_j)!} \right) \right\} \right] \\
&= \sum_{j=1}^{d} \log \left\{ \frac{R_j!}{(R_j + a_j)!} \right\} + \mathbb{E} \left[\log \left\{ \frac{(R_j - r_j + a_j)!}{(R_j - r_j)!} \right\} \right],
\end{aligned}
\tag{4.18}
$$

where $\mathbb{E}$ is the expectation of $\log(f_{N,R,n}/f_{N,R+a,n})$ with respect to $f_{N,R,n}$. The following Lemma 4.2, which has a proof in Appendix A, determines the minimum Kullback–Leibler divergence from model $p_{N,R,n}$.

Lemma 4.2. *Consider the d-dimensional multivariate Hypergeometric distribution $f_{N,R,n}$, where parameters N and n, with probability mass function as specified in (4.16). If we consider the Hypergeometric distribution $f_{N,R',n}$ which differs from $f_{N,R,n}$ by the composition of the unknown d-dimensional parameter vector R, then the Kullback–Leibler divergence between $f_{N,R,n}$ and $f_{N,R',n}$ is minimum when $R' = R + c$, where c is a vector of dimension d with $d - 1$ zeroes and, in correspondence of the element of R closer to $N/2$, has a minus or plus one depending if the "closeness" is, respectively, from above or below $N/2$.*

To summarise the results obtained in the proof of Lemma 4.2, and to generalise to any multivariate Hypergeometric distribution, we have that the smallest difference between $f_{N,R,n}$ and $f_{N,R+c,n}$ is obtained when only one of the components of R is changed. In particular, when the change is an increase or decrease of one unit. Therefore, c will have $d - 1$ elements equal to zero and the remaining equal to plus or minus one. From the analysis of the Kullback–Leibler divergence between two bivariate Hypergeometric models the nearest model to $f_{N,R,n}$ corresponds to the model $f_{N,R+c,n}$, where c will have $d - 1$ null elements and the remaining one, in position i (where i is the index of element R_i of R nearest to $N/2$, either from below or above), and will have value one if $R_i \leq N/2$, and minus one if $R_i \geq N/2$.

The objective prior for the R can be found from the result of the previous paragraph, and it has the form

$$\pi(R) \propto \exp\{D_{KL}(f_{N,R,n} \| f_{N,R+a,n})\} - 1,$$

where vector a is determined so that the divergence is minimised, as discussed in Lemma 4.2.

To have a better understanding of the prior we propose, we have computed it for the specific case of a bivariate Hypergeometric distribution. In this way it is possible to have a useful graphical aid to capture the main characteristics of the prior.

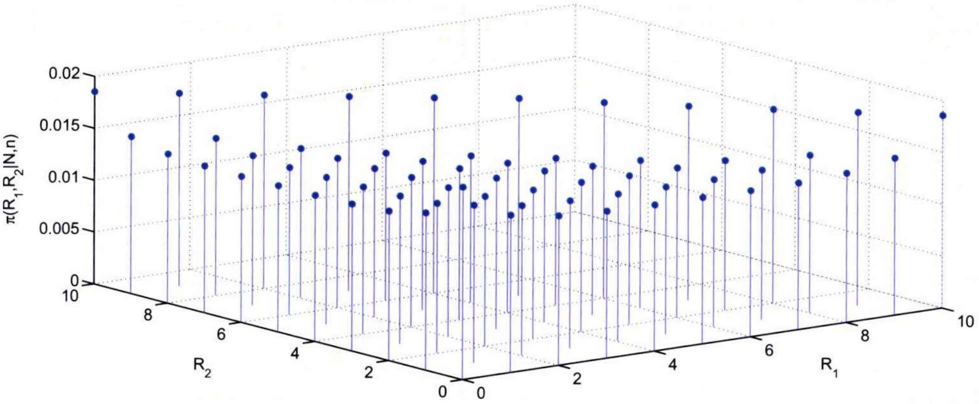

Figure 4.4: Graphical representation of the normalised prior, obtained following our approach, for the bivariate Hypergeometric model with parameters $N = 10$, $n = 3$ and unknown $R = (R_1, R_2)$.

In particular, we have considered a bivariate distribution with population $N = 10$ and sample size $n = 3$. In Figure 4.4 we have plotted the normalised prior distribution $\pi(R)$, with $R = (R_1, R_2)$. We note, from the figure, that the symmetry properties are reflected in the distribution of the prior mass. It can be seen that the largest mass is put at the edges of the parameter spaces; that is, at the points $(0, 0)$, $(0, 10)$ and $(10, 0)$. The mass decreases toward the "centre" of the distribution, that is when either or both R_1 and R_2 approach $N/2 = 5$.

Jeffreys' objective prior for $R = (R_1, R_2)$, as reported in Berger et al. (2012), is computed on continuous parameters. Therefore, it is not possible to perform a complete comparison with the Villa–Walker prior. In particular, due to its nature, the prior in (4.17) would give infinite values at the borders of the parameter space. However, through graphical representation, it is still possible to capture some similarities to the Villa–Walker in the behaviour. In Figure 4.5 we have plotted the surface (left) of the prior density function for parameter vector R, and the contour lines (right).

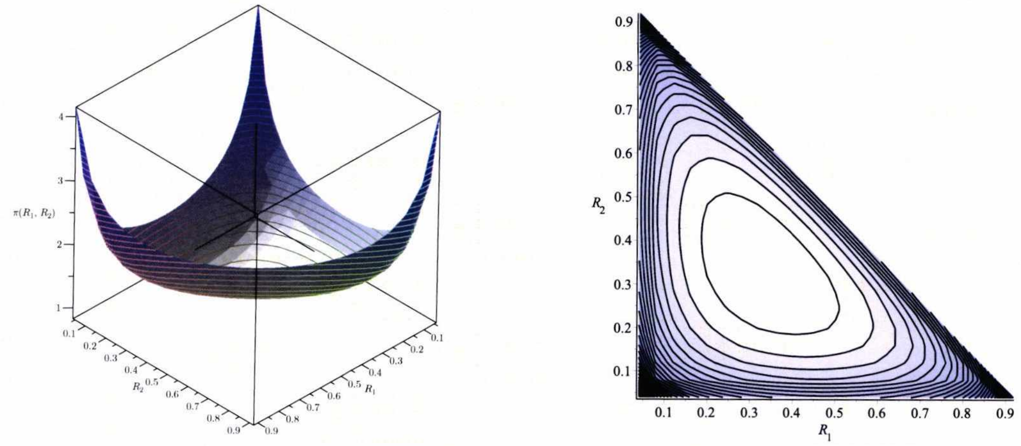

Figure 4.5: Graphical representation of the Jeffreys' prior for the parameter $R = (R_1, R_2)$ of a bivariate Hypergeometric model with $N = 25$. The plot on the left represents the surface of the distribution with highlighted the contours. The contours are shown in the right plot, where the lighter colour corresponds to low density regions, and the dark colour to high density regions.

By inspecting Figure 4.5, we note the symmetry of the prior, which is remarkably similar (at least in terms of behaviour) to the one we obtained considering our approach. In fact, the highest density is in correspondence to the three vertices of the parameter space. And, as expected, the central area of the triangular parameter space has associated relatively low density.

4.4 Binomial-Beta model

Let us now consider the Binomial-Beta model. This particular model arises as the marginal distribution of a Binomial model with parameters $n = 1, 2, \ldots$ and

$p \in (0,1)$, where p is in turn modelled through a Beta density. That is, $x|n,p \sim Bin(n,p)$ and $p \sim Be(a,b)$. Thus, the Binomial-Beta distribution is given by

$$\begin{aligned} f(x|n) &= \int_0^1 \frac{n!}{(n-x)!\,x!}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{x+a-1}(1-p)^{n-x+b-1}dp \\ &= \frac{n!}{(n-x)!\,x!}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 p^{x+a-1}(1-p)^{n-x+b-1}dp \\ &= \frac{n!}{(n-x)!\,x!}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(n+a+b)}, \qquad (4.19) \end{aligned}$$

for $x = 0, 1, \ldots, n$. The parameter of interest is the number of trials n.

The uniform prior on n is not an appropriate solution (Berger et al., 2012). In fact, we note that the tail of the marginal likelihood (for fixed x) has the form

$$\begin{aligned} f(x|n) &\propto \frac{n!}{(n-x)!}\frac{\Gamma(n-x+b)}{\Gamma(n+a+b)} \\ &\approx \frac{1}{n^{a+b-1}(n-x)^{(}1-b)} \approx \frac{1}{n^a}, \end{aligned}$$

that is, $f(x|n)$ can be approximated by $1/n^a$, which is not integrable for $a \leq 1$. Therefore, if the prior of n is uniform, the posterior would be improper.

The objective prior proposed by Berger et al. (2012) is obtained by applying the Approach 3 discussed in Section 2.3.1. Recalling that this approach aims to apply reference prior theory with a consistent estimator, the following non efficient estimator has been identified

$$\hat{n} = \frac{a+b}{ak}\sum_{j=1}^{k} x_j,$$

where k is the number of independent samples from (4.19). As the mean and the variance of $\hat{n}$ are given by, respectively, $\mathbb{E}[\hat{n}|n] = n$ and $Var[\hat{n}|n] = n(n+a+b)a^{-2}(a+b+1)^{-1}k^{-1}$, by applying the central limit theorem we have

$$f(\hat{n}|n) \approx N\left(\hat{n}\,\middle|\, n, \frac{n(n+a+b)}{a^2(a+b+1)k}\right).$$

Pretending that n is continuous, and applying reference analysis, we obtain

$$\begin{aligned}\pi_1(n) &\propto \left(\frac{n(n+a+b)}{a^2(a+b+1)k}\right)^{-1/2}\\ &\propto \frac{1}{\sqrt{n(n+a+b)}},\end{aligned} \tag{4.20}$$

which coincides with the prior obtained by applying Jeffreys' rule.

Although the prior distribution in (4.20) is the one chosen by Berger et al. (2012), it is worth to mention that the decision is the result of the comparison with the prior obtained by applying Approach 4 in Section 2.3.1. With this approach, the prior obtained was $\pi_2(n) \propto 1/n$. In both circumstances the prior distribution for n is suitable, in the sense that the posterior is proper. However, Berger et al. (2012) compare π_1 and π_2 on the basis of the respective frequentist coverage of credible sets of the posteriors. In particular, for values $a = b = 5$, $a = b = 20$ and $a = b = 50$, the approximate frequentist coverage and the average posterior coverage are compared, resulting in a better performance of π_1 over π_2. It can then be concluded that the fact that $\pi_1(n)$ depends on the the value of the parameters of the Beta, a and b, gives a superior prior distribution.

4.4.1 The Villa–Walker prior for the Binomial-Beta model

We now illustrate the prior distribution for parameter n of the Binomial-Beta model in (4.19) obtained by applying our objective criterion. In order to perform a sensible comparison with the result presented by Berger et al. (2012), we apply the same conditions to the parameters of the Beta distribution. That is, $a, b > 1$.

To find the model which is nearest, in terms of Kullback–Leibler divergence, to $f_n = f(x|n)$, we first note that, as $P(x = n|n') = 0$ for $n > n'$, we have

$$D_{KL}(f_n \| f_{n'}) = \infty \qquad n > n'.$$

Therefore, the nearest model will have the parameter representing the number of trials larger than n. The Kullback–Leibler divergence between model f_n and f_{n+j}, $j = 1, 2, \ldots$, is given by

$$\begin{aligned}
D_{KL}(f_n \| f_{n+j}) &= \sum_{x=0}^{n} f_n \log\left(\frac{f_n}{f_{n+j}}\right) \\
&= \sum_{x=0}^{n} f_n \log\left(\frac{n!}{(n-x)!}\frac{\Gamma(n-x+b)}{\Gamma(n+a+b)}\right) \\
&\quad -\sum_{x=0}^{n} f_n \log\left(\frac{(n+j)!}{(n+j-x)!}\frac{\Gamma(n+j-x+b)}{\Gamma(n+j+a+b)}\right) \\
&= \log\left(\frac{n!}{(n+j)!}\frac{\Gamma(n+j+a+b)}{\Gamma(n+a+b)}\right) \\
&\quad +\mathbb{E}\left[\log\left(\frac{(n+j-x)!}{(n-x)!}\frac{\Gamma(n-x+b)}{\Gamma(n+j-x+b)}\right)\right]. \qquad (4.21)
\end{aligned}$$

The Kullback–Leibler divergence in (4.21) is minimum when $n' = n+1$. This is a sensible result given that model $f(x|n)$ and model $f(x|n+1)$ are more similar to each other than $f(x|n)$ and $f(x|n+2)$. This has been computationally verified by calculating the difference $D_{KL}(f_n \| f_{n+2}) - DKL(f_n \| f_{n+1})$, for different values of the parameters a and b, and noting that it is positive.

In detail. The divergence between f_n and f_{n+1}, by setting $j = 1$ in (4.21) is

$$D_{KL}(f_n \| f_{n+1}) = \log\left(\frac{n+a+b}{n+1}\right) + \mathbb{E}\left[\log\left(\frac{n+1-x}{n-x+b}\right)\right], \qquad (4.22)$$

and the divergence between f_n and f_{n+2} is

$$\begin{aligned}
D_{KL}(f_n \| f_{n+2}) &= \log\left(\frac{(n+a+b)(n+a+b+1)}{(n+1)(n+2)}\right) \\
&\quad +\mathbb{E}\left[\log\left(\frac{(n+1-x)(n+2-x)}{(n-x+b)(n-x+b+1)}\right)\right]. \qquad (4.23)
\end{aligned}$$

In Figure 4.6 we have plotted the difference between the two divergences, that is (4.22) and (4.23), for $n = 1, \ldots, 70$, and for $a = b = 5$ (continuous-red curve), $a = b = 20$ (dashed-blue curve) and $a = b = 50$ (dotted-black curve). It is easy to note that the value of $D_{KL}(f_n \| f_{n+2}) - DKL(f_n \| f_{n+1})$, is always positive.

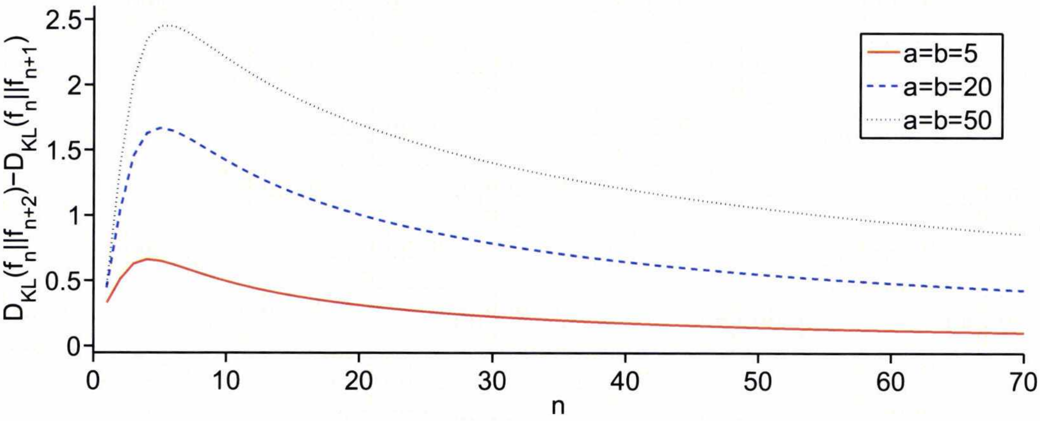

Figure 4.6: Value of the difference between the Kullback–Leibler divergence $D_{KL}(f_n \| f_{n+2})$ and the Kullback–Leibler divergence $D_{KL}(f_n \| f_{n+1})$. The difference has been computed for different values of the parameters of the Beta distribution. That is, $a = b = 5$ (continuous red line), $a = b = 20$ (dashed blue line) and $a = b = 50$ (dotted black line).

By applying our approach, the objective prior for n based on (4.22), is given by

$$\begin{aligned} \pi(n) &\propto \exp\left\{D_{KL}(f_n \| f_{n+1})\right\} - 1 \\ &= \frac{n+a+b}{n+1} \exp\left\{\mathbb{E}\left[\log\left(\frac{n+1-x}{n-x+b}\right)\right]\right\} - 1, \end{aligned} \tag{4.24}$$

or, equivalently

$$\pi(n) \propto \left[\frac{n+a+b}{n+1} \prod_{x=0}^{n} \left(\frac{n+1-x}{n-x+b}\right)^{p_n}\right] - 1. \tag{4.25}$$

The prior in (4.25) is improper. The following Theorem 4.2 shows that, with only one observation, the posterior distribution is proper.

Theorem 4.2. *Let us assume that we observe the data point x_1 from a Binomial-Beta distribution with parameters $a > 1$, b, n and p. Also, assume a prior distribution for the parameter n as $\pi(n) \propto \exp\{D_{KL}(p_n \| p_{n+1})\} - 1$. Then, the posterior* *distribution given by*

$$\pi(n|x_1) \propto \left[\frac{n+a+b}{n+1}\prod_{x=0}^{n}\left\{\left(\frac{n+1-x}{n-x+b}\right)^{p_n}\right\}-1\right] \times \binom{n}{x_1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x_1+a)\Gamma(n-x_1+b)}{\Gamma(n+a+b)},$$

is proper.

Proof. As discussed in Berger et al. (2012), and mentioned at the beginning of this section, the tail of the marginal likelihood of n behaves like $1/n^a$ which, for $a>1$, converges. Thus, we have

$$\begin{aligned}\sum_{n=1}^{\infty}\pi(n|x_1) &\approx \sum_{n=1}^{\infty}\left\{\left[\frac{n+a+b}{n+1}\prod_{x=0}^{n}\left(\frac{n+1-x}{n-x+b}\right)^{p_n}-1\right]\frac{1}{n^a}\right\}\\ &= \sum_{n=1}^{\infty}\left\{\frac{1}{n^a}\frac{n+a+b}{n+1}\prod_{x=0}^{n}\left(\frac{n+1-x}{n-x+b}\right)^{p_n}\right\}-\sum_{n=1}^{\infty}\frac{1}{n^a}. \qquad (4.26)\end{aligned}$$

As $(n+1-x)/(n-x+b)<1$, because of the condition $b>1$, the product term is small or equal to one. Therefore

$$\sum_{n=1}^{\infty}\left\{\frac{1}{n^a}\frac{n+a+b}{n+1}\prod_{x=0}^{n}\left(\frac{n+1-x}{n-x+b}\right)^{p_n}\right\}\le\sum_{n=1}^{\infty}\left\{\frac{1}{n^a}\frac{n+a+b}{n+1}\right\}<\infty,$$

as $a>1$. And, as $\sum_{n=1}^{\infty}\{1/n^a\}<\infty$ in (4.26), the theorem statements follows. □

The following Theorem 4.3, which proof in in Appendix A, shows that the posterior distribution for n is consistent.

Theorem 4.3. *Consider the family of Binomial-Beta distributions f_n, with $n=1,2,\dots$ and common parameters a and b. We also assume that the true value of n is n_0. Given the prior distribution $\pi(n)$ and the set of observations from f_{n_0}, $x=(x_1,\dots,x_k)$, the mass of the posterior corresponding to n_0 converges to one almost surely. That is,*

$$\pi(n_0|n\ge n_0,x_1,\dots,x_k)\to 1,$$

for $k \to \infty$.

To have a feeling of the behaviour of the prior in (4.24), we have computed its mass for different values of the parameters a and b. In particular, as shown in the bottom plot of Figure 4.7, we have computed the prior for $a = b = 5$ (continuous-red curve), $a = b = 20$ (dashed-blue curve) and $a = b = 50$ (dotted-black curve). The choice of the parameter values has been done in order to compare our result with the one obtained by Berger et al. (2012), where objective priors for the same values of a and b have been computed. These priors are showed in the top plot of Figure 4.7.

By inspecting the bottom plot of Figure 4.7, we note that the prior obtained with our approach puts more mass on the lower values of the parameter n. The value of the mass, then, rapidly decreases toward zero as n increases. We can interpret this behaviour as the fact that Binomial-Beta models with a small value of n have more *worth* than models with a large value of n. Also, when n is sufficiently large, models tend to have similar importance. Furthermore, the three curves a very similar, suggesting that the parameters of the Beta distribution for p, that is a and b, play a marginal role in the determination of the prior distribution.

As expected, on the basis of the discussion in Berger et al. (2012), which we have repeated at the beginning of this section, the prior distribution based on (4.20) is sensible to the value of the parameters a and b. In particular, from the top plot in Figure 4.7, we note that the main difference occurs for values of n relatively low, where values of a and b relatively large, produce relatively small prior mass. However, as n grows larger, the three priors tend to assign similar mass to the same value of n.

4.5 Binomial model

The last model we discuss in this chapter in the Binomial distribution. Let us assume that the random variable x is binomially distributed with number of trials n and probability of success p. Its probability mass function is

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \dots, n.$$

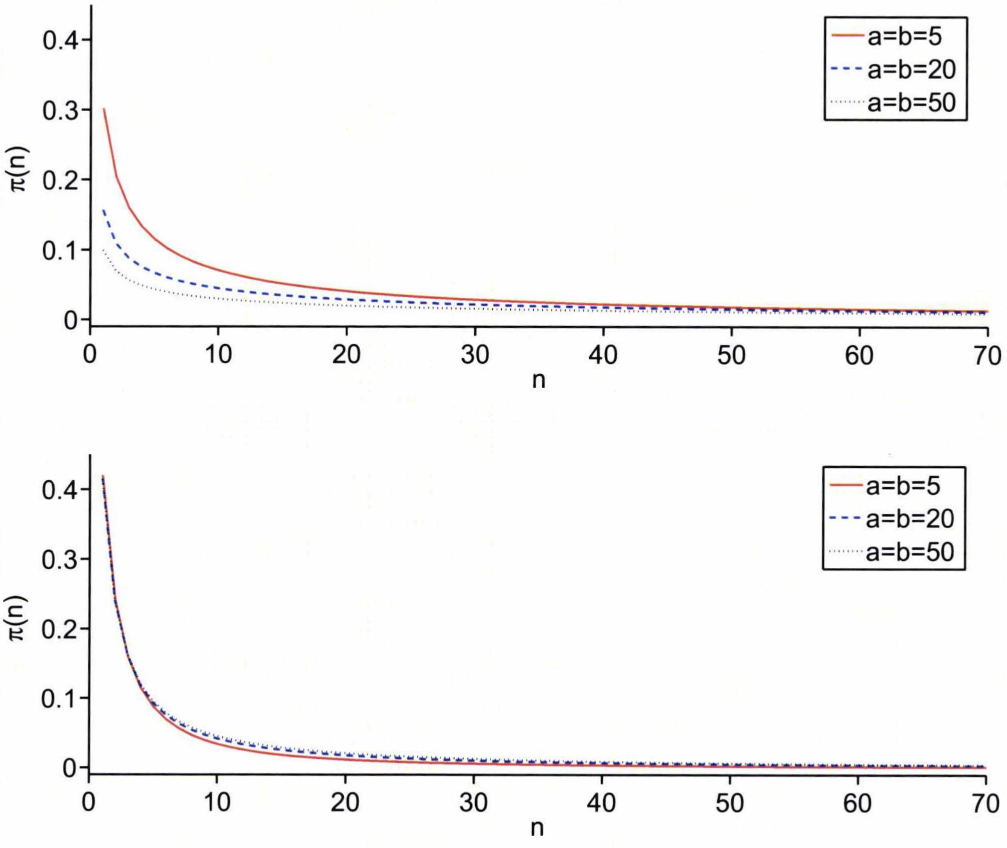

Figure 4.7: Objective prior obtained by applying our method (bottom) and objective prior obtained by Berger et al. (2012) (top). For both approaches, the prior has been computed for $a = b = 5$ (continuous red line), $a = b = 20$ (dashed blue line) and $a = b = 50$ (dotted black line).

We assume p known, which means that we will initially discuss the prior for n, given p. The goal is to define a function which assigns positive mass to all the possible values of n, where $n = 1, 2, \ldots$. In the following discussion, we indicate the prior for n as $\pi(n)$, whether it depends functionally on p or not. In other words, some of the following priors are for n given p, but we consider this as implicit in the notation $\pi(n)$.

The natural objective prior mass function for n would be the uniform, that is

$$\pi(n) \propto 1, \tag{4.27}$$

as this gives equal weight to each possible value of n, and this would hold whether p is known or unknown. However, as n is theoretically infinite, the posterior obtained from (4.27) would be improper when p is unknown. This issue has been discussed, for example, in Berger et al. (1999) and Berger et al. (2012).

In Draper and Guttman (1971), we find the following objective prior mass function on the parameter n, considering p as known

$$\pi(n) = 1/N, \tag{4.28}$$

where N is a large preselected integer with $1 \leq n \leq N$. For example, if n were the number of units with a certain characteristic within a population, N could represent the size of the population. The authors show that the estimate of the parameter n obtained by using the prior (4.28), is represented by the mode of the posterior distribution $\pi(n|x)$, where x represents the observations. Furthermore, without defining in explicit form the prior mass function for n, it is shown that the estimate is the same when p is considered unknown; this is done by assuming the parameters to be independent and defining a beta prior mass function for p.

Raftery (1988) adopts a hierarchical Bayesian approach. In particular, the issue of n being discrete is overcome by first assuming that the parameter has a Poisson distribution, and then by assigning an objective prior to the continuous hyper-parameter. The result is a prior mass function of the form

$$\pi(n) \propto 1/n. \tag{4.29}$$

Note that prior (4.29) coincides with the generic Jeffreys' prior for unbounded integer parameters, as introduced in Section 2.3.

Raftery's prior is improper, as it diverges for n going to infinity. Even though this does not represent a problem in a Bayesian approach, as long as the posterior is proper, Rissanen (1983) (by applying his general method to obtain objective priors for integer parameter spaces discussed in Section 2.3) proposes a modified version of it by reducing the dominating factor $1/n$ just enough to make the whole prior proper.

The objective prior for n we examine in details, is the one in Berger et al.

(2012). The prior is derived by applying Approach 3 in Section 2.3.1. In particular, the following linear estimate of n is considered

$$\hat{n} = \frac{1}{p\,k}\sum_{j=1}^{k} x_j,$$

where k is the number of independent samples from a Binomial distribution with parameters p and n; that is, $(x_1, \ldots, x_k)$. As

$$\mathbb{E}[\hat{n}|n,p] = n \quad \text{and} \quad Var(\hat{n}|n) = [n(1-p)]/[k\,p],$$

for the central limit theorem, we have

$$p(\hat{n}|n,p) \approx N\left(\hat{n}\middle| n, \sqrt{\frac{n(1-p)}{k\,p}}\right).$$

By considering n as continuous, and applying Jeffreys' rule, the prior for n will be

$$\pi_1(n) \propto 1/\sqrt{n}. \tag{4.30}$$

The prior in (4.30) is not the only that can be derived through the continuous embedding technique illustrated in Berger et al. (2012). In fact, not only by using another approach (e.g. Approach 4, discussed in Section 2.3.1) a different prior for n can be derived. It is possible to derive a different objective prior for the parameter by using a different consistent (and inefficient) estimator for n. For example, by considering

$$\hat{n} = (\sqrt{1+16S^2} - 1)/2,$$

with $S^2 = \sum x_i^2/k$. The relative prior is then

$$\pi(n) \propto \frac{1}{\sqrt{n}} \times \frac{2pn+1-p}{\sqrt{4p^2n^2 + 2pn(3-5p) + 1 - 6p(1-p)}}. \tag{4.31}$$

Similarly as done in Section 4.4, the "optimal" prior is selected by comparing the frequentist performances of the posterior distribution. In this important to note that priors (4.30) and (4.31) result in posteriors with very similar distributions, although the priors themselves are different. The argument in support of the first one, as discussed in Berger et al. (2012), is on the basis that (4.30) has a simpler functional form than (4.31).

4.5.1 The Villa–Walker prior for the Binomial model

Let us consider the following two binomial distributions: f_n and $f_{n'}$, with $n = 1, 2, \ldots$, $n' = 1, 2, \ldots$ and $n \neq n'$, where n and n' represent the number of trials for each distribution. Furthermore, we assume that both distributions have the same value of the known parameter p. The Kullback–Leibler divergence between f_n and $f_{n'}$ is given by

$$D_{KL}(f_n \| f_{n'}) = +\infty,$$

if $n' < n$; else

$$\begin{aligned} D_{KL}(f_n \| f_{n'}) &= \sum_{x=0}^{n} \left\{ f_n \log\left(\frac{f_n}{f_{n'}}\right) \right\} \\ &= \sum_{x=0}^{n} \left\{ \binom{n}{x} p^x (1-p)^{n-x} \times \log\left[\frac{\binom{n}{x} p^x (1-p)^{n-x}}{\binom{n'}{x} p^x (1-p)^{n'-x}} \right] \right\} \\ &= \mathbb{E}\left\{ \log\binom{n}{x} \right\} - \mathbb{E}\left\{ \log\binom{n'}{x} \right\} + (n-n')\log(1-p), \quad (4.32) \end{aligned}$$

where $\mathbb{E}$ represents the expected value with respect to f_n. The following Lemma 4.3, which proof is in Appendix A, shows that the nearest model to f_{n_0} is f_{n_0+1}.

Lemma 4.3. *Consider the binomial distribution f_{n_0}. If we indicate by f_n, with $n > n_0$, the generic binomial distribution that differs from f_{n_0} only by the number of trials, then the Kullback–Leibler divergence from f_{n_0} to f_n is minimum when $n = n_0 + 1$.*

Thus, the minimum divergence from model f_n is obtained by setting $n' = n+1$ in expression (4.32), which results in the following Kullback–Leibler divergence

$$
\begin{aligned}
D_{KL}(f_n\|f_{n+1}) &= \sum_{x=0}^{n}\left\{\binom{n}{x}p^x(1-p)^{n-x}\log\binom{n}{x}\right\}\\
&-\sum_{x=0}^{n}\left\{\binom{n}{x}p^x(1-p)^{n-x}\log\binom{n+1}{x}\right\}\\
&+[n-(n+1)]\log(1-p)\\
&=\sum_{x=0}^{n}\left\{\binom{n}{x}p^x(1-p)^{n-x}\log\left(\frac{n!}{x!\,(n-x)!}\,\frac{x!\,(n+1-x)!}{(n+1)!}\right)\right\}\\
&-\log(1-p)\\
&=\sum_{x=0}^{n}\left\{\log(n+1-x)\binom{n}{x}p^x(1-p)^{n-x}\right\}-\log(n+1)-\log(1-p).
\end{aligned}
$$

Therefore, the prior distribution is obtained by applying (4.6) and has the form

$$
\pi(n)\propto\frac{1}{(n+1)(1-p)}\exp\left\{\sum_{x=0}^{n}\log(n+1-x)\binom{n}{x}p^x(1-p)^{n-x}\right\}-1,
$$

which can alternatively be written as

$$
\pi(n)\propto\frac{1}{(n+1)(1-p)}\prod_{x=0}^{n}\left\{(n+1-x)^{\binom{n}{x}p^x(1-p)^{n-x}}\right\}-1, \qquad (4.33)
$$

The form of the prior in (4.33) is useful to employ in some analytical procedures.

The objective prior for n obtained is improper. In fact, we have that

$$
\sum_{n=1}^{\infty}\left[\frac{1}{(n+1)(1-p)}\prod_{x=0}^{n}\left\{(n+1-x)^{\binom{n}{x}p^x(1-p)^{n-x}}\right\}-1\right]\geq
$$
$$
\sum_{n=1}^{\infty}\left[\frac{1}{(n+1)(1-p)}-1\right], \qquad (4.34)
$$

as $\prod_{x=0}^{n}\left\{(n+1-x)^{\binom{n}{x}p^x(1-p)^{n-x}}\right\}/(1-p)\geq 1$. Therefore, the left-hand side of (4.34) diverges as well, given that $\sum_{n=1}^{\infty}[1/\{(n+1)(1-p)\}-1]=\infty$. Therefore, given that the prior distribution is improper, we need to show that the resulting posterior is proper. This is stated in Theorem 4.4 below.

Theorem 4.4. *Assume that we observe the data point x_1 from a binomial distribution with parameters n and p. Also, assume a prior distribution for the parameter n as $\pi(n) \propto \exp\{D_{KL}(f_n \| f_{n+1})\}$. Then, the posterior distribution is given by*

$$\pi(n|x_1) \propto \left[\frac{1}{(n+1)(1-p)} \prod_{x=0}^{n} \left\{(n+1-x)^{\binom{n}{x}p^x(1-p)^{n-x}}\right\} - 1\right] \\ \times \binom{n}{x_1} p^{x_1}(1-p)^{n-x_1}, \quad (4.35)$$

is proper.

Proof. The likelihood function in the posterior (4.35) converges for $n \to \infty$. In fact we have

$$\sum_{n=x_1}^{\infty} \left\{\binom{n}{x_1} p^{x_1}(1-p)^{n-x_1}\right\} = \frac{1}{x_1!}\left[\left(\frac{p}{1-p}\right)^{x_1} \times \sum_{n=x_1}^{\infty}\left\{\frac{n!}{(n-x_1)!}(1-p)^n\right\}\right],$$

and as $n!/(n-x_1)! = n \times (n-1) \times \cdots \times (n-x_1+1) \leq n^{x_1}$, it follows

$$\frac{1}{x_1!}\left(\frac{p}{1-p}\right)^{x_1} \times \sum_{n=x_1}^{\infty}\left\{\frac{n!}{(n-x_1)!}(1-p)^n\right\} \leq \frac{1}{x_1!}\left(\frac{p}{1-p}\right)^{x_1} \times \sum_{n=x_1}^{\infty}\left\{n^{x_1}(1-p)^n\right\}.$$

To prove that $\sum_{n=1}^{\infty}\{n^{x_1}(1-p)^n\}$ converges, we show that

$$\lim_{n \to +\infty} \frac{(n+1)^{x_1}(1-p)^{n+1}}{n^{x_1}(1-p)^n} < 1.$$

In fact, we have

$$\lim_{n\to\infty} \frac{(n+1)^{x_1}(1-p)^{n+1}}{n^{x_1}(1-p)^n} = \lim_{n\to\infty}\left\{\left(\frac{n+1}{n}\right)^{x_1}(1-p)\right\} = 1 \cdot (1-p) < 1,$$

Therefore, the series

$$\sum_{n=x_1}^{\infty}\{n^{x_1}(1-p)^n\} < \infty, \quad (4.36)$$

converges. From the result in Lemma A.1 in Appendix A, we have

$$\frac{1}{(n+1)(1-p)}\prod_{x=0}^{n}\left[(n+1-x)^{\binom{n}{x}p^x(1-p)^{n-x}}\right]-1<\infty. \tag{4.37}$$

By combining the results in (4.36) and (4.37), we conclude that

$$\sum_{n=x_1}^{\infty}\left\{\frac{1}{(n+1)(1-p)}\prod_{x=0}^{n}\left[(n+1-x)^{\binom{n}{x}p^x(1-p)^{n-x}}\right]-1\right\} \\ \times\left\{\binom{n}{x_1}p^{x_1}(1-p)^{n-x_1}\right\}<\infty,$$

that is that the posterior distribution for n becomes proper after only one observation. □

The consistency of the posterior distribution is examined in the following Theorem 4.5. The proof in in Appendix A.

Theorem 4.5. *Consider the family of binomial distributions f_n, with $n = 1, 2, \ldots$ and common parameter p. Also assume that the true value of the parameter n is n_0. Given the prior distribution $\pi(n)$ and the set of observations from f_{n_0}, $x = (x_1, \ldots, x_k)$, the mass of the posterior corresponding at n_0 converges to 1 almost surely. That is,*

$$\pi(n_0|n \geq n_0, x_1, \ldots, x_k) \to 1,$$

for $k \to \infty$.

To have an understanding of the objective prior for n, we have computed and plotted it, for a given value of $p = 0.5$, and for $n = 1, \ldots, 100$. The result is shown in Figure 4.8. We see that the highest mass is put on $n = 1$, as the largest divergence is $D_{KL}(f_1\|f_2)$, and that it will decrease as n increases. This is obvious, as the difference between a binomial with n number of trials and the binomial with $n+1$ number of trials, diminishes as n tends to infinity; as such, the Kullback–Leibler divergence measured between the two models tends to zero and so does the prior.

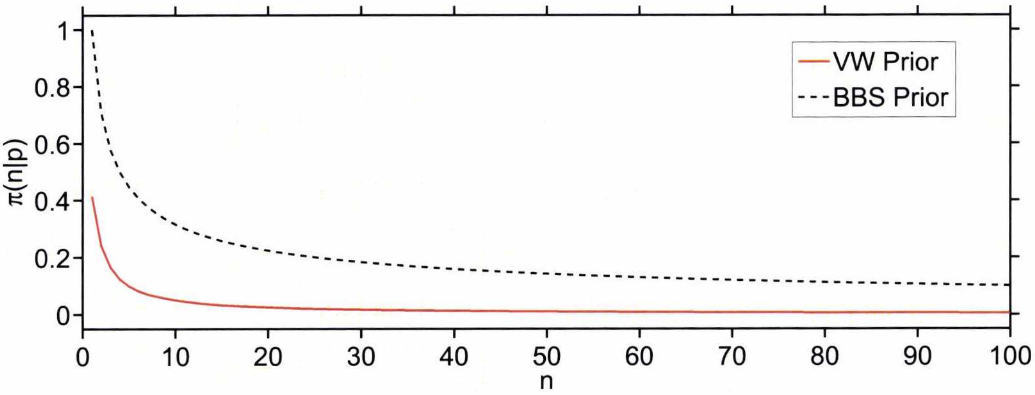

Figure 4.8: Comparison of our prior (continuous red line) and the one obtained by Berger et al. (2012) (dashed black line). The priors refer to parameter n given $p = 0.5$.

The comparison between the prior we propose and the prior defined by BBS is straight forward. By simply inspecting expression (4.30), we notice that the BBS prior distribution assigns large mass to small values of n, and that it decreases for n becoming large (Figure 4.8). In particular, the behaviour is similar to the one of our prior. Even though our prior depends on the value of p, by computation, we have seen that the changes in the prior are negligible.

Unknown p

In the case parameter p is considered as unknown, the joint prior for n and p proposed by Berger et al. (2012), derives from the application of Approach 4 (refer to Section 2.3.1), in combination with the common prior for the probability of success of a Binomial model: Jeffreys' rule prior. The prior for n, when p is unknown, will have the form $\pi(n) \propto 1/n$, whilst the prior for p is the common Beta with both parameters equal to 1/2. That is

$$\pi(n, p) \propto \frac{1}{n} \frac{1}{\sqrt{p(1-p)}}. \tag{4.38}$$

Berger et al. (2012) argue that, although in Jeffreys writings there is no reference to a prior for the parameter vector (n, p) of a Binomial distribution, it is quite plausible to believe that he would have chosen a prior of the form of (4.38). In fact, it is well known that Jeffreys' prior for p is a $Be(1/2, 1/2)$; also, as discussed in Section 2.3, Jeffreys proposed $1/n$ as a prior for an infinite positive parameter. Therefore, prior (4.38) can also be interpret by the product of the priors that Jeffreys has (or would have had) proposed for this particular problem.

Raftery (1988) proposed as prior for n and p, simply $1/n$. This prior has the tail of the posterior that becomes sharper and sharper as n grows, as shown in Berger et al. (1999).

Although it is not in the scope of this thesis to discus in details the application to our approach to continuous parameter spaces, it is not complicated to combine the results for n discussed in this chapter with the common Beta prior for p. In fact, we can define the prior on n and p and $\pi(n,p) \propto \pi(n|p)\pi(p)$, where $\pi(p) \sim Be(1/2, 1/2)$ (Jeffreys' prior) would be a natural choice. That is

$$\pi(n,p) \propto \left[\frac{1}{(n+1)(1-p)} \exp\left\{\sum_{x=0}^{n} \log(n+1-x)\binom{n}{x} p^x (1-p)^{n-x}\right\} - 1\right] \times \frac{1}{\sqrt{p(1-p)}},$$

which, even though improper, it yields to a proper posterior.

Chapter 5

Objective Prior for the Number of Degrees of Freedom of a t Distribution

The content of this chapter, included the illustrations, is taken from Villa and Walker (2013c).

In this section we introduce an objective prior for the number of degrees of freedom of a Student's t probability density function. From now on, the model will be simply identified as t distribution or t density.

The objective prior we propose, based on the approach thoroughly discussed in Section 3, assumes that the parameter ν, representing the number of degrees of freedom, is discrete. We will motivate our choice later in the chapter.

The parameter ν is typically problematic to estimate. In particular, a problem in objective Bayesian inference is that improper priors lead to improper posteriors, whilst proper priors may dominate the likelihood (Fonseca et al., 2008).

The prior that we construct takes into consideration an important property of the t distribution: its convergence to a Normal density when ν tends to infinity. Actually, the approximation to normality reaches remarkable levels for relatively small values of the number of degrees of freedom. It is in fact common practice to assume the approximation as acceptable for $\nu \geq 30$. As a consequence, the prior we propose is truncated.

5.1 Introduction

In disciplines such as finance and economics, extreme values tend to occur at a probability rate that is too high to be effectively modelled by distributions with appealing analytical properties, such as the normal. This is the case, for example, of financial asset returns and market index values, whose behaviour of extreme values is better represented by distributions with tails heavier than the normal distribution; in particular, see Fabozzi et al. (2010), the t distribution represents an appealing alternative. Furthermore, in Maronna (1976), Lange et al. (1989) and West (1984), it is pointed out that heavy-tailed data are more efficiently handled by regression models for which the error term is assumed to be t-distributed. In fact, it is shown that the influence of outliers is significantly reduced, leading to a more robust analysis; in particular, the smaller the number of degrees of freedom, the more robust the analysis tends to be. As such, the possibility of discerning between t distributions with different numbers of degrees of freedom, especially when the value of this parameter is small, represents an important step of the regression analysis and, in general, whenever a t model is deemed to be the most suitable in representing the observations of interest.

By considering the three-parameter representation of the t density, we introduce an objective Bayesian prior mass function for the degrees of freedom ν of a t distribution, conditional on the mean parameter μ and variance parameter σ^2. Hence, it will be of the form $\pi(\nu|\mu, \sigma^2)$. However, we first review some of the most important priors for ν existing in the literature.

Let us define by $\pi(\nu)$ the prior distribution for the number of degrees of freedom of a t density. In the case the number of degrees of freedom are considered as a continuous quantity , we have $\nu \in (0, \infty)$. If ν is consider to take values in a countable set, then we would have $\nu = 1, 2, \ldots$.

The likelihood for ν given μ and σ^2 tends to a positive constant as $\nu \to +\infty$ (Anscomber, 1967). As such, to have a proper posterior, the prior distribution has to tend to 0 as $\nu \to +\infty$. Therefore, the natural objective prior

$$\pi(\nu) \propto 1,$$

cannot be adopted as the posterior would be improper. In fact, as shown in Fernández and Steel (1999b), this behaviour of the likelihood function may lead, in general, to an improper posterior when the prior distribution is improper.

To overcome this issue, Jacquier et al. (2004) proposed a truncated uniform prior on the discrete integer degrees of freedom. In particular, they note that the variance of a t density exists only for values of $\nu \geq 3$. Furthermore, for values of $\nu \in [41, 50]$, the model does not have significant changes in behaviour and therefore, their discrete uniform prior is

$$\pi(\nu) \propto 1, \qquad 3 \leq \nu \leq 40.$$

According to Fonseca et al. (2008), this type of priors is inappropriate, because the estimate of the number of degrees of freedom is sensitive to the chosen truncation point.

Geweke (1993) proposes a prior distribution that is exponential. In this case, the parameter ν is considered continuous and the distribution depends on a value c, which is strictly positive

$$\pi(\nu) \propto c \exp\{-c\nu\} \qquad \nu > 0.$$

This prior, in our opinion, cannot be considered as strictly objective. In fact, different values of c will lead to a different distribution of the mass over small values of ν, where it is more critical to be able to estimate the number of degrees of freedom. Furthermore, as shown in Fonseca et al. (2008), the exponential prior tends to dominate the data.

In Fonseca et al. (2008), a linear regression model with p covariates and error term t-distributed is considered. The authors define two prior distributions for ν, both based on Jeffreys' prior Jeffreys (1961): the independence Jeffreys prior

$$\pi_I(\nu) \propto \left(\frac{\nu}{\nu+3}\right)^{1/2} \left\{\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right\}^{1/2} \qquad \nu > 0, \qquad (5.1)$$

and the Jeffreys-rule prior

$$\pi_J(\nu) \propto \pi^I(\nu) \left(\frac{\nu+1}{\nu+3}\right)^{p/2} \qquad \nu > 0. \qquad (5.2)$$

It is shown that both priors are proper, and that they lead to proper posteriors.

Prior distributions, though not objective, for the number of degrees of freedom of a t distribution, are given by Juárez and Steel (2010), where a non-hierarchical and a hierarchical prior are considered. The first is a particular gamma, with parameters 2 and 1/100, leading to the density

$$\pi_1(\nu) = \frac{\nu}{100} e^{-\nu/10}. \qquad (5.3)$$

This prior has the property of covering a large range of relevant values of degrees of freedom and allows for all prior moments to exist. The hierarchical prior is obtained by considering and exponential distribution for the scale parameter of the gamma, with shape parameter 2. The resulting density is

$$\pi_2(\nu) = 2k \frac{\nu}{(\nu+k)^3},$$

where $k > 0$ is the hyper-parameter. The authors compared the performance of their priors with the Jeffreys' independent prior proposed by Fonseca et al. (2008), noting that there were no significant differences for values of ν below 50.

It has to be noted that in Geweke (1993), Fonseca et al. (2008) and Juárez and Steel (2010), the number of degrees of freedom is considered as continuous.

5.2 Preliminaries

We make here some fundamental preliminary considerations.

We consider the parameter space of ν to be discrete, that is restricted to positive integers. The motivation is practical. In fact, the Kullback–Leibler divergence between contiguous densities rapidly decreases to zero, making necessary large amount of information about ν (i.e. observations) in order to discern between different t distributions (Jacquier et al., 2004). We could make more dense the parameter space, for example $\nu = \{1, 1.5, 2, 2.5, \dots\}$ (or even more dense), and apply our criterion to derive a prior, but the resulting increase in precision of the estimate of ν would not be of any practical use, as, for example, there is no sensible difference in having a t density with 7 degrees of freedom and one with 7.1 degrees of freedom.

The second remark we would like to discuss originates from the well known property of the t distribution to converge to a normal distribution when the degrees of freedom tend to infinity. That is, from a certain point in the parameter space of degrees of freedom, the distribution can be considered as normal. The key point we wish to make is that it is not fundamental where the quantification of this *turning point* is (i.e. where a t distribution turns into a normal), but the fact that there is one, and that every t distribution with a value of ν equal or larger than this *turning point* is considered the same model, that is, a normal distribution. We take this point to be 30 based on theoretical results, see Chu (1956), and also Section 5.3. It follows that the set of parameter values on which the prior $\pi(\nu)$ is built becomes a finite set of models and ν translates to a label associated to each model. If we indicate the turning point as ν_{max}, the set of models is represented by $\{f_1, f_2, \dots, f_{\nu_{max}-1}, f_{\nu_{max}}\}$, where the first $(\nu_{max} - 1)$ models are t distributions with degrees of freedom $\nu = 1, 2, \dots, \nu_{max} - 1$, and $f_{\nu_{max}} \approx N(\mu, \sigma^2)$.

A direct consequence of this consideration is that it reveals an important conceptual gap common to other objective approaches to derive $\pi(\nu)$. Even though it is theoretically possible to discern between two t distributions with any number of degrees of freedom, provided a sufficiently large number of observations is available, this task loses meaning when the number of degrees of freedom is large enough. It follows that, if we want to assign prior mass to models, for example, in intervals $[f_{200}, \dots, f_{299}]$ and $[f_{300}, \dots, f_{399}]$, this mass has to be the same for each element, as these models are in practice not distinguishable. As such, if we define a prior of ν for values that go from one to infinity, this prior has to be uniform in the interval $[\nu_{max}, +\infty)$, and therefore improper. But, as we have discussed above, all the models in this interval are (approximatively) represented by a normal distribution and, as a result, the set of options has to be finite with the *last* element equal to a normal. Furthermore, as all the models from $f_{\nu_{max}}$ onwards are virtually the same model (i.e. normal), if $\pi(\nu)$ is defined over the whole sample space, it means that a large amount of mass is put on the normal model. And there is no apparent justification for this approach.

If random variable x has a t distribution with degrees of freedom ν, location parameter μ and scale parameter σ^2, its probability density function is represented by

$$f(x|\nu, \mu, \sigma^2) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(\frac{1}{\nu\sigma^2}\right)^{\frac{1}{2}} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \qquad -\infty < x < \infty,$$

where $B(\cdot,\cdot)$ is the beta function. Both location and scale parameters are continuous, with $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The density of x can equivalently be expressed in terms of the precision parameter $\lambda = 1/\sigma^2$ as follows

$$f(x|\nu, \mu, \lambda) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(\frac{\lambda}{\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad -\infty < x < \infty.$$

We mainly focus on the particular case where $\mu = 0$ and $\sigma^2 = 1$; it is always possible to move from a t distribution with $\mu = 0$ and $\sigma^2 = 1$ to a t distribution with any value of the parameters (and vice versa) by simply applying the relationship $x_{\nu,\mu,\sigma^2} = \mu + \sigma x_{\nu,0,1}$. In any case, as we are interested in comparing t distributions that differ only in the number of degrees of freedom, to avoid a cumbersome notation, the t model with ν degrees of freedom and parameters μ and σ^2 is represented as f_ν in lieu of $f(x|\nu, \mu, \sigma^2)$.

Let us consider the following t distributions: f_ν and $f_{\nu'}$, with $\nu \neq \nu'$. Also, we assume that location and scale parameters are equal for both densities, with $\mu = 0$ and $\sigma^2 = 1$. The Kullback–Leibler divergence between f_ν and $f_{\nu'}$ is given by

$$
\begin{aligned}
D_{KL}(f_\nu \| f_{\nu'}) &= \int_{-\infty}^{\infty} f_\nu \log\left(\frac{f_\nu}{f_{\nu'}}\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\nu}B(1/2,\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \log\left\{\frac{\frac{1}{\sqrt{\nu}B(1/2,\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\frac{1}{\sqrt{\nu'}B(1/2,\nu'/2)}\left(1+\frac{x^2}{\nu'}\right)^{-\frac{\nu'+1}{2}}}\right\} dx \\
&= \log\left\{\frac{\sqrt{\nu'}B(\frac{1}{2},\frac{\nu'}{2})}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}\right\} - \frac{\nu+1}{2}\mathbb{E}_\nu\left[\log\left(1+\frac{x^2}{\nu}\right)\right] + \frac{\nu'+1}{2}\mathbb{E}_\nu\left[\log\left(1+\frac{x^2}{\nu'}\right)\right],
\end{aligned}
\tag{5.4}
$$

where $\mathbb{E}_\nu$ represents the expected value with respect to f_ν. To identify the nearest model, in terms of Kullback–Leibler divergence, we have numerically computed the expression in (5.4), for $\nu > 1$ to compare $D_{KL}(f_\nu \| f_{\nu-1})$ and $D_{KL}(f_\nu \| f_{\nu+1})$. Figure 5.1 shows that $D_{KL}(f_\nu \| f_{\nu-1}) > D_{KL}(f_\nu \| f_{\nu+1})$, for any ν, and that the divergence decreases as the number of degrees of freedom tend to infinity. The result obtained is independent from the choice of μ and σ^2.

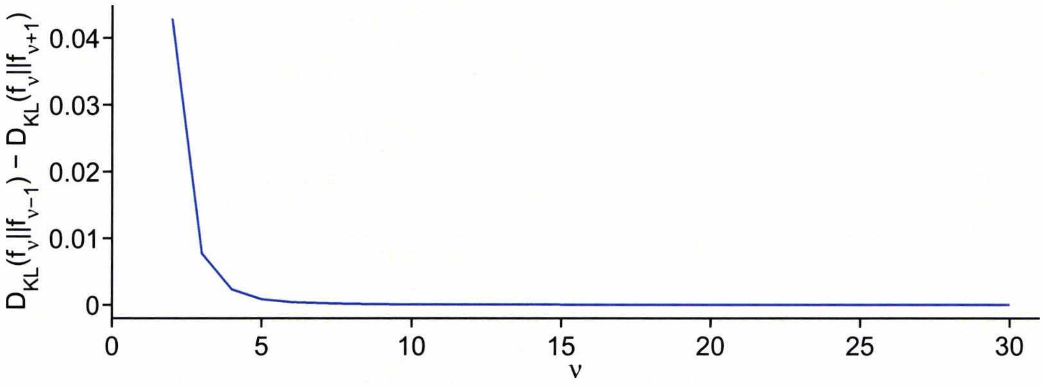

Figure 5.1: Numerical computation of $D_{KL}(f_\nu \| f_{\nu-1}) - D_{KL}(f_\nu \| f_{\nu+1})$, for $\nu = 2, \ldots, 30$. The result does not depend on μ and σ^2.

We have anticipated that the prior we propose is truncated, and that this is done to avoid assigning more mass than appropriate to the normal model. As such, the Kullback–Leibler divergence at the points of the parameter space near to and at the truncation have to be discussed separately. First, we note that the minimum Kullback–Leibler divergence at the truncation point is given by

$$\begin{aligned} D_{KL}(N_{0,1}\|f_\nu) &= \int_{-\infty}^{\infty} N_{0,1}\log\left(\frac{N_{0,1}}{f_\nu}\right)dx \\ &= \log\left\{\frac{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}{\sqrt{2\pi}}\right\} - \frac{1}{2}\mathbb{E}_N\left[x^2\right] + \frac{\nu+1}{2}\mathbb{E}_N\left[\log\left(1+\frac{x^2}{\nu}\right)\right], \end{aligned} \tag{5.5}$$

where $N_{0,1}$ is the standard Normal, and $\mathbb{E}_N$ represents the expected value with respect to $N_{0,1}$. If we indicate by $f_{\nu_{max}}$ the Normal model at the truncation point, the nearest distribution to $f_{\nu_{max-1}}$ is $f_{\nu_{max-2}}$, as the numerical computation in Table 5.1 shows. The results can be summarised as follows. If the set of densities is given by $\{f_1, f_2, \ldots, f_{\nu_{max-1}, f_{\nu_{max}}}\}$, with $f_{\nu_{max}} \approx N(0,1)$, the minimum divergence for $\nu = 1, \ldots, \nu_{max-2}$ is $D_{KL}(f_\nu\|f_{\nu+1})$; for $f_{\nu_{max-1}}$ and $f_{\nu_{max}}$ it is $D_{KL}(f_\nu\|f_{\nu-1})$.

ν	$D_{KL}(f_{\nu_{max-1}}\|f_{\nu_{max-2}})$	$D_{KL}(f_{\nu_{max-1}}\|f_{\nu_{max}})$
30	2.0399×10^{-06}	0.0021
60	1.3121×10^{-07}	0.0005×10^{-04}
90	2.6168×10^{-08}	0.0002×10^{-04}
120	8.3194×10^{-09}	0.0001×10^{-04}
150	3.4174×10^{-09}	7.9029×10^{-05}
180	1.6513×10^{-09}	5.4735×10^{-05}

Table 5.1: Comparison of the Kullback–Leibler divergence from $f_{\nu_{max-1}}$ to $f_{\nu_{max-2}}$ and from $f_{\nu_{max-1}}$ to $f_{\nu_{max}}$, with $f_{\nu_{max}} \approx N(0,1)$. It can be noted that the last t distribution is closer to the t distribution on its left than to the standard normal.

5.3 The Villa–Walker prior for ν

To define the prior mass function for the degrees of freedom ν of a t distribution, we need to make the following considerations. We assume that the location parameter μ and the scale parameter σ^2 (or, equivalently, the precision λ) are known. Let us consider a random variable x with a t distribution with parameters ν, μ and σ^2.

Therefore, for $\nu \to +\infty$ we have $x \xrightarrow{d} N(\mu, \sigma^2)$. It is common practice to assume normality for $\nu \geq 30$. Chu (1956) shows that the proportional error in using the distribution function of a standard normal, $\Phi(x)$, as an approximation to the distribution function of x, $F(x)$, is smaller than $1/\nu$ for every $\nu \geq 8$, where the proportional error is defined as $E = |(F(x)/\Phi(x)) - 1|$. In fact, the approximation of a t distribution to a normal density is always to a certain level of precision and, apart from computational limitations, it is always possible to find a sample size large enough to be able to discriminate the two distributions for a given precision level. In any case, the prior mass function for the parameter ν is defined over a set of models composed by t distributions with increasing number of degrees of freedom and, as a final model, a normal distribution. This normal distribution can be seen as the model that incorporates all the remaining t distributions for which we assess that the value of ν is too high to make them distinguishable from a normal. Therefore, as introduced in Section 5.2, the prior $\pi(\nu)$ is a function that associates a mass to each model in the finite set $\{f_1, f_2, \ldots, f_{\nu_{max-1}}, f_{\nu_{max}}\}$, where f_ν (for $\nu = 1, \ldots, \nu_{max-1}$) is a t distribution with ν degrees of freedom, and $f_{\nu_{max}}$ is the normal distribution $N(\mu, \sigma^2)$.

For the remainder of this section, without loss of generality, we focus on the special case where $\mu = 0$ and $\sigma^2 = 1$. Our criterion is based on the fact that, if the true model is removed from the set of all possible models, then the posterior distribution will tend to accumulate on the nearest model in terms of the Kullback–Leibler divergence. Then

$$l(\nu) = \begin{cases} -D_{KL}(f_\nu \| f_{\nu-1}) & \text{if } \nu \geq \nu_{max} - 1 \\ -D_{KL}(f_\nu \| f_{\nu+1}) & \text{if } \nu < \nu_{max} - 1, \end{cases}$$

and the derivation of the prior probability from this loss is given by the *self-information* loss function

$$-\log \pi(\nu) = \begin{cases} -D_{KL}(f_\nu \| f_{\nu-1}) & \text{if } \nu \geq \nu_{max} - 1 \\ -D_{KL}(f_\nu \| f_{\nu+1}) & \text{if } \nu < \nu_{max} - 1. \end{cases}$$

The prior mass to be put on each model in the set of options is then obtain by applying (3.6), and is given by

$$\pi(\nu) \propto \begin{cases} \exp\left\{D_{KL}(f_\nu \| f_{\nu-1})\right\} & \text{if } \nu \geq \nu_{max} - 1 \\ \exp\left\{D_{KL}(f_\nu \| f_{\nu+1})\right\} & \text{if } \nu < \nu_{max} - 1. \end{cases} \tag{5.6}$$

The prior for values of $\nu < \nu_{max} - 1$ is obtained by replacing equation (5.4) in the first of (5.6), for which we set $\nu' = \nu + 1$

$$\begin{aligned} \pi(\nu) \propto \frac{\sqrt{\nu+1}B\left(\frac{1}{2}, \frac{\nu+1}{2}\right)}{\sqrt{\nu}B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \exp\bigg\{ &-\frac{\nu+1}{2}\mathbb{E}_\nu\left[\log\left(1+\frac{x^2}{\nu}\right)\right] \\ &+\frac{\nu+2}{2}\mathbb{E}_\nu\left[\log\left(1+\frac{x^2}{\nu+1}\right)\right]\bigg\}. \end{aligned} \tag{5.7}$$

The prior mass for $\nu_{max} - 1$ is again obtained by replacing (5.4) in the first of (5.6), for which we set $\nu' = \nu_{max} - 2$

$$\begin{aligned} \pi(\nu_{-1}) \propto \frac{\sqrt{\nu_{-1}-1}B\left(\frac{1}{2}, \frac{\nu_{-1}-1}{2}\right)}{\sqrt{\nu_{-1}}B\left(\frac{1}{2}, \frac{\nu_{-1}}{2}\right)} \exp\bigg\{ &-\frac{\nu_{-1}+1}{2}\mathbb{E}_{\nu_{-1}}\left[\log\left(1+\frac{x^2}{\nu_{-1}}\right)\right] \\ &+\frac{\nu_{-1}}{2}\mathbb{E}_{\nu_{-1}}\left[\log\left(1+\frac{x^2}{\nu_{-1}-1}\right)\right]\bigg\}. \end{aligned} \tag{5.8}$$

Note that in equation (5.8) we have replaced $\nu_{max} - 1$ by ν_{-1}. Finally, the prior for ν_{max} is obtained by replacing (5.5), for which $\nu = \nu_{max} - 1$, in the second equation of (5.6), obtaining

$$\pi(\nu_{max}) \propto \frac{\sqrt{\nu_{-1}}B\left(\frac{1}{2}, \frac{\nu_{-1}}{2}\right)}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\mathbb{E}_N\left(x^2\right) + \frac{\nu_{-1}}{2}\mathbb{E}_N\left[\log\left(1+\frac{x^2}{\nu_{-1}}\right)\right]\right\}. \tag{5.9}$$

To have a picture of the prior on ν, we have plotted its behaviour for three distinctive values of ν_{max}; in particular, in Figure 5.2 we have explored the cases where the prior has been truncated at $\nu = 30, 60$ and 90. The prior puts the highest value of mass on the first model, the t distribution with 1 degree of freedom, and gradually decreases toward 1 as ν increases. This is a direct consequence of the fact that the models become more and more similar to each other, resulting in a

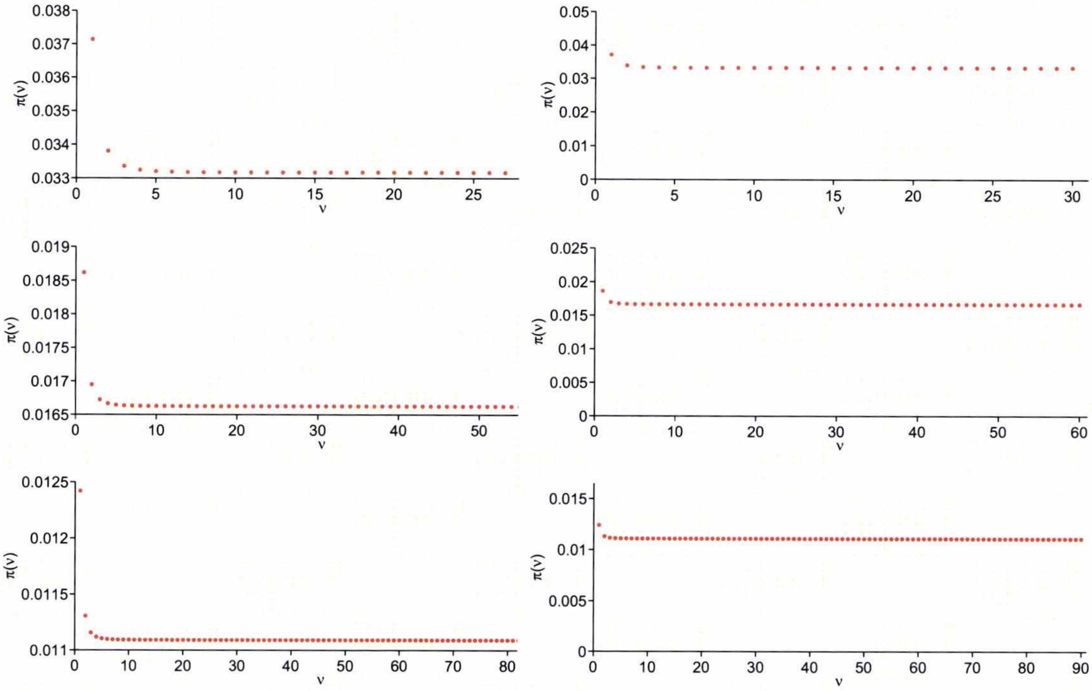

Figure 5.2: Normalised prior distributions for ν truncated at $\nu_{max} = 30$, $\nu_{max} = 60$ and $\nu_{max} = 90$. In the left column we show the distributions on a non-zero scale graph, whilst in the right column the graphs are scaled to zero.

Kullback–Leibler divergence converging to 0. The priors look uniform for $\nu > 5$; however this is a perception caused by the fact that the scale is distorted by the larger values of the prior for the small values. While the prior does look uniform, it is not and the subtle differences are sufficient for the prior not to behave as a uniform prior. And something close to uniform for high degrees of freedom is coherent. For if mass $\pi(\nu)$ has been put on ν then one would expect the mass on $\pi(\nu + 1)$ to be very similar simply because the f_ν and $f_{\nu+1}$ are almost the same density.

The prior distribution has also been analysed for t distributions with different values of μ and σ^2. We have observed that the prior is not affected by changes in the location parameter μ. Although the scale parameter σ^2 has some effect on the prior, that is a larger mass is assigned to values of $\nu \leq 5$ for increasing values of σ^2, there is no change in the tail of the distribution. However, the posterior is not significantly affected by this, given that the main effect of the prior on the posterior is in the tails, where the priors are remarkably similar, refer for example to (Berger et al., 2012).

5.4 Posterior analysis

We now analyse the posterior for the degrees of freedom when the prior obtained with our approach is used on simulate data. We first analyse the result on simulated data from an identically distributed sample. Then, we assume to deal with a regression model where the error term is supposed to follow a t distribution.

5.4.1 Sampling algorithm

By combining the likelihood function for parameter ν (given μ and σ^2) for a t distribution, that is

$$L(\nu|\mu,\sigma^2,x) = \prod_{i=1}^{n}\left\{\frac{1}{B\left(1/2,\nu/2\right)}\left(\frac{1}{\nu\sigma^2}\right)^{1/2}\left(1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}\right\},$$

with the appropriate prior for ν in (5.7), (5.8) or (5.9), in which we have included parameters μ and σ^2, we obtain, respectively, the following three posterior distributions

$$\begin{aligned}\pi(\nu|\mu,\sigma^2,x) \propto \prod_{i=1}^{n}&\left\{\frac{1}{B\left(1/2,\nu/2\right)}\left(\frac{1}{\nu\sigma^2}\right)^{1/2}\left(1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}\right\}\\ &\frac{\sqrt{\sigma^2(\nu+1)}B\left(\frac{1}{2},\frac{\nu+1}{2}\right)}{\sqrt{\sigma^2\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)}\exp\left\{-\frac{\nu+1}{2}\mathbb{E}_\nu\left[\log\left(1+\frac{(x-\mu)^2}{\sigma^2\nu}\right)\right]\right.\\ &\left.+\frac{\nu+2}{2}\mathbb{E}_\nu\left[\log\left(1+\frac{(x-\mu)^2}{\sigma^2(\nu+1)}\right)\right]\right\},\end{aligned}$$

for values of $\nu = 1,\dots,\nu_{max}-2$

$$\pi(\nu_{-1}|\mu,\sigma^2,x) \propto \prod_{i=1}^{n}\left\{\frac{1}{B\left(1/2,\nu/2\right)}\left(\frac{1}{\nu\sigma^2}\right)^{1/2}\left(1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}\right\}$$

$$\begin{aligned}\frac{\sqrt{\sigma^2(\nu_{-1}-1)}B\left(\frac{1}{2},\frac{\nu_{-1}-1}{2}\right)}{\sqrt{\sigma^2\nu_{-1}}B\left(\frac{1}{2},\frac{\nu_{-1}}{2}\right)}\exp\Bigg\{&-\frac{\nu_{-1}+1}{2}\mathbb{E}_{\nu_{-1}}\left[\log\left(1+\frac{(x-\mu)^2}{\sigma^2\nu_{-1}}\right)\right]\\ &+\frac{\nu_{-1}}{2}\mathbb{E}_{\nu_{-1}}\left[\log\left(1+\frac{(x-\mu)^2}{\sigma^2(\nu_{-1}-1)}\right)\right]\Bigg\},\end{aligned}$$

for $\nu=\nu_{max}-1$, and

$$\begin{aligned}\pi(\nu_{max}|\mu,\sigma^2,x)\propto\prod_{i=1}^{n}&\left\{\frac{1}{B\left(1/2,\nu/2\right)}\left(\frac{1}{\nu\sigma^2}\right)^{1/2}\left(1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}\right\}\\ &\frac{\sqrt{\sigma^2\nu_{-1}}B\left(\frac{1}{2},\frac{\nu_{-1}}{2}\right)}{\sqrt{2\pi\sigma^2}}\exp\Bigg\{-\frac{1}{2}\mathbb{E}_N\left[(x-\mu)^2/\sigma^2\right]+\\ &\qquad\frac{\nu_{-1}}{2}\mathbb{E}_N\left[\log\left(1+\frac{(x-\mu)^2}{\sigma^2\nu_{-1}}\right)\right]\Bigg\}.\end{aligned}$$

for $\nu=\nu_{max}$. It has to be noted that the posterior distribution is proper as it is finite. Furthermore, the actual posterior for the general case, that is when $\mu\neq 0$ and $\sigma^2\neq 1$, needs to take into consideration the priors for these parameters. We have chosen proper priors for both the location and the scale parameters so that the posterior distributions are proper as well. In particular, $\pi(\mu)$ is normally distributed and $\pi(\sigma^2)$ has an inverse gamma distribution, both with relatively large variance. However, we have also run the simulations with the well known objective priors, that is $\pi(\mu)\propto 1$ and $\pi(\sigma^2)\propto 1/\sigma^2$, and no significant differences were seen.

The above expressions are not analytically tractable. Thus, to study the posterior distribution of the number of degrees of freedom ν, it necessary to use Monte Carlo methods. The main part of the sampling structure is a Gibbs sampler, used to sample from the conditional distribution of each parameter ν, μ and σ^2; at each step, given that the conditional distributions are complex as well, we perform Metropolis-Hastings algorithms.

5.4.2 Independent and identically distributed samples

For the first simulation study, we have considered the t distribution with $\nu=3$, $\mu=0$ and $\sigma^2=1$: $x\sim t(3,0,1)$. We have obtained a random sample of $n=100$

observations and considered a prior truncated at $\nu_{max} = 31$; that is, $f_{31} \approx N(0,1)$. For the parameter of interest ν, in Figure 5.3 we have plotted the histogram of the posterior distribution.

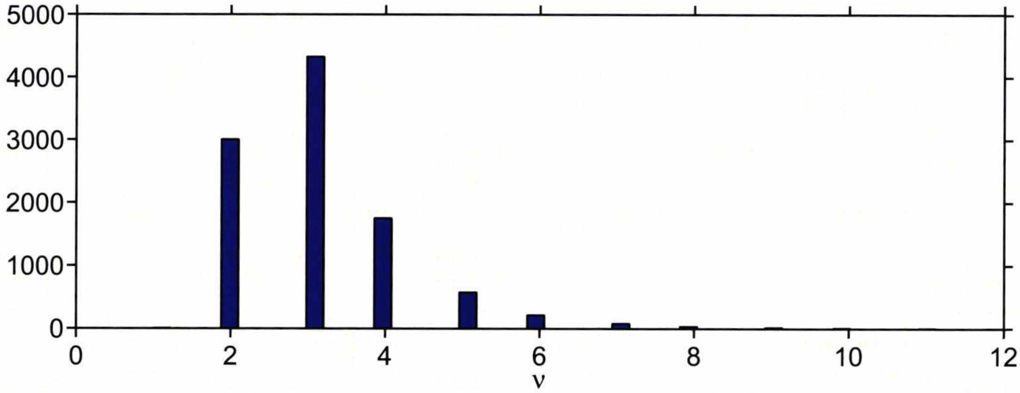

Figure 5.3: Posterior histogram of the parameter ν for an independent sample of size $n = 100$ drawn from a t distribution with $\nu = 3$, $\mu = 0$ and $\sigma^2 = 1$.

The statistics of the posterior distribution are reported in Table 5.2. The posterior distribution is skewed. As such, the most appropriate statistics to be used as the estimate of ν is the median. We note that, for the parameter of interest ν, the median of the posterior is in line with the degrees of freedom of the distribution and that is included the 95% credible interval.

Parameter	Mean	Median	C.I. (95%)
ν	3.12	3	(2, 6)
μ	0.08	0.08	(-0.17, 0.33)
σ^2	1.09	1.07	(0.70, 1.59)

Table 5.2: Posterior mean, median and 95% credible interval for the simulated data from a t distribution with $\nu = 3$, $\mu = 0$ and $\sigma^2 = 1$.

To show that the truncation does not affect the estimate of the number of degrees of freedom, for this case, we have performed simulations for different values of ν_{max}. In particular $\nu_{max} = 35, 45, 60$ and 90. In Figure 5.4 we have plotted the posterior histogram for ν with the above truncation points. It can be seen that the distributions are remarkably similar. However, we would like to stress one more time that a truncation point of 30 (or 31) gives sensible results; therefore, the above analysis is for illustrative purposes only.

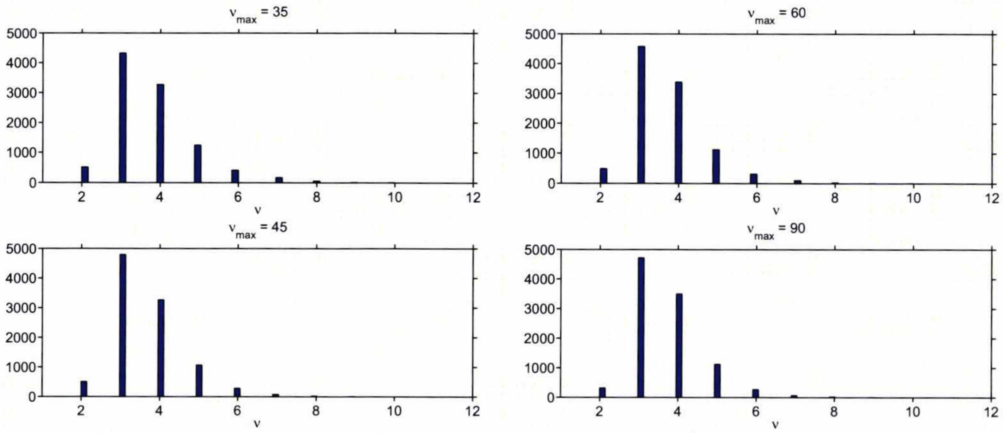

Figure 5.4: Histogram of the posterior distribution of ν obtained by truncating the prior at different points. In particular, at $\nu_{max} = 35$ (top left), $\nu_{max} = 45$ (top right), $\nu_{max} = 60$ (bottom left) and $\nu_{max} = 90$ (bottom right).

Large number of degrees of freedom

Let us now analyse the performance of our objective prior when the data is simulated from a t density with $\nu = 20$, that is a relatively high number of degrees of freedom. In this region of the parameter space consecutive t models are significantly close, therefore difficult to discern. The histogram of the posterior is plotted in Figure 5.5.

We note that the posterior is not very informative, although the median value is $\nu = 20$. The reason, as anticipated at the beginning of this section, has to be ascribed to the relative closeness between t models when the number of degrees of freedom is relatively large. An objective prior, which intent is to "let the data speak", would not heavily contribute to the posterior, in terms of information. Of course, should the number of observations increase, the posterior would be definitely more concentrated around the true value of the parameter, $\nu = 20$. We

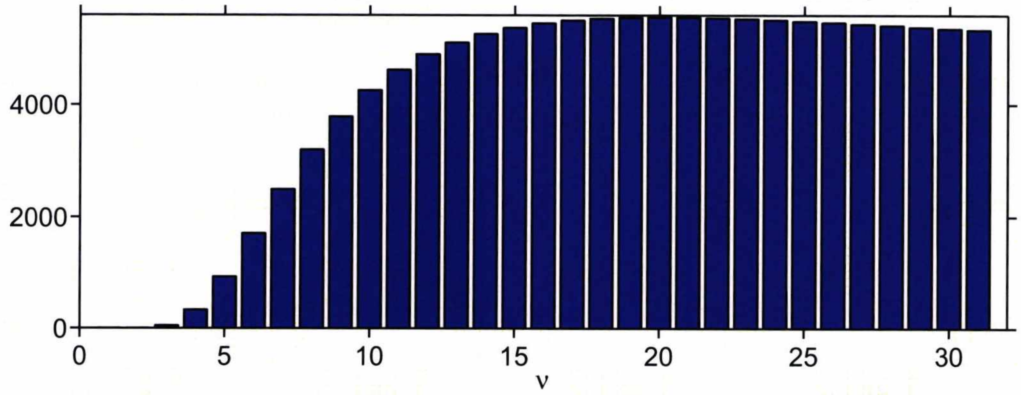

Figure 5.5: Histogram of the posterior distributions of the parameter ν with data simulated from a t density with 20 degrees of freedom.

do not discuss in the detail how a posterior distribution has to be interpreted, as it is not in scope for this thesis. However, the top histogram in Figure 5.5, clearly indicates a more likely value of the number of degrees of freedom at $\nu = 20$. This is also the mode of the distribution.

The wrong model

In Section 5.3, where we have discussed the motivation that lead to a truncated prior for the parameter ν, we argued that this is advisable as the t distribution rapidly converges in distribution to a Normal model, for $\nu \to +\infty$. It is then advisable to analyse how our prior performs when simulated data is sampled from a t distribution with a number of degrees of freedom above the truncation point.

We have then analysed the behaviour of our prior distribution, truncated at $\nu_{max} = 31$, with data simulated from a t distribution with 50 degrees of freedom. It can then be assumed as if the data was originated by a standard Normal model. In Figure 5.6 we have plotted the histogram of the posterior distribution for ν. The posterior tends to accumulate at the truncation point. This behaviour would suggest that the data analysed derives from a model with a number of degrees of freedom that is at least as large as $\nu = 31$ (i.e. the truncation point). The interpretation then, would be to assert that we are examining a phenomenon which is normally distributed. Obviously, we know that the data has been simulated from

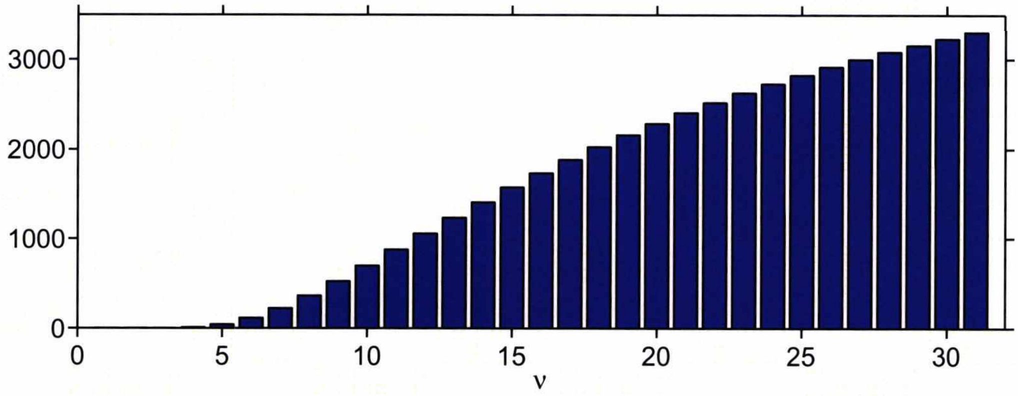

Figure 5.6: Histogram of the posterior distributions of the parameter ν with data simulated from a t density with 50 degrees of freedom.

a t density; but this density has a number of degrees of freedom sufficiently large to be approximated by a Normal model. Given that the prior has been truncated at $\nu_{max} = 31$, this reflects the fact that we believe that from that point (included) onward, the model can be assumed as normal. Hence, the appropriateness of the conclusion.

Small sample size

It is well known that Bayesian analysis has better pay off for relatively small sample sizes. Therefore, we thought appropriate to test the performance of our prior distribution with an i.i.d. sample of size $n = 30$. The sample has been obtained from a t model with parameters $\nu = 5$, $\mu = 0$ and $\sigma^2 = 1$, and the prior was truncated at $\nu_{max} = 31$. The results of the simulation are in Figure 5.7 and Table 5.3. The posterior for ν, although showing a tendency to accumulate on the true value of the parameter, presents a positively skewed behaviour stronger than the case of large sample size. This is reflected by the summary statistics: whilst the median gives a sensible result, the mean indicates a higher value for ν. As said, in the presence of skewed distribution, the central value would be more effectively represented by the median, rather than the mean; therefore, the median is the estimate of ν.

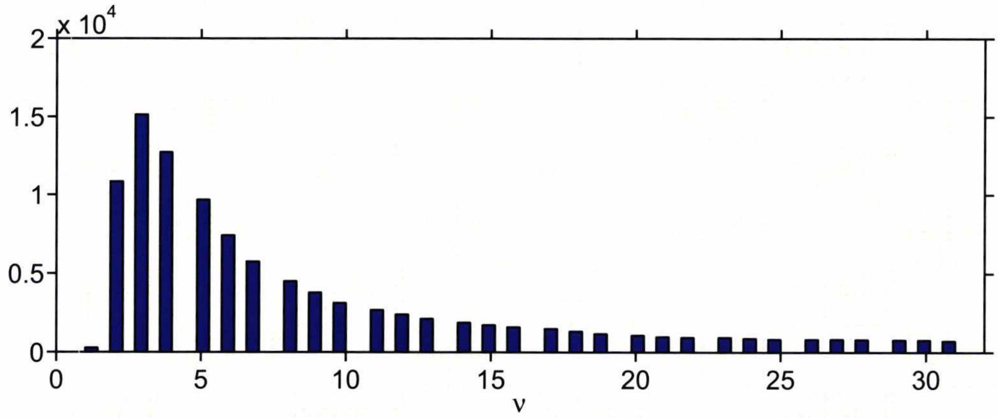

Figure 5.7: Histogram of the posterior sample of ν. Observations ($n = 30$) are sampled from a t distribution with $\nu = 5$, $\mu = 0$ and $\sigma^2 = 1$.

Parameter	Mean	Median	C.I. (95%)
ν	8.33	5	(2,28)
μ	0.12	0.12	(-0.25,0.47)
σ^2	1.17	1.12	(0.52,2.10)

Table 5.3: Posterior mean, median and 95% credible interval for the simulated data ($n = 50$) from a t distribution with $\nu = 5$, $\mu = 0$ and $\sigma^2 = 1$. The posterior for the number of degrees of freedom is positively skewed, therefore the median represents a more suitable estimate of the true value of the parameter.

5.4.3 Regression model

In Section 5.1, at the beginning of this chapter, we have mentioned the importance of assuming t-distributed the error term of a regression model when there are outliers. West (1984) discusses in detail the effect of outliers in Bayesian linear regression. Let us consider the linear regression model

$$y_i = x_i'\beta + \varepsilon_i \qquad i = 1, \dots, n,$$

where $y_1, \dots, y_n$ are the observations of the dependent variable, $x_1, \dots, x_n$ is a set of $p + 1$ vectors representing the covariates, β is the vector with the $p + 1$ parameters (including the intercept) and, finally, $\varepsilon_1, \dots, \varepsilon_n$ is a set of exchangeable random variables with common distribution with one mode ad symmetric around zero. That is, $\varepsilon_i \sim g(0, \sigma)$, where σ is a scale parameter, in general unknown. The idea is that, if we choose function $g(\cdot)$ to be heavy-tailed, relatively to the Normal distribution, the regression model would be less influenced by outliers in the observations.

We consider the case where the scale parameter σ is known only, as this is the assumption for our analysis. Let $\pi(\beta)$ be the prior for the vector of parameters. Thus, by representing as D_n the set of n observations of the dependent variable and the covariates (i.e. the data), we have that the score function of the posterior $\pi(\boldsymbol{\beta}|D_n)$ is given by

$$\frac{d}{d\beta} \log \pi(\beta|D_n) = \frac{d}{d\beta} \log \pi(\beta) + \sum_{i=1}^{n} x_i h(y_i - x_i'\beta),$$

where function $h(\cdot)$ is the so-called *influence function* (from M-estimation theory) of $g(\varepsilon)$ and has the form $h(\varepsilon) = -d/d\varepsilon \log g(\varepsilon)$. The influence function is what determines the effect on the posterior distribution carried by an observation y_i. In particular, when the distribution of the error term is *outlier-prone* (O'Hagan, 1979), the corresponding influence function would be in a such a way that, the more the observation of the dependent variable is far from its centrality, the less the influence in the estimate will have. On the other hand, if $g(\varepsilon)$ is *outlier-resistant*, the effect of an observation y_i relatively distant from the centre will not be negligible.

O'Hagan (1979) introduces clear conditions under which a distribution is either outlier-prone or outlier resistant. On the basis of this, he shows that, among the others, the t density is an outlier-prone distribution, therefore suitable to model the regression error term when there are relevant outliers in the data set. He also shows that the Normal distribution, more commonly used as the distribution of the error term, is outlier-resistant. Thus, not appropriate to handle outliers. Although the appropriate choice for representing the behaviour of the error term in a regression model seems to be a t density, it is still more common to assume that this distribution is Normal. This is because the latter is more tractable than the former.

For our simulation, we consider a linear regression model with one covariate, that is

$$y_i|x_i \sim t(\beta_0 + \beta_{1i}x_i, \sigma^2|\nu) \qquad i = 1, \ldots, n, \tag{5.10}$$

where β_0 and β_1 are the regression parameters, σ^2 the regression variance and ν the number of degrees of freedom of the t-distributed errors. For the purpose of this simulation, we have set $\beta_0 = 10$, $\beta_1 = 10$, $\sigma^2 = 4$ and $\nu = 5$. We have generated $n = 100$ observations from a uniform with parameters 0 and 1: $x_i \sim U(0, 1)$. Then, values y_i have been obtained according to the model in (5.10).

The prior for ν is the one obtained according to our approach. In particular, for this simulation we have considered a prior truncated at $\nu_{max} = 31$. As for the independent sample, we have used a Gibbs sampler with Metropolis-Hastings steps for each parameter. Figure 5.8 shows the histogram of the posterior for

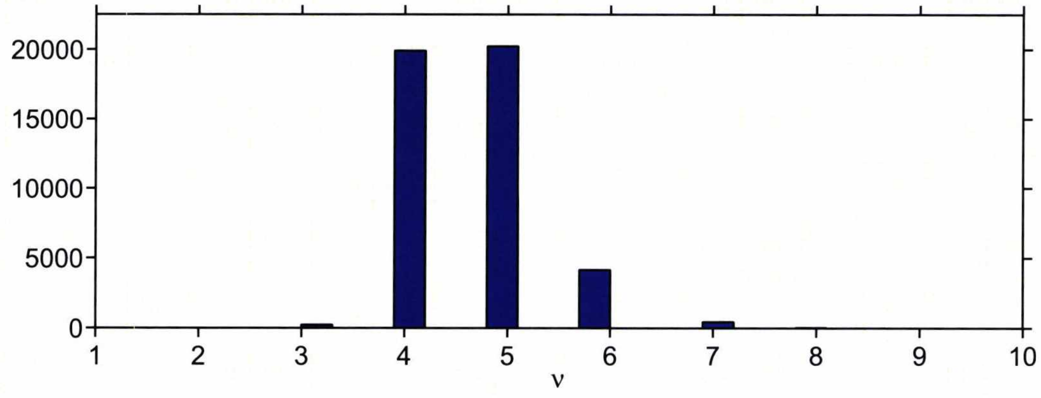

Figure 5.8: Histogram of the posterior distributions for ν in a linear regression study. The parameters of the regression model where the data has been sampled from, were $\nu = 5$, $\beta_0 = 10$, $\beta_1 = 10$ and $\sigma^2 = 4$.

Parameter	Mean	Median	C.I. (95%)
ν	5	4.67	(4, 6)
β_0	9.99	9.99	(9.70, 10.26)
β_1	10.17	10.17	(9.68, 10.67)
σ^2	3.89	3.87	(3.36, 4.50)

Table 5.4: Posterior median and 95% credible interval for the regression simulation. The parameters were set to $\nu = 5$, $\beta_0 = 10$, $\beta_1 = 10$ and $\sigma^2 = 4$.

ν, and Table 5.4 the statistics of the simulation. By inspecting both sources of information, we conclude that the prior has performed well.

5.5 Application

To illustrate the proposed prior on real data, we analyse a sample of the daily closing values of the Dow Jones Industrial Average index of the U.S. stock market. In particular, the data from 11 November 2008 to 4 May 2009, that is 98 observations. This data sample is part of a wide sample analysed in Lin et al. (2012), which ranged from 22 October 2008 to 22 October 2009. Given that the objective of Lin et al. (2012) was to estimate variance change-points in the series, we have focussed our analysis on a subset with estimated constant variance.

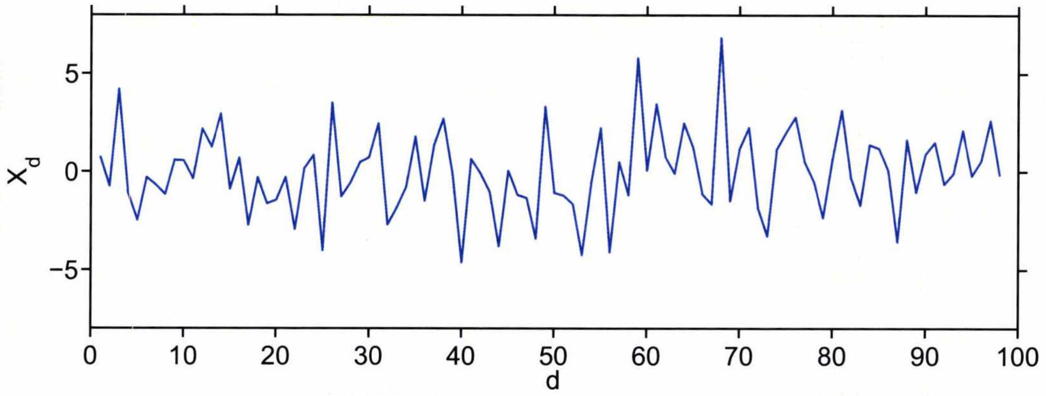

Figure 5.9: Daily return (multiplied by 100) of the closing Dow Jones index from 11 November 2008 to 4 May 2009.

The actual analysis has been performed on the daily returns, multiplied by 100. That is, $X_d = \left[\left(Y_{d+1} - Y_d \right) / Y_d \right] 100$, where Y_d is the market index at day d. The transformed data, for the period of interest, is plotted in Figure 5.9. It can be noted that the series is stationary, and that its variance can be reasonably considered as constant (for the period).

In Table 5.5 we have reported some basic descriptive statistics of the series. The kurtosis is larger than 3 and even though the distribution of the returns does not have tails much heavier than a normal, it seems to be appropriate to consider a t model.

Mean	0.0035
Variance	4.4813
Skewness	0.3216
Kurtosis	3.5626

Table 5.5: Descriptive statistics of the daily Dow Jones index returns from 11 November 2008 to 4 May 2009.

Specifically, the model is

$$X_d = \mu + \varepsilon_d \qquad d = 1, \dots, 98,$$

where $\varepsilon_d \sim t(0, \sigma^2, \nu)$. The result of the simulation are compared, when appropriate, with the ones obtained in (Lin et al., 2012).

We have obtained the posterior distributions for the three parameters by Monte Carlo methods. In Figure 5.10 we have plotted the sample, the progressive median and the histogram of the posterior of the number of degrees of freedom ν only. As the posterior distribution of ν is skewed, the median represents the sensible estimate of the true value of the parameter. The posterior statistics of the parameters are reported in Table 5.6. The results from (Lin et al., 2012) are, $\nu = 8.4873$, $\mu = -0.0406$ and $\sigma^2 = 3.3749$. The authors, as a prior for ν, have used the one proposed by Geweke (1993) (refer to (5.1)) with hyperparameter $\lambda = 0.3$. Therefore, $\pi(\nu) \sim Exp(0.3)$. It has to be noted that the estimate of the degrees of freedom and the mean μ are relative to a larger data set, in particular, for the first 133 observations. However, the authors conclude that the number of degrees of freedom for the whole data set is homogeneous in the range 6.68–8.49 and the mean is zero. The median of the posterior distribution, representing our estimate of the parameter value, is 8 degrees of freedom. We can then conclude that our estimate of ν is in agreement with Lin et al. (2012).

We have analysed the data by adopting priors different from ours. In addition to the independence Jeffreys' prior (5.1) and the Jeffreys-rule prior (5.2) proposed

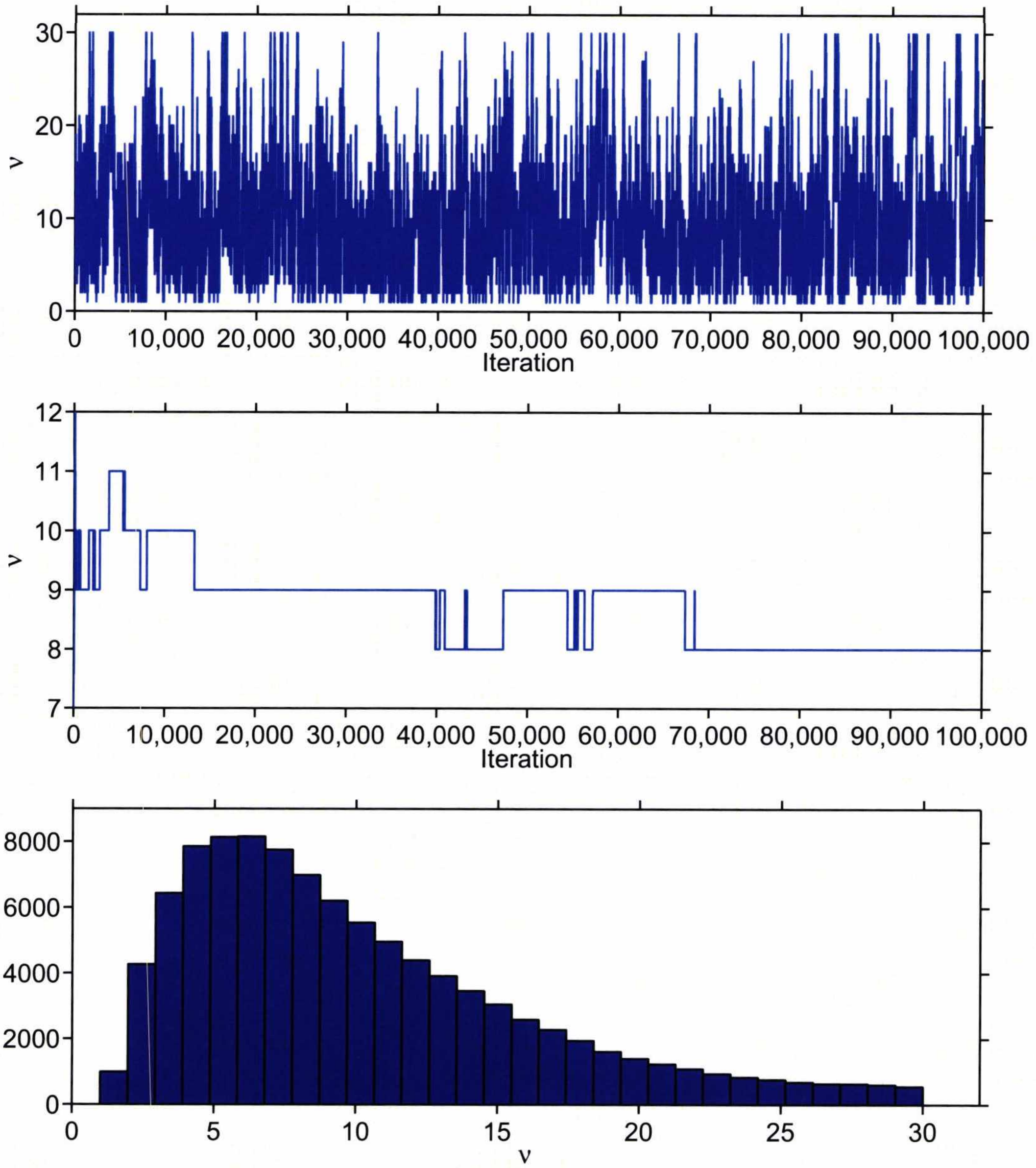

Figure 5.10: Posterior samples (top), posterior progressive median (middle) and posterior histogram (bottom) for the parameter ν.

by Fonseca et al. (2008), we have considered the non-hierarchical prior proposed by Juárez and Steel (2010) in (5.3). The resulting posterior statistics are summarised in Table 5.7. We see that both the independence Jeffreys' and the Jeffreys-rule prior give estimation results that do not differ from ours, considering that our prior

Parameter	Mean	Median	C.I. (95%)
ν	9.96	8	(2, 26)
μ	-0.05	-0.05	(-0.45, 0.36)
σ^2	3.07	3.21	(0.03, 5.61)

Table 5.6: Median and credible interval for the number of degrees of freedom, location and scale parameters for the daily returns of the Dow Jones index, from 11 November 2008 to 4 May 2009.

Prior	Median	C.I. (95%)
$\pi_I(\nu)$	7.30	(3.80, 25.44)
$\pi_J(\nu)$	8.63	(3.46, 31.98)
$\pi_1(\nu)$	15.32	(4.90, 28.89)

Table 5.7: Posterior statistics obtained by using the independence Jeffreys' prior $\pi_I(\nu)$, the Jeffreys-rule prior $\pi_J(\nu)$ and the non-hierarchical gamma prior proposed by Juárez and Steel (2010) $\pi_1(\nu)$.

assumes ν discrete whilst both Jeffreys' do not. However, the credible interval of the Jeffreys-rule prior is larger than the one obtained with our prior and the independence Jeffreys'. For the Dow Jones index data analysed here, the posterior median of ν obtained by applying the gamma prior proposed by Juárez and Steel (2010) is in contrast with our results.

5.6 Discussion

The adoption of t distributed models is an important area of application in finance. This can either be the application of t-distributed random variable to model a certain quantity, such as financial returns, or the assumption that the errors of a linear regression model should have heavier tails than the ones of the more commonly adopted normal distribution. While objective priors for continuous parameters, such as the mean or the variance, can be obtained with various approaches, the estimation of the number of degrees of freedom of a t distribution is not so straightforward.

An important aspect of the objective prior for the number of degrees of freedom we propose, is that it is truncated. This is a consequence of the fact that the t density converges in distribution to the Normal density. Therefore, for a sufficiently large number of degrees of freedom, the model can be considered as normal and it represents the last element in the set of the option models. An important property of the proposed prior is that its estimation performance is not sensible to the point of truncation. We also add that taking the truncation point up to, say, 60 implies an interest in discriminating between a t_{45} and a t_{50}, for example. This is not practical or desirable.

The efficiency of the designed prior for the number of degrees of freedom of a t distribution has been demonstrated through two types of simulation. The first one is based on data simulated from a t density with given parameter values, and the second from data simulated from a given regression model. For the first type we have considered a wide range of scenarios, including relatively large value of ν, "wrong" model (i.e. value of ν above the truncation point) and small sample size.

The analysis on real data appears to give comforting results about the prior. In fact, for the Dow Jones data sample, the estimation of the parameters of the model are in line with the ones obtained by using a different prior. Furthermore, our analysis using two well known objective priors supports again the results obtained.

Chapter 6

Objective Model Selection

6.1 Introduction

The content of this chapter is taken from Villa and Walker (2013b).

In this chapter we introduce a novel approach to objectively determine model prior probabilities for model selection problems. A particular type of model selection, that is variable selection, will be discussed in Chapter 7. The approach is based on our objective criterion (refer to Section 3.1) where we move from the parameter space Θ to the model space $\mathcal{M}$.

We focus on the case where the prior is the pair $\{f(x|\theta), \pi(\theta)\}$, where $f(x|\theta)$ is a probability distribution, characterised by parameter θ (possibly a vector of parameters), and $\pi(\theta)$ is the prior distribution representing beliefs on the model parameters. We assume both $f(x|\theta)$ and $\pi(\theta)$ specified.

A model selection problem is as follows. We have a set of n observations $x = (x_1, \ldots, x_n)$, and a set of possible k models indicated by

$$M_j = \{f_j(x|\theta_j), \pi(\theta_j)\}, \qquad j = 1, \ldots, k.$$

The set of the k models is sometimes identified as the models space, that is $\mathcal{M} = \{M_1, \ldots, M_k\}$.The general aim is to compare the k models. The usual way to perform this comparison is to compute pairwise Bayes factors between the models in the model space. Thus, the Bayes factor between model M_j and model M_i is given by

$$B_{ji} = \frac{m_j(x)}{m_i(x)} = \frac{\int f_j(x|\theta_j)\pi_j(\theta_j)\, d\theta_j}{\int f_i(x|\theta_i)\pi_i(\theta_i)\, d\theta_i}, \qquad i \neq j \in \{1, \dots, k\},$$

where $m_j(x)$ and $m_i(x)$ are the marginal densities of x under, respectively, model M_j and model M_i. We can then see that the Bayes factor B_{ji} is a weighted likelihood ratio (for the observed data) of M_j over M_i, where the weights are represented by the prior probabilities $\pi_j(\theta_j)$ and $\pi_i(\theta_i)$. Then, given model prior probabilities, $P(M_j)$, $j = 1, \dots, k$, the posterior mass for each element in the model space, given the data x, is

$$\begin{aligned} P(M_j|x) &= \frac{P(M_i)m_i(x)}{\sum_{j=1}^{k} P(M_j)m_j(x)} \\ &= \left[\sum_{j=1}^{k} \frac{P(M_j)}{P(M_i)} B_{ji} \right]^{-1}. \end{aligned}$$

Although we focus on the model priors, it is still appropriate to examine how Bayes factors and posterior model probabilities can be interpreted and used.

Bayes factors can be seen as the "odds provided by the data for M_j versus M_i" (Berger and Pericchi, 2001). In other words, they show what are the odds that the observations have been generated by model M_j with respect to model M_i. A Bayes factor larger than one, would indicate that it is more likely that model M_j has generated the data than model M_i. And, the larger the value of B_{ji}, the more strong is this "statement". On the contrary, a value of the Bayes factor smaller than one, would indicate as more likely model M_i (with respect to model M_j). Obviously, the closer to zero the value of B_{ji}, the stronger this indication is.

For what in concerns model posterior probabilities $P(M_j|x)$, $j = 1, \dots, k$, they can be used in different ways. It would seem appropriate that, should one of these probabilities be definitely higher in value than the remaining ones, then the associated model has to be chosen. However, especially when the number of models is large, posterior probabilities tend to be all small in value. In this case, (Chipman et al., 2001), show that a decision theory approach to select the appropriate model can be applied. A utility function $u(\alpha, \Delta)$ is chosen, where α is the action of choosing model M_j, and Δ is an unknown quantity of interest, such as a prediction of x. Thus, the model is selected on the basis of the action α maximising the expected utility

$$\mathbb{E}\{u(\alpha, \Delta)\} = \int u(\alpha, \Delta) P(\Delta|x) \, d\Delta,$$

where $P(\Delta|x)$ is the predictive distribution of Δ given x

$$P(\Delta|x) = \sum_{j=1}^{k} P(\Delta|M_j, x) P(M_j|x).$$

$P(\Delta|M_j, x)$ represents the probability of Δ given model M_j and data x. Note that, in this case, the strategy that is used to select a model will depend on the utility function $u(\alpha, \Delta)$ adopted in the process.

If the objective is solely prediction, Bayesian model averaging could represent an appropriate solution (Hoeting et al., 1999). The general idea is to consider the posterior probability of the quantity of interest (given the data), as the average of the posterior probabilities under each model in the model space, weighted by the model posterior probabilities. Thus, we can compute appropriate indexes, such as mean and variance, of the posterior distribution of Δ. That is

$$\mathbb{E}(\Delta|x) = \sum_{j=1}^{k} \mathbb{E}(\Delta|M_j, x) P(M_j|x),$$

and

$$\begin{aligned} Var(\Delta|x) &= \sum_{j=1}^{k} \left[\left\{ Var(\Delta|M_j, x) + \mathbb{E}(\Delta|M_j, x) P(M_j|x)^2 \right\} P(M_j|x) \right] \\ &\quad -\mathbb{E}(\Delta|x)^2. \end{aligned}$$

Sometimes, model averaging is restricted to a subset of the model space. In this case, only models with a relatively high posterior probability are considered in computing the weighted posterior of Δ and its indexes.

There are several reasons why Bayesian model selection has to be preferred to a classical approach. Berger and Pericchi (2001) discuss this in detail. Among the advantages, we have an easiness in interpretation of the Bayes factors with respect to the widely criticised p values. For example, see Sellke et al. (2001).

Bayesian model selection is consistent, in the sense that, if the true model is in the set of all possible models, with a sufficient amount of data, this model will be selected by the procedure. In addition, if the true model is not in the models space, the result in Berk (1966) shows that (asymptotically) the selection process will point to the model which is the closest to the true one, in terms of Kullback–Leibler divergence.

As discussed in Scott and Berger (2010), the Bayesian procedure is an automatic Occam's razor: the selection is always in favour of the simpler model.

Other positive results in adopting a Bayesian model selection approach include: the procedure is the same if the model space has two, three or more elements. Nested models, standard distributions or regular asymptotics, are not required. Model uncertainty is accounted for; thus, it is not necessary to use part of the data for parameter estimation and the remaining for prediction.

On the downside, in the specific when an objective Bayesian model selection approach is considered, the following difficulties have to be considered (Berger and Pericchi, 2001). Computational issues can arise in the calculation of Bayes factors when parameter spaces are large. Similarly, difficulties are encountered for selection problems where the number of models is considerably large.

Use of improper priors for the parameters of the models is not possible, in general. Given that most objective priors are improper, this leads to strong challenges for an objective approach. Even the use of "arbitrary" vague priors is not advisable, as the results will strongly depend on the level of "arbitrariness" chosen for the prior.

Finally, even though some models may have parameters in common, their meaning may be different. Thus, prior distributions for these parameters have to reflect the difference as well.

6.2 Current objective approaches

Except for variable selection problems (Berger and Pericchi, 2001; Scott and Berger, 2010), it appears that the main effort in determining objective prior probabilities is concentrated on $\pi_j(\theta_j)$. Comprehensive discussions on the various approaches in determining the prior distribution for the parameters of the models can be found in Berger and Pericchi (2001), Chipman et al. (2001), Pérez and Berger (2002), Stracham and van Dijk (2003), and the references included in the papers. On the other hand, very little discussion has been given to the prior mass to be put of the model space, and the usual objective prior is the uniform one; that is, $P(M_j) = 1/k$, for $j = 1, \ldots, k$. In other words, the claimed objective approach assigns equal importance to each model in the set of all the possible models.

6.3 The Villa-Walker objective model prior

As anticipated, we obtain the model prior on the basis of the criterion discussed in Chapter 3, where we replace the parameter θ with the model M. We then objectively assign a *worth* to each model via Berk's result (Theorem 3.1), and link it to the prior mass via the *self-information* loss function. We need to take into consideration that the Kullback–Leibler divergence is minimised in expectation, where the expectation is with respect to the prior on the model parameters.

We introduce the idea in a simple model selection problem with two possible models only. Let us assume that we have to select between models

$$M_1 = \{f_1(x|\theta_1), \pi_1(\theta_1)\} \quad \text{and} \quad M_2 = \{f_2(x|\theta_2), \pi_2(\theta_2)\},$$

where we assume that the prior of the parameter $\theta_1 \in \Theta_1$, $\pi_1(\theta_1)$, and the prior on the parameter $\theta_2 \in \Theta_2$, $\pi_2(\theta_2)$, are known and proper. Following the criterion, the prior mass on M_1, $P(M_1)$, is determined on the basis on what is lost if model M_1 is removed, and it is the true one. $P(M_1)$ is then proportional to the expected minimum loss between the models. Hence,

$$P(M_1) \propto \exp\left\{\int_{\Theta_1} \min_{\theta_2} D_{KL}\Big(f_1(x|\theta_1)\|f_2(x|\theta_2)\Big)\pi_1(\theta_1)\, d\theta_1\right\}. \tag{6.1}$$

Similarly, the mass associated to M_2, $P(M_2)$, is proportional to the expected minimum loss between model M_2 and model M_1, given by

$$P(M_2) \propto \exp\left\{\int_{\Theta_2} \min_{\theta_1} D_{KL}\Big(f_2(x|\theta_2)\|f_1(x|\theta_1)\Big)\pi_2(\theta_2)\, d\theta_2\right\}. \qquad (6.2)$$

The expressions at the exponential in (6.1) and in (6.2), can also be written, respectively, as

$$\begin{aligned} P(M_1) &\propto \exp\left\{\mathbb{E}\left[\min_{\theta_2} D_{KL}\Big(f_1(x|\theta_1)\|f_2(x|\theta_2)\Big)\right]\right\} \quad \text{and} \\ P(M_2) &\propto \exp\left\{\mathbb{E}\left[\min_{\theta_1} D_{KL}\Big(f_2(x|\theta_2)\|f_1(x|\theta_1)\Big)\right]\right\}. \end{aligned}$$

where the expectations are taken with respect to the prior distributions.

The most general scenario is represented by a model space of k elements, where each model is specified by a vector of parameters of finite dimension. Let us consider a model selection problem with model space $\mathcal{M} = \{M_1, \ldots, M_k\}$, with $M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\}$, $j = 1, \ldots, k$. A compact notation for the prior mass for model M_j is then given by

$$P(M_j) \propto \exp\left\{\mathbb{E}\left[\min_{m\neq j} D_{KL}\Big(f_j(x|\theta_j)\|f_m(x|\theta_m)\Big)\right]\right\}, \qquad m, j = 1, \ldots, k,$$

In other words, the prior assigned to model M_j can be seen as if it is obtained by measuring the divergence between $f_j(x|\theta_j)$ and any other model, and selecting the smaller one.

In the following sections we discuss some illustrations for the non-nested and the nested model selection case. To simplify the notation, unless otherwise specified, the numbering of the various models (including the probability and prior distribution that form them) starts afresh in each illustration.

6.4 Non-nested models

In this section we consider scenarios where the elements of the model space are non-nested. In the first illustration we compare two discrete models; a Poisson and a Geometric probability mass function. Next, we consider a model selection problem with two multiparameter continuous densities: Weibull and Log-normal. Finally, in the third illustration, we extend the continuous problem to a three model selection problem by adding a Gamma density.

6.4.1 Illustration 1: Poisson and Geometric models

Let us assume that we have observed a set of observations x from a phenomenon we know to have support $\mathcal{X} = \{0, 1, 2, \ldots\}$. We want to compare the following two models

$$M_1 = \left\{f_1(x|\theta) = \theta^x e^{-\theta}/x!\,, \pi_1(\theta)\right\} \quad \text{and} \quad M_2 = \{f_2(x|\phi) = \phi(1-\phi)^x, \pi_2(\phi)\}\,,$$

that is, M_1 is a Poisson distribution with rate parameter $\theta \in (0, +\infty)$, and M_2 is a Geometric distribution with probability of success $\phi \in (0, 1)$.

Following the objective approach we have outlined in Section 6.3, we first consider the mass to be assigned to model M_1. By applying (6.1) we have

$$P(M_1) \propto \exp\left\{\int \min_{\phi} D_{KL}\Big(f_1(x|\theta)\|f_2(x|\phi)\Big)\pi_1(\theta)\, d\theta\right\}. \tag{6.3}$$

To determine the mass in (6.3), we first find the Kullback–Leibler divergence between a Poisson distribution with parameter θ and a Geometric distribution with parameter ϕ. As shown by Theorem B.1 in Appendix B, this is given by

$$\begin{aligned} D_{KL}(f_1(x|\theta)\|f_2(x|\phi)) &= \sum_{x=0}^{\infty}\left[\frac{\theta^x}{x!}e^{-\theta}\log\left\{\frac{e^{-\theta}\theta^x/x!}{\phi(1-\phi)^x}\right\}\right] \\ &= \theta\log\theta - \sum_{x=0}^{\infty}\left(\log x!\,\frac{\theta^x}{x!}e^{-\theta}\right) \\ &- \theta - \log\phi - \theta\log(1-\phi). \end{aligned} \tag{6.4}$$

The divergence (6.4) is minimised, with respect to ϕ, by $\phi = 1/(1+\theta)$. By replacing this result into (6.4), we obtain the minimum Kullback–Leibler divergence between a Poisson and a Geometric distributions

$$\min_{\phi} D_{KL}(f_1(x|\theta)\|f_2(x|\phi)) = -\theta + \theta\log(1+\theta) + \log(1+\theta) - \sum_{x=0}^{\infty}\left(\log x!\,\frac{\theta^x}{x!}e^{-\theta}\right).$$

For this illustration, we have considered a Gamma prior on the parameter θ, with shape and scale parameter both equal to one; that is, $\pi_1(\theta) \sim Ga(1,1) = \exp(-\theta)$. Therefore

$$\begin{aligned} P(M_1) &\propto \exp\left\{\int \min_{\phi} D_{KL}\Big(f_1(x|\theta)\|f_2(x|\phi)\Big)e^{-\theta}d\theta\right\} \\ &= \exp(0.09) = 1.09. \end{aligned} \tag{6.5}$$

The result in (6.5) is obviously affected by the choice of the prior. In particular, we note that if the variance of $\pi_1(\theta)$ increases, corresponding to an increase of uncertainty about the true value of the parameter, the mass assigned to model M_1 increases. For example, if we chose the prior to be $\pi_1(\theta) \sim Ga(10,1)$ (corresponding to a variance of 10), the corresponding mass on M_1 would be $P(M_1) \propto 2.16$. Similarly, if the variance decreases, therefore the uncertainty about the parameter is more limited, the approach will assign a lower mass. For example, for $\pi_1(\theta) \sim Ga(1,5)$ (variance equal to 0.04), we have $P(M_1) \propto 1.01$. Intuitively, if we have a relatively high uncertainty about the true value of the parameter, the loss (in expectation) we would incur in choosing the wrong model would be relatively large. Hence, the model assumes more importance in the overall scenario. Vice versa, if our prior knowledge about the true value of the parameter is relatively precise (i.e. low uncertainty), the loss of information in choosing the wrong model would be (in expectation) relatively low.

With a similar procedure, by applying (6.2) we obtain the mass for model M_2. In fact, the Kullback–Leibler divergence between a Geometric distribution and a Poisson distribution is given by

$$D_{KL}(f_2(x|\phi)\|f_1(x|\theta)) = \sum_{x=0}^{\infty}\left[\phi(1-\phi)^x\log\left\{\frac{\phi(1-\phi)^x}{e^{-\theta}\theta^x/x!}\right\}\right]$$

$$
\begin{aligned}
&= \log\phi + \frac{1-\phi}{\phi}\log(1-\phi) - \frac{1-\phi}{\phi}\log\theta + \theta \\
&\quad + \sum_{x=0}^{\infty}\Big\{\phi(1-\phi)^x \log x!\Big\}, \qquad (6.6)
\end{aligned}
$$

which is minimised by $\theta = (1-\phi)/\phi$ (refer to Theorem B.1 in Appendix B). We replace this result in (6.6), and obtain

$$
\min_{\theta} D_{KL}(f_2(x|\phi)\|f_1(x|\theta)) = \log\phi + \frac{1-\phi}{\phi}\log\phi + \frac{1-\phi}{\phi} + \sum_{x=0}^{\infty}\Big\{\phi(1-\phi)^x \log x!\Big\}.
$$

The prior for the parameter ϕ has been selected to be a Beta distribution with both shape parameter values equal to two. That is, $\pi_2(\phi) \sim Be(2,2) \propto \phi(1-\phi)$. Thus, the mass to be put on model M_2 is determined to be

$$
\begin{aligned}
P(M_2) &\propto \exp\left\{\int \min_{\theta} D_{KL}\Big(f_2(x|\phi)\|f_1(x|\theta)\Big)\phi(1-\phi)\, d\phi\right\} \\
&= \exp(0.47) = 1.60. \qquad (6.7)
\end{aligned}
$$

Also in the computation of $P(M_2)$ we see that the prior mass assigned to the model depends on the variance of the prior distribution for ϕ. In particular, similarly to the computation of $P(M_1)$, the larger the variance the more the mass, and vice versa.

Results in (6.5) and (6.7) can be normalised. The resulting prior distribution for this model selection problem (i.e. given the chosen models and the prior distributions of the respective parameters), is $P_N(M_1) = 0.41$ and $P_N(M_2) = 0.59$. It is not possible to perform a direct comparison between the variances of the two prior distributions, $\pi_1(\theta)$ and $\pi_2(\phi)$. However, it is plausible to assume that there is always the possibility to chose them in a way that the prior masses on the models are equal. In fact, if we consider as prior distribution for θ a Gamma with shape parameter 5 and rate parameter 1, and as prior for ϕ a Beta with both parameters equal to two, we obtain $P(M_1) \propto 1.59$ and $P(M_2) \propto 1.60$. Normalising, we have the uniform prior of the models given by $P_N(M_1) = 0.50$ and $P_N(M_2) = 0.50$. Under these circumstances, we can assume that the level of uncertainty about θ

and ϕ is virtually the same.

It is also interesting to examine what happens when the uncertainty about the parameter value of one model is much larger than the uncertainty on the parameter of the other model. For example, let us keep the prior on ϕ fixed, that is $\pi_2(\phi) \sim Be(2,2)$, and set $\pi_1(\theta) \sim Ga(20, 1/2)$. In this case, the variance of $\pi_1(\theta)$ is equal to 80, which is a much larger value than the case where $\pi_1(\theta) \sim Ga(1,1)$. Thus, we have that $P(M_1) \propto \exp(1.43) = 4.17$. Normalising, $P_N(M_1) = 0.72$ and $P_N(M_2) = 0.28$.

6.4.2 Illustration 2: Weibull and Log-normal models

In this illustration we consider a scenario where the quantity of interest x has a continuous support $\mathcal{X} = (0, +\infty)$. We also show how the approach can be applied to models with dimension of the parameter space larger than one. We consider model M_1 to be a Weibull density with scale parameter $\lambda > 0$ and shape parameter $\kappa > 0$. Model M_2 is a Log-normal density with location parameter $\mu \in \mathbb{R}$ (in the log-scale), and shape parameter σ^2. These distributions are often considered as option to model data, for example, in survival analysis studies (Klein and Moeschberger, 1997). Note that we will consider the parametrisation expressed with the precision parameter $\tau = 1/\sigma^2 > 0$. Therefore

$$M_1 = \left\{ f_1(x|\lambda,\kappa) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right], \pi_1(\lambda,\kappa) \right\},$$

$$M_2 = \left\{ f_2(x|\mu,\tau) = \frac{1}{x}\left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left[-\frac{1}{2}\tau(\log x - \mu)^2\right], \pi_2(\mu,\tau) \right\}.$$

On the basis of our approach, the prior mass to be assigned to model M_1 and model M_2 is determined, respectively, by

$$P(M_1) \propto \exp\left\{ \int\int \min_{\mu,\tau} D_{KL}\Big(f_1(x|\lambda,\kappa) \| f_2(x|\mu,\tau)\Big) \pi_1(\lambda,\kappa)\, d\lambda d\kappa \right\}, \qquad (6.8)$$

and

$$P(M_2) \propto \exp\left\{ \int\int \min_{\lambda,\kappa} D_{KL}\Big(f_2(x|\mu,\tau) \| f_1(x|\lambda,\kappa)\Big) \pi_2(\mu,\tau)\, d\mu d\tau \right\}.$$

To compute the mass for model M_1, we first obtain the Kullback–Leibler divergence between a Weibull density and a Log-normal density, as shown in Theorem B.2 in Appendix B.

$$\begin{aligned}
D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau)) &= \int_0^\infty f_1(x|\lambda,\kappa)\log\left\{\frac{f_1(x|\lambda,\kappa)}{f_2(x|\mu,\tau)}\right\}dx\\
&= \log\kappa + \kappa\,\mathbb{E}(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa) - \frac{1}{2}\log\tau + \frac{1}{2}\log(2\pi)\\
&+ \frac{1}{2}\tau\,\mathbb{E}(\log^2 x) - \tau\mu\,\mathbb{E}(\log x) + \frac{1}{2}\tau\mu^2, \qquad (6.9)
\end{aligned}$$

where the expectations are with respect to $f_1(x|\lambda,\kappa)$, with $\mathbb{E}(\log x) = \log\lambda - \gamma/\kappa$ ($\gamma \approx 0.5772$ is the Euler's constant), $\mathbb{E}(x^\kappa) = \lambda^\kappa$, and $\mathbb{E}(\log^2 x) = \pi^2/(6\kappa^2) + (\log\lambda - \gamma/\kappa)^2$ ($\pi^2/(6\kappa^2)$ is the variance of the logarithm of x, that is $Var(\log x) = \pi^2/(6\kappa^2)$). The minimum of the divergence in (6.9), with respect to parameters μ and τ, is attained at $\mu = \mathbb{E}(\log x) = \log\lambda - \gamma/\kappa$ and $\tau = 1/Var(\log x) = 6\kappa^2/\pi^2$. Recalling that, if random variable x is log-normally distributed with parameters μ and τ, then random variable $y = \log x$ has a normal distribution with mean μ and precision τ, we see that the minimum divergence between a Weibull and a Log-normal is attained when, in the log-scale, both densities have the same mean and variance. And this is a sensible result. Thus, by replacing the expressions of the expectations of the functions of x into equation (6.9), we have

$$\begin{aligned}
\min_{\mu,\tau} D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau)) &= \log\kappa + \kappa\,\mathbb{E}(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa)\\
&\quad + \frac{1}{2}\log\{Var(\log x)\} + \frac{1}{2}\log(2\pi) + \frac{1}{2}\\
&= \frac{1}{2}\log(2\pi) + \log\pi - \gamma - \frac{1}{2}\log 6 - \frac{1}{2}.
\end{aligned}$$

We note that the minimum divergence, with respect to μ and τ, from a Weibull density to a Log-normal density, does not depend on the values of parameters λ and κ, and it has value 0.09. An important aspect of this result is that the mass to be assigned to model M_1 does not depend on the choice of the priors for λ and κ. By applying (6.8), the prior mass for the Weibull density is $P(M_1) \propto \exp(0.09) = 1.09$.

With an analogous approach, we compute the value of $P(M_2)$. The Kullback–

Leibler divergence from a Log-normal density with parameters μ and τ, and a Weibull density with parameters λ and κ (refer to Theorem B.2 in Appendix B) is given by

$$\begin{aligned} D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa)) &= \int_0^\infty f_2(x|\mu,\tau)\log\left\{\frac{f_2(x|\mu,\tau)}{f_1(x|\lambda,\kappa)}\right\}dx \\ &= \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\,\mathbb{E}(\log^2 x) + \tau\mu\,\mathbb{E}(\log x) - \\ &\quad \frac{1}{2}\tau\mu^2 - \log\kappa - \kappa\,\mathbb{E}(\log x) + \kappa\log\lambda + \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa), \quad (6.10) \end{aligned}$$

where in this case the expectations are with respect to the Log-normal density. In particular, $\mathbb{E}(\log x) = \mu$, $\mathbb{E}(x^\kappa) = \exp\{\kappa^2/(2\tau) + \mu\kappa\}$ and $\mathbb{E}(\log^2 x) = 1/\tau + \mu^2$. The divergence in (6.10) has minimum for $\lambda = \exp\{1/(2\sqrt{\tau}) + \mu\}$ and $\kappa = \sqrt{\tau}$, giving

$$\begin{aligned} \min_{\lambda,\kappa} D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa)) &= \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\left(\frac{1}{\tau} + \mu^2\right) + \tau\mu^2 \\ &- \frac{1}{2}\tau\mu^2 - \frac{1}{2}\log\tau - \sqrt{\tau}\mu + \sqrt{\tau}\left(\frac{1}{2\sqrt{\tau}} + \mu\right) + 1 \\ &= 1 - \frac{1}{2}\log(2\pi). \end{aligned}$$

Again, we note that the minimum divergence between the models is a constant, and its value is of 0.08. As such, the choice of $\pi_2(\mu,\tau)$ does not have impact on the prior mass that, in accordance to our approach, is assigned to model M_2. We then compute this mass as $P(M_2) \propto \exp(0.08) = 1.08$.

By normalising, we have that $P_N(M_1) = 0.50$ and $P_N(M_2) = 0.50$, which is uniform and, in this case, traces back to the common objective approach to assign equal prior probability to two models.

The result deriving from Theorem B.2 is easy to derive and it is discussed, for example, in Dumonceaux and Antle (1973) and Dumonceaux et al. (1973). In essence, the Kullback–Leibler divergence between two models with location and scale parameters, when minimised with respect to the parameters of either model, is independent of the parameter values from the other model. In the light of our approach, this means that the choice of the prior distribution for the parameters has no influence on the value of prior mass assigned to each model. Furthermore, for the Weibull and Log-normal models, the Kullback–Leibler divergences are very similar, resulting in a uniform model prior.

6.4.3 Illustration 3: Weibull, Log-normal and Gamma models

The approach we propose can be applied to model spaces with a number of elements as large as necessary. To illustrate this, we consider the case where, in addition to the two models introduced in Section 6.4.2, we add a third one. In particular, a Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. This distribution as well, is considered as an option to model survival analysis data (Klein and Moeschberger, 1997). The model space is then formed by the following three models

$$
\begin{aligned}
M_1 &= \left\{ f_1(x|\lambda,\kappa) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right], \pi_1(\lambda,\kappa) \right\}, \\
M_2 &= \left\{ f_2(x|\mu,\tau) = \frac{1}{x}\left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left[-\frac{1}{2}\tau(\log x-\mu)^2\right], \pi_2(\mu,\tau) \right\}, \\
M_3 &= \left\{ f_3(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x), \pi_3(\alpha,\beta) \right\}.
\end{aligned}
$$

Given that our approach assigns mass on a model on the basis of what it is lost if the model is removed from the model space and it is the true model, we have to identify, for each model M_j, $j = 1, 2, 3$, the model M_i, $j \neq i$ that is nearer (in terms of the expected Kullback–Leibler divergence).

Let us first consider the Weibull model M_1. The $\log P(M_1)$ is proportional to the minimum value between

$$
\left\{ \mathbb{E}\left[\min_{\mu,\tau} D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau))\right], \mathbb{E}\left[\min_{\alpha,\beta} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta))\right] \right\}, \tag{6.11}
$$

where the expectation are taken with respect to the prior $\pi_1(\lambda,\kappa)$. From Section 6.4.2, we know that the value of the first element in (6.11) is 0.09, as the minimum divergence from a Weibull density to a Log-normal density does not depend on

$\pi_1(\lambda, \kappa)$. To compute the expected minimum divergence from model M_2 to model M_3, we proceed as seen in Section 6.4.2. First, we determine the Kullback–Leibler divergence from M_1 to M_3, as shown in Theorem B.3 in Appendix B, which gives

$$\begin{aligned} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta)) &= \int_0^\infty f_1(x|\lambda,\kappa)\log\left\{\frac{f_1(x|\lambda,\kappa)}{f_3(x|\alpha,\beta)}\right\} dx \\ &= \log\kappa + \kappa\,\mathbb{E}(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa) - \alpha\log\beta \\ &+ \log\Gamma(\alpha) - \alpha\,\mathbb{E}(\log x) + \beta\,\mathbb{E}(x). \end{aligned} \tag{6.12}$$

The minimum of (6.12), with respect to the parameter α and β of the Gamma density, is found by solving the following system of equations

$$\begin{cases} \mathbb{E}(\log x) = \Psi(\alpha) - \log\beta \\ \mathbb{E}(x) = \alpha/\beta, \end{cases} \tag{6.13}$$

from which we see that the two densities are nearer when they have equal expectation for x and $\log x$ (refer to Theorem B.3 in Appendix B). In fact, if a random variable has Gamma distribution with shape parameter α and rate parameter β, its expectation is α/β and the expectation of its logarithm is $\Psi(\alpha) - \log\beta$; where $\Psi(\alpha) = d\{\log\Gamma(\alpha)\}/d\alpha$ is the digamma function. System (6.13) is solved with numerical methods, and the minimum divergence between a Weibull and a Gamma has the form

$$\begin{aligned} \min_{\alpha,\beta} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta)) &= \log\kappa - \gamma - 1 - \alpha\log\beta + \log\Gamma(\alpha) \\ &- \alpha\log\lambda + \alpha\frac{\gamma}{\kappa} + \beta\lambda\Gamma\left(1+\frac{1}{\kappa}\right), \end{aligned}$$

where we have considered that, if x has a Weibull distribution with parameters λ and κ, then $\mathbb{E}(x) = \Gamma(1/\kappa)\lambda/\kappa$. In this illustration we assume that the parameters of the Weibull are independent. Therefore, the prior $\pi_1(\lambda, \kappa)$ is the product of the marginal prior assigned on each parameter, which have been chosen to be identical and, in particular, Gamma distributed with shape parameter equal to 25 and rate parameter equal to 1. That is, distributions with relatively large variance.

With this prior, we have obtained $\mathbb{E}\left\{\min_{\alpha,\beta} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta))\right\} = 0.05$. Thus, as this result gives a smaller expected divergence in comparison to the one measured to the Log-normal (as computed in Section 6.4.2), the mass to be assigned to model M_1 is $P(M_1) \propto \exp(0.05) = 1.06$.

It is legitimate to wonder if it is possible, by selecting a different prior $\pi_1(\lambda,\kappa)$, to define a Weibull density which is nearer to the Log-normal than to the Gamma. For example, if we chose the Gamma distributions for λ and κ with the rate parameter equal to 2, the expected minimum divergence would have value 0.14, and the prior mass for M_1 would be based on the expected minimum divergence with respect to the Log-normal density.

To determine the prior probability for model M_2, we need to identify the minimum between

$$\left\{\mathbb{E}\left[\min_{\lambda,\kappa} D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa))\right],\ \mathbb{E}\left[\min_{\alpha,\beta} D_{KL}(f_2(x|\mu,\tau)\|f_3(x|\alpha,\beta))\right]\right\}.$$

In Section 6.4.2 we have shown that the first term does not depend on the parameters of the Log-normal ad has value 0.08. The Kullback–Leibler divergence between M_2 and M_3 is (refer to Theorem B.4 in Appendix B)

$$\begin{aligned} D_{KL}(f_2(x|\mu,\tau)\|f_3(x|\alpha,\beta)) &= \int_0^\infty f_2(x|\mu,\tau)\log\left\{\frac{f_2(x|\mu,\tau)}{f_3(x|\alpha,\beta)}\right\}dx \\ &= \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\,\mathbb{E}(\log^2 x) + \tau\mu\,\mathbb{E}(\log x) \\ &\quad - \frac{1}{2}\tau\mu^2 - \alpha\log\beta + \log\Gamma(\alpha) - \alpha\,\mathbb{E}(\log x) + \beta\,\mathbb{E}(x). \end{aligned} \tag{6.14}$$

The minimum of (6.14), with respect to α and β, is attained when simultaneously $\mathbb{E}(x) = \alpha/\beta$ and $\mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$; that is, when the two densities have equal mean and equal expectation of the logarithm of x. Note that this result is analogous to the one obtained when we have determined the minimum divergence from M_1 to M_3. To compute the expected minimum Kullback–Leibler divergence between the Log-normal density and the Gamma density, for coherence, we have again assumed the parameters as independent: $\pi_\mu(\mu) \sim Ln(0, 0.1)$ and $\pi_\tau(\tau) \sim Ga(25, 1)$. We have obtained $\mathbb{E}\left\{\min_{\alpha,\beta} D_{KL}(f_2(x|\mu,\tau)\|f_3(x|\alpha,\beta))\right\} = 0.06$. With this prior, the mass for model M_2 is determined on the basis of its distance to the Gamma density, and it is $P(M_2) \propto \exp(0.06) = 1.06$.

We note that, by increasing the uncertainty around the parameters, this mass increases as well. For example, by setting the rate parameter of the prior for τ to 1/4, we would have an expected minimum divergence of 0.09. In this case, the prior probability for the Log-normal would be based on the distance with respect to the Weibull.

For the prior probability of model M_3, we need to compare

$$\left\{\mathbb{E}\left[\min_{\lambda,\kappa} D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa))\right], \mathbb{E}\left[\min_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau))\right]\right\}.$$

First, we see that the divergence from model M_3 to model M_1 is given by

$$\begin{aligned} D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa)) &= \int_0^\infty f_3(x|\alpha,\beta)\log\left\{\frac{f_3(x|\alpha,\beta)}{f_1(x|\lambda,\kappa)}\right\} dx \\ &= \alpha\log\beta - \log\Gamma(\alpha) + \alpha\,\mathbb{E}(\log x) - \beta\,\mathbb{E}(x) - \log\kappa \\ &- \kappa\,\mathbb{E}(\log x) + \kappa\log\lambda + \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa), \end{aligned} \qquad (6.15)$$

where $\mathbb{E}(x) = \alpha/\beta$, $\mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$ and $\mathbb{E}(x^\kappa) = \beta^{-\kappa}\Gamma(\kappa+\alpha)/\Gamma(\alpha)$ (refer to Theorem B.2 in Appendix B). The minimum of (6.15), with respect to λ and κ, is found by solving

$$\begin{cases} \mathbb{E}(x^\kappa) = \lambda^\kappa \\ \Psi(\kappa+\alpha) - 1/\kappa = \Psi(\alpha). \end{cases} \qquad (6.16)$$

Solving system (6.16), with numerical methods, the minimum Kullback–Leibler divergence between the Gamma density and the Weibull density has the following expression

$$\begin{aligned} \min_{\lambda,\kappa} D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa)) &= -\log\Gamma(\alpha) + \alpha\Psi(\alpha) - \alpha - \log\kappa - \kappa\Psi(\alpha) \\ &+ \kappa\log\beta + \kappa\log\lambda + 1. \end{aligned}$$

Assuming α and β independent, prior $\pi_3(\alpha,\beta)$ can be set as the product of two

Gamma distributions. For coherence with previous decisions, we have chosen both Gamma with shape parameter equal to 25, in order to have relatively high variance, thus relatively high uncertainty about the parameter values. The expected minimum divergence is $\mathbb{E}\left\{\min_{\lambda,\kappa} D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa))\right\} = 0.02$.

To assess $\mathbb{E}\left\{\min_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau))\right\}$, we consider the Kullback–Leibler divergence between the two models (refer to Theorem B.4 in Appendix B)

$$\begin{aligned}
D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau)) &= \int_0^\infty f_3(x|\alpha,\beta)\log\left\{\frac{f_3(x|\alpha,\beta)}{f_2(x|\mu,\tau)}\right\}\,dx \\
&= \alpha\log\beta - \log\Gamma(\alpha) + \alpha\,\mathbb{E}(\log x) - \beta\,\mathbb{E}(x) - \frac{1}{2}\log\tau \\
&\quad + \frac{1}{2}\log(2\pi) + \frac{1}{2}\tau\,\mathbb{E}(\log^2 x) - \tau\mu\,\mathbb{E}(\log x) + \frac{1}{2}\tau\mu^2. \quad (6.17)
\end{aligned}$$

The divergence in (6.17) is minimised with respect to μ and τ when, simultaneously, $\mu = \mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$ and $\tau = 1/Var(\log x) = 1/\Psi'(\alpha)$, with $\Psi'(\alpha) = d\left\{\Psi(\alpha)\right\}/d\alpha$ the trigamma function. We note that the two models are at their nearest distance when expectation and variance (of the logarithm) are equal. The expression of this minimum divergence is

$$\begin{aligned}
\min_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau)) &= -\log\Gamma(\alpha) + \alpha\Psi(\alpha) - \alpha + \frac{1}{2}\log\Psi'(\alpha) \\
&\quad + \frac{1}{2}\log 2\pi + \frac{1}{2}.
\end{aligned}$$

We used the same prior we have used to compute the expected minimum divergence between M_3 and M_1. The result is $\mathbb{E}\left\{\min_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau))\right\} = 0.06$. Therefore, the prior mass for M_3 is based on the "distance" from the Gamma to the Weibull, and has value $P(M_3) \propto \exp(0.02) = 1.02$. The expected minimum divergence depends only on the value of the shape parameter.

Unlike for the previous two comparisons, when we have to assign a mass to M_3, it appears impossible to define prior distributions such that we can invert the relationship between the two expected minimum divergences. In fact, both of them depend at least on one parameter and, unless there are specific (and strong) reasons to justify a different level of uncertainty in the two cases, we should adopt

	M_1	M_2	M_3
M_1		0.08	0.06
M_2	0.09		0.02
M_3	0.05	0.03	
$P(M_j)$	1.06	1.03	1.02
$P_N(M_j)$	0.34	0.33	0.33

Table 6.1: Expected minimum Kullback-Leibler divergence (by column) among the models M_1 (Weibull), M_2 (Log-normal) and M_3 (Gamma). The divergences have been computed on the basis of the prior distributions on the parameters of the models as specified in Section 6.4.3. The mass is proportional to the exponential of the minimum divergence, and the last two rows show this mass for each model: non-normalised $P(M_j)$ and normalised $P_N(M_j)$, $j = 1, 2, 3$.

the same prior.

Table 6.1 summarises the expected minimum divergences among models M_1, M_2 and M_3 and, as previously computed, the appropriate prior mass. The normalised prior for this particular model selection problem, and given the selected priors for the parameter of the models, are $P_N(M_1) = 0.34$, $P_N(M_2) = 0.33$ and $P_N(M_3) = 0.33$. Even though is not possible to make a direct comparison among the level of uncertainty that we have expressed for the parameters of each model (via the appropriate prior distributions), we note that, by keeping variances relatively large, the model prior is practically uniform.

6.5 Nested models

Let us now consider the case where models are nested. The simplest scenario is when we have only two models, where we can identify an inner (or simple) model and an outer (or complex) model. Logic dictates that, if we select the outer model when is the inner one the true one, there would be no loss (in terms of information), for the inner is a special case of the outer. Therefore, the prior mass to be assigned to the inner model will be proportional to one. The mass for the outer model will be determined with a procedure analogous to the one we have repeatedly examined in Section 6.4. Consider the following example.

Example 6.1. *Let us assume that we want to select between a standard normal density and a normal density with the same precision but with the mean that is allowed to be different from zero; that is*

$$M_1 = \left\{ f(x|0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)\right\} \quad \textit{and}$$
$$M_2 = \left\{ f(x|\mu,1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2\right], \pi(\mu)\right\}.$$

The general expression of the Kullback–Leibler divergence between two normal densities with different means and precisions, say $f(x|\mu_1,\tau_1) = N(\mu_1,\tau_1)$ *and* $f(x|\mu_2,\tau_2) = N(\mu_2,\tau_2)$, *is given by*

$$\begin{aligned} D_{KL}(f(x|\mu_1,\tau_1)\|f(x|\mu_2,\tau_2)) &= \int_{-\infty}^{\infty} f(x|\mu_1,\tau_1)\log\left\{\frac{f(x|\mu_1,\tau_1)}{f(x|\mu_2,\tau_2)}\right\} dx \\ &= \frac{\tau_2}{2}(\mu_1-\mu_2)^2 + \frac{1}{2}\left(\frac{\tau_2}{\tau_1} - 1 - \log\frac{\tau_2}{\tau_1}\right) \quad (6.18)\end{aligned}$$

To assign a mass to M_1, *we have to find the minimum of* $D_{KL}(f(x|0,1)\|f(x|\mu,1))$ *which, considering* (6.18), *is attained for* $\mu = 0$, *resulting in a divergence equal to zero. As such,* $P(M_1) \propto 1$. *For* M_2, *we note that* $D_{KL}(f(x|\mu,1)\|f(x|0,1)) = \mu^2/2$, *which is also the minimum, given* μ. *Therefore, the minimum expected divergence, with respect to the prior* $\pi(\mu)$, *is given by*

$$\begin{aligned}\int D_{KL}\Big(f(x|\mu,1)\|f(x|0,1)\Big)\pi(\mu)\,d\mu &= \int \frac{\mu^2}{2}\pi(\mu)\,d\mu \\ &\propto \mathbb{E}(\mu^2).\end{aligned}$$

Thus, taking $\mathbb{E}(\mu) = 0$, *we have* $\mathbb{E}(\mu^2) = Var(\mu)$. *So* $P(M_2) \propto \exp\{Var(\mu)\}$. *The result is coherent with what is expected. First, we note that the mass associated with the simpler model is proportional to one. Second, the mass on the more complex model is related to the variance of the prior distribution for* μ. *If* $Var(\mu) = 0$ *(i.e. we put a point mass at* $\mu = 0$*), we have that* $P(M_1) = P(M_2) = 1/2$, *as it should be. On the other hand, if* $Var(\mu) \to \infty$, $P(M_2)$ *increases, as we believe more and more that model* M_1 *is wrong. In particular, the larger the variance* *(i.e. the more uncertainty about the parameter we have), the larger the mass associated to the larger model. Furthermore, our approach allows us to avoid the so called Jeffreys-Lindley paradox (Lindley, 1957), by assigning a model prior that depends on the model, namely $\{f(x|\mu,1),\pi(\mu)\}$. In fact, the paradox arises when $P(M_1)=P(M_2)=1/2$, and the uncertainty on μ is not zero. For detailed discussions on the paradox see, for example, Shafer (1982) and Bernardo (1999).*

To generalise, let us assume that we have to select between the following two nested models

$$M_1=\{f(\cdot|\theta_1),\pi_1(\theta_1)\} \quad \text{and} \quad M_2=\{f(\cdot|\theta_1,\theta_2),\pi_2(\theta_2|\theta_1)\pi_1(\theta_1)\},$$

with $\theta_1\in\Theta_1$, $\theta_2\in\Theta_2$, and where the prior distributions for the parameters are supposed to be known. The fact that model M_1 is nested into model M_2, implies that $D_{KL}(f(\cdot|\theta_1)\|f(\cdot|\theta_1,\theta_2))$ is minimised, with respect to the pair (θ_1,θ_2), when θ_2 is removed and M_2 degenerates into M_1. As such, $P(M_1)\propto 1$.

The prior mass to be put on model M_2, according to our approach, will be found in the following way. First, we note that it is not necessary to identify the minimum Kullback–Leibler divergence from model M_2 to model M_1, as parameter θ_1 would have the same value for both models. Hence, if we indicate by ϕ the parameter in model M_1, in order to distinguish it from θ_1 in model M_2, we have that

$$\min_{\phi} D_{KL}\left(f(\cdot|\theta_1,\theta_2)\|f(\cdot|\phi)\right)=D_{KL}(f(\cdot|\theta_1,\theta_2)\|f(\cdot|\theta_1)),$$

that is, when $\phi=\theta_1$. Thus, the mass to be assigned to M_2 is given by

$$P(M_2)\propto\exp\left\{\int\int D_{KL}(f(\cdot|\theta_1,\theta_2)\|f(\cdot|\theta_1))\pi_2(\theta_2|\theta_1)\pi_1(\theta_1)\,d\theta_2d\theta_1\right\}.$$

This result can be further generalised if we consider a set of models nested one into each other. In this case the mass assigned to each model, except for the largest one (i.e. the most complex), will be proportional to one. Furthermore, the only mass that has to be actually computed is the one to be assigned to the largest model.

From this result, we note that when a model is nested into another one, say

M_1 nested in M_2, the prior mass on the simpler model will never be larger than the prior mass on the more complex model. That is, $P(M_2) \geq P(M_1)$. The complex model expresses a more detailed representation of the phenomenon than the simple model (i.e. it caries more information). Therefore, in general, it has to be $P(M_2) > P(M_1)$, and we would have $P(M_2) = P(M_1) = 1/2$ if and only if M_1 and M_2 are the same model.

Let us now see an example on how the general approach is applied to the selection of two nested models of the same family. This example is a generalisation the previous one.

Example 6.2. *Let us assume that we are interested in selecting between a normal density with mean μ and precision one, and a normal density with the same mean parameter and precision τ. The models are*

$$M_1 = \left\{ f(x|\mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2\right], \pi_1(\mu) \right\},$$
$$M_2 = \left\{ f(x|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau}{2}(x-\mu)^2\right], \pi_2(\mu, \tau) \right\}.$$

Applying (6.18), *we have that $D_{KL}(f(x|\mu,1)\|f(x|\mu,\tau)) = \tau(\mu-\theta)^2/2 + (\tau - 1 - \log\tau)/2$, where the mean in M_2 has been rewritten as θ in order to distinguish it from the mean of model M_1. By differentiating with respect to θ and τ, we find that the minimum is attained when $\theta = \mu$ and $\tau = 1$. And, as expected, the value of the divergence at this point is zero. Thus, $P(M_1) \propto 1$. The prior mass for M_2 is based on the divergence $D_{KL}(f(x|\mu,\tau)\|f(x|\mu,1)) = \tau(\theta-\mu)^2/2 + (1/\tau - 1 + \log\tau)/2$. We can see that this is minimised, with respect to μ, when the two means are equal, and the value is $\min_\mu D_{KL}(f(x|\mu,\tau)\|f(x|\mu,1)) = (1/\tau - 1 + \log\tau)$. Therefore, we have*

$$P(M_2) \propto \exp\left\{ \int \int \frac{1}{2}\left(\frac{1}{\tau} - 1 + \log\tau\right) \pi_2(\mu,\tau)\, d\mu d\tau \right\}. \qquad (6.19)$$

Similarly as seen in Example 6.1, we note that the farther π_2 is from a point mass at one, the larger $P(M_2)$ becomes. This is shown by the fact that $(1/\tau + \log\tau)$ in (6.19) *is minimised at $\tau = 1$. And this expresses the idea that the more uncertain we are about the simpler model being the true one, the more mass we assign to* *the more complex model. As an illustration, we consider the prior for τ to be a Gamma distribution with shape parameter 5 and rate parameter 1. We then obtain that $P(M_2) \propto \exp(0.38) = 1.46$. With this result, the normalised prior mass is $P_N(M_1) = 0.41$ and $P_N(M_2) = 0.59$. It is of course possible, by changing the prior π_2, to obtain a different prior mass for M_1 and M_2.*

Again, we note from Example 6.2 that, when we consider nested models, the *worth* of the larger model is, at least, as large as the *worth* of the inner model, which is intuitive.

In the examples we have seen, we have considered nested models belonging to the same family. In the rest of the section, we examine model selection scenarios where the alternative models do not belong to the same family, strictly speaking.

6.5.1 Illustration 4: Normal and Student's t models

The first illustration for nested models not belonging to the same family of distributions, considers a Normal density and a Student's t density. We have already discussed properties and relations between the two models in Chapter 5. We then consider model M_1 to be a Normal distribution with mean μ and precision τ, and model M_2 to be a t distribution with location parameter θ, precision parameter λ and parameter ν representing the number of degrees of freedom. That is

$$M_1 = \left\{ f_1(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau}{2}(x-\mu)^2\right], \pi_1(\mu,\tau) \right\},$$

$$M_2 = \left\{ f_2(x|\theta,\lambda,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\lambda}{\nu\pi}\right)^{1/2} \left[1 + \frac{\lambda}{\nu}(x-\theta)^2\right]^{-\frac{\nu+1}{2}}, \pi_2(\theta,\lambda,\nu) \right\}.$$

The t distribution converges to a normal distribution when the number of degrees of freedom tends to infinity (Chu, 1956); as such, the two models can be considered nested models which differ from the number of degrees of freedom only (for example, see Casellas et al. (2008)). Therefore, as discussed above, we have that the minimum Kullback–Leibler divergence between M_1 and M_2 is zero, resulting in a prior mass on the normal model $P(M_1) \propto 1$.

To determine the mass for M_2, following our approach, we consider that, as shown in Theorem B.5 in Appendix B, we have

$$\begin{aligned}
D_{KL}(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau)) &= \int_{-\infty}^{\infty} f_2(x|\theta,\lambda,\nu)\log\left\{\frac{f_2(x|\theta,\lambda,\nu)}{f_1(x|\mu,\tau)}\right\} dx \\
&= \log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{1}{2}\log\lambda - \frac{1}{2}\log\nu \\
&- \frac{\nu+1}{2}\mathbb{E}\left\{\log\left(1+\frac{\lambda}{\nu}(x-\theta)^2\right)\right\} - \frac{1}{2}\log\tau \\
&+ \frac{1}{2}\log 2 + \frac{1}{2}\tau\,\mathbb{E}(x^2) - \tau\mu\,\mathbb{E}(x) + \frac{1}{2}\tau\mu^2. \qquad (6.20)
\end{aligned}$$

The divergence in (6.20) is minimised, with respect to μ and τ, when $\mu = \mathbb{E}(x) = \theta$ and $\tau = 1/Var(x) = \lambda$. That is, when the two distributions have location parameter and scale parameter of the same value. The minimum divergence is then

$$\begin{aligned}
\min_{\mu,\tau} D_{KL}(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau)) &= \log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu+1}{2} \\
&\mathbb{E}\left\{\log\left(1+\frac{\lambda}{\nu}(x-\theta)^2\right)\right\} + \frac{1}{2}\log\nu \\
&-\frac{1}{2}\log(\nu-2) + \frac{1}{2}.
\end{aligned}$$

To compute the prior mass for M_2, we consider the following prior distribution $\pi_2(\theta,\lambda,\nu) = \pi_{2,1}(\nu)\pi_{2,2}(\lambda)\pi_{2,3}(\theta|\lambda)$. Where $\pi_{2,1}(\nu)$ is an Exponential distribution (Geweke, 1993) with rate parameter equal to 1, $\pi_{2,2}(\lambda)$ is a Gamma with shape parameter 25 and rate parameter 1, and $\pi_{2,3}(\theta|\lambda)$ is a Normal distribution with mean zero and precision determined by the prior on λ. Thus

$$\begin{aligned}
P(M_2) &\propto \exp\left\{\int\int\int \min_{\mu,\tau} D_{KL}\left(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau)\right)\pi_2(\theta,\lambda,\nu)\, d\theta\lambda d\nu\right\} \\
&= \exp(0.23) = 1.26,
\end{aligned}$$

where the result has been obtained through numerical methods. By normalising, we have $P_N(M_1) = 0.44$ and $P_N(M_2) = 0.56$, which shows that more mass is given to the outer model. This is in line with the idea that, in relation to the other model, M_2 has more *worth*.

6.5.2 Illustration 5: Nested and non-nested models

In this final illustration, we consider a realistic model selection problem where the model space has both nested and non-nested elements, and a total of four models. We do this by adding an Exponential model to the selection scenario analysed in Section 6.4.3. That is, M_4 is an Exponential density with rate parameter θ

$$M_4 = \left\{ f_4(x|\theta) = \theta e^{-\theta x}, \pi_4(\theta) \right\}.$$

To identify the prior mass for model M_1, in addition to the results in Section 6.4.3, we need to consider the expected minimum Kullback–Leibler divergence with respect to the Exponential density. This is given by (refer to Theorem B.6 in Appendix B)

$$\begin{aligned} D_{KL}(f_1(x|\lambda,\kappa)\|f_4(x|\theta)) &= \int_0^\infty f_1(x|\lambda,\kappa) \log\left\{\frac{f_1(x|\lambda,\kappa)}{f_4(x|\theta)}\right\} dx \\ &= \log\kappa + \kappa\,\mathbb{E}(\log x) - \mathbb{E}(\log x) - \kappa\log\lambda \\ &\quad -\frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa) - \log\theta + \theta\,\mathbb{E}(x), \end{aligned}$$

which is minimised for $\theta = 1/\mathbb{E}(x) = \lambda^{-1}\Gamma(1+1/\kappa)^{-1}$. As expected, the two densities have minimum distance when the respective first moments are equal. Then $\min_\theta D_{KL}(f_1(x|\lambda,\kappa)\|f_4(x|\theta)) = \log\kappa - \gamma + \gamma/\kappa + \log\Gamma(1+1/\kappa)$. We note that the minimum Kullback–Leibler divergence between the Weibull and the Exponential densities does not depend on the scale parameter λ. To compute the expected minimum divergence, we have adopted the same prior distributions for the parameter of the Weibull we have used in Section 6.4.3. The result is $\mathbb{E}\{\min_\theta D_{KL}(f_1(x|\lambda,\kappa)\|f_4(x|\theta))\} = 0.05$. Given that this is the smallest expected divergence for model M_1 (refer to Table 6.2), we have $P(M_1) \propto (0.05) = 1.06$.

With a similar process, we find the Kullback–Leibler divergence between model

M_2 (Log-normal) and the model M_4 (refer to Theorem B.6 in Appendix B)

$$\begin{aligned} D_{KL}(f_2(x|\mu,\tau)\|f_4(x|\theta)) &= \int_0^\infty f_2(x|\mu,\tau)\log\left\{\frac{f_2(x|\mu,\tau)}{f_4(x|\theta)}\right\}\,dx \\ &= -\mathbb{E}(\log x)+\frac{1}{2}\log\tau-\frac{1}{2}\log(2\pi)-\frac{1}{2}\tau\,\mathbb{E}(\log^2 x) \\ &\quad +\tau\mu\,\mathbb{E}(\log x)-\frac{1}{2}\tau\mu^2-\log\theta+\theta\,\mathbb{E}(x), \end{aligned}$$

which is minimised for $\theta = 1/\mathbb{E}(x) = 1/\exp\{\mu+1/(2\tau)\}$, as expected. The minimum divergence is $\min_\theta D_{KL}(f_2(x|\mu,\tau)\|f_4(x|\theta)) = \{\log\tau-\log(2\pi)+1+\tau\}/2$, which does not depend on the location parameter μ of the Log-normal density. With the same priors for μ and τ considered in Section 6.4.3, we have obtained $\mathbb{E}\{\min_\theta D_{KL}(f_2(x|\mu,\tau)\|f_4(x|\theta))\} = 0.05$. As the smallest expected divergence for model M_2 remains the one with respect to the Gamma density (refer to Table 6.2), we have $P(M_2) \propto \exp(0.03) = 1.03$.

For the Gamma model M_3, we have (refer to Theorem B.6 in Appendix B)

$$\begin{aligned} D_{KL}(f_3(x|\alpha,\beta)\|f_4(x|\theta)) &= \int_0^\infty f_3(x|\alpha,\beta)\log\left\{\frac{f_3(x|\alpha,\beta)}{f_4(x|\theta)}\right\}\,dx \\ &= \alpha\log\beta-\log\Gamma(\alpha)+\alpha\,\mathbb{E}(\log x)-\mathbb{E}(\log x) \\ &\quad -\beta\,\mathbb{E}(x)-\log\theta+\theta\,\mathbb{E}(x), \end{aligned}$$

which is minimised by $\theta = 1/\mathbb{E}(x) = \beta/\alpha$. Therefore, we obtain the minimum divergence as $\min_\theta D_{KL}(f_3(x|\alpha,\beta)\|f_4(x|\theta)) = -\log\Gamma(\alpha)+\alpha\Psi(\alpha)-\Psi(\alpha)-\alpha+\log\alpha+1$. The expected minimum divergence has been computed using the same priors for α and β defined in Section 6.4.3, obtaining $\mathbb{E}\{\min_\theta D_{KL}(f_3(x|\alpha,\beta)\|f_4(x|\theta))\} = 0.05$. In this case as well, the divergence with respect to the Exponential distribution does not constitute the minimum (refer to Table 6.2), so we have $P(M_3) \propto \exp(0.02) = 1.02$.

To compute the prior mass for model M_4, we note that, being the Exponential nested into the Weibull and the Gamma models (it is in fact a special case of these two densities), we obviously have $D_{KL}(f_4(x|\theta)\|f_1(x|\lambda,\kappa)) = 0$ and $D_{KL}(f_4(x|\theta)\|f_3(x|\alpha,\beta)) = 0$. Therefore, we can conclude that $P(M_4) \propto 1$. However, for completion, we have: $\mathbb{E}\{\min_\theta D_{KL}(f_4(x|\theta)\|f_2(x|\mu,\tau))\} = 0.41$ (refer to

Theorem B.7 in Appendix B); where the expectation has been computed with respect to the priors for μ and τ defined in Section 6.4.3.

	M_1	M_2	M_3	M_4
M_1		0.08	0.06	0.00
M_2	0.09		0.02	0.41
M_3	0.05	0.03		0.00
M_4	0.05	0.05	0.05	
$P(M_j)$	1.05	1.03	1.02	1.00
$P_N(M_j)$	0.26	0.25	0.25	0.24

Table 6.2: Expected minimum Kullback–Leibler divergence (by column) among the models M_1 (Weibull), M_2 (Log-normal), M_3 (Gamma) and M_4 (Exponential). The divergences are computed considering the priors for the parameter of the models as defined in Section 6.4.3 and Section 6.5.1. The prior mass is proportional to the exponential of the minimum divergence, and the last two rows report this mass for each model, non-normalised $P(M_j)$ and normalised $P_N(M_j)$, $j = 1, 2, 3, 4$.

Table 6.2 summarised the results for this particular selection problem. We note that all the normalised prior probabilities are close to $1/4$. Given that we have kept the prior uncertainty about the parameter of the models at a relatively high level, the result is sensible. However, as we have already discussed in the previous illustrations, a change in the informational content within the prior distribution on the parameters will cause, in general, a different prior over the model space.

Another interesting consideration is that, by inspecting Table 6.2, we note that the expected Kullback–Leibler divergence between models M_1, M_2 and M_3 and model M_4 is constant and it is equal to 0.05. Recalling the results in this section, we have that the minimum divergences (in these three particular cases) depend only on the shape parameter of the models, respectively κ, τ and α. As we have used identical prior distributions for these parameters, the result obtained is sensible in the light of the prior information considered.

6.6 Discussion

Similarly to parameter spaces, we can assign prior mass over a model space by quantifying the *worth* that each model has, in relation to the others. It is important to highlight that we don't have to assume that one of the models is correct: we evaluate the *worth* by thinking what is lost if we remove a model and it is correct. By doing this for all models, we obtain an objective value of each of their worth, which is then linked to the prior mass via the *self-information* loss function.

An important result we have obtained is that the prior on a model should depend on the model itself; that is $\{f(x|\theta), \pi(\theta)\}$. This is evident in Example 6.1, where we compare the Normal density $N(\mu, 1)$ to the Normal density $N(0, 1)$. It is well known that assigning equal probability to the two models may result in the so called Jeffreys-Lindley paradox. This is because the uncertainty on the parameter value (i.e. μ) is not "fairly" represented by the equal masses. A common solution to avoid the paradox, see for example (Bernardo, 1999), is to set $\pi(\mu) \propto 1$, which is an expression of maximum uncertainty about the parameter value. And our approach is in line with this, by assigning more and more mass to the model $N(\mu, 1)$ as the uncertainty on μ increases.

The proposed method can be applied to any selection problem. Particular results have been obtained for nested models. In fact, when a model is nested into another, among all models taken into consideration, its prior mass is zero. The result is not surprising, as the more complex model is (at least) as good as the simpler one, and there is no information loss in removing the simpler model.

Chapter 7

Variable Selection in Linear Regression Models

In this chapter we discuss variable selection in linear regression models. In particular, by considering the approach discussed in Chapter 3, we illustrate how regression model prior distributions can be defined on the basis of losses.

We begin with a brief overview of objective Bayesian procedures that deal with this type of problem. The review includes objective priors for models as well as objective priors for the parameters of the model. We then present the prior for the model, given a specific type of parameter-specific prior: the g-prior (Zellner, 1986). Although g-priors have some weaknesses (e.g. the *information paradox* discussed in Section 7.1.1), they represent a good compromise, as they allow for a closed form representation of the marginal and the posterior distributions.

In the last section we present a result that raises some questions about Bayesian procedures for variable selection. The result is not within the scope of the thesis, so it will be simply introduced and generally discussed. It is however noteworthy as it opens the door to further specific research in the field of Bayesian variable selection.

7.1 Introduction

Consider the linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where y is the response variable, $x_1, \ldots, x_p$ are p covariates and ε is the error term, with $\varepsilon_i \overset{iid}{\sim} N(0, \lambda)$, $\lambda = 1/\sigma^2$ representing the regression precision. We assume $n > p$ and the design matrix X to have full rank. For simplicity in the notation, we also assume $\beta_0 = 0$ for all regression models. Variable selection problems consist in finding how many and which one of the p covariates have a significant impact on the response variable, and therefore should be included in the regression model. We can indicate the generic model by M_γ, where γ is a p-dimensional binary vector. Each element of vector γ corresponds to a covariate, such that

$$\gamma_j = \begin{cases} 0 & \text{if } \beta_j = 0, \\ 1 & \text{if } \beta_j \neq 0. \end{cases} \qquad j = 1, \ldots, p.$$

Note that γ can be considered as a random variable taking values in the space $\{0, 1\}^p$, and each model has associated a dimension $\dim(\gamma)$ representing the number of covariates included. The model prior probability is indicated by $P(M_\gamma)$, and it represents the prior belief that model M_γ is the true one. In a Bayesian framework, inference is based on the model posterior probabilities, which are obtained by combining the model prior and the marginal likelihood of the observations y under each model.

$$f(y|M_\gamma) = \int f(y|\beta_\gamma, \lambda)\pi(\beta_\gamma, \lambda)\, d\beta_\gamma d\lambda. \tag{7.1}$$

In (7.1), $\pi(\beta_\gamma, \lambda)$ represents the prior assigned to the parameters of the model: the coefficients β_γ, and the precision λ. The prior $\pi(\beta_\gamma, \lambda)$ is identified in the literature as the prior for *model-specific parameters.* By applying Bayes theorem, the model posterior probability is given by

$$P(M_\gamma|y) \propto f(y|M_\gamma)P(M_\gamma).$$

7.1.1 Prior on model parameters

Although our primary interest is in model prior $P(M_\gamma)$, we deem as appropriate to briefly discuss priors for the model-specific parameters. It is clear from (7.1) that the marginal likelihood depends on the choice of the priors for β_γ and λ. The literature on the subject is extensive, and its detailed discussion is beyond the scope of this work. Recall that Bayesian model selection, and therefore variable selection, can be performed by means of Bayes factors (as seen in Chapter 6), we note the following challenges (Berger and Pericchi, 2001):

1. The number of Bayes factors that have to be computed rapidly grows when the number of covariates grows. For a p covariates case, the number of possible models is 2^p, clearly resulting in a large number of computations that have to be performed even for moderate values of p;

2. Objective (improper) priors can only be used for the parameters common to the two models compared through the Bayes factor. The arbitrary constant we could multiply each improper prior would not cancel out for the non-common parameters, resulting in an indeterminate Bayes factor;

3. Vague (proper) priors must not be used. In this case, the resulting Bayes factor would be affected by the arbitrary level of "vagueness" of the prior, rendering the analysis ineffective in practice;

4. Either in subjective or objective Bayes, common parameters can change meaning for different models. In theory, the prior distribution on the common parameters should change in order to reflect the different meaning. This issue has not easy solution. Refer, for example, to Berger and Pericchi (2001) and the references therein.

The first work in defining objective priors for variable selection problems can be found in Jeffreys (1961). The idea is to use objective (improper) priors for the common parameters and proper (but not vague) priors for the non-common parameters. Jeffreys proposed a Cauchy distribution centred at zero, with scale parameter σ^2, as the prior for β; the well known objective prior $\pi(\sigma) \propto 1/\sigma$ for the standard deviation of the regression.

A popular prior is the g-prior proposed by Zellner (1986). In this case, we have $\pi(\lambda) \propto \lambda^{-1}$ for the precision, and $\pi(\beta_\gamma|\lambda) \sim N(0, g/\lambda \left(X'_\gamma X_\gamma\right)^{-1})$ for the coefficients of the regression model. The prior has the undesirable property that, if the true model is M_γ, the Bayes factor in favour of this model would (asymptotically) tend to a constant rather than to infinity. That is, it converges to $(1+g)^{m-p-1}$. This is known as the *information paradox*. However, as this is the only prior that yields a closed form expression for marginal likelihoods (and because the above constant is generally very large in value), g-priors are appealing prior distributions. For this reason, we will be considering g-priors for our discussions.

A g-like prior, which does not generate the above *information paradox*, was proposed by Zellner and Siow (1980). The idea, is to have a Cauchy prior for the coefficients: $\pi(\beta_\gamma|\lambda) \sim Ca(0, n/\lambda \left(X'_\gamma X_\gamma\right)^{-1})$; thus, the Cauchy distribution is expressed as a scale mixture of normals, $\pi(\beta_\gamma|\lambda) \propto \int N(\beta_\gamma|0, g/\lambda(X'_\gamma X_\gamma)^{-1})\pi(g)\,dg$, and a prior assigned to g. In particular, $\pi(g)$ would be an Inverse-Gamma with parameters $1/2$ and $n/2$. The downside of this prior is that, unlike the g-prior, it does not yield closed form expressions for the marginal likelihoods.

The prior proposed by Zellner and Siow (1980), can be seen as part of the hyper-g priors discussed in Liang et al. (2008) (also discussed in Cui and George (2008)), where other prior distributions for g, besides the Inverse-Gamma, are illustrated.

A different way to deal with the problem of model selection is suggested by Berger and Pericchi (1996). They propose to solve the issue of not being able to use objective improper priors for the non-common parameters by working with a particular form of Bayes factor: the Intrinsic Bayes Factor (IBF). The idea is to use part of the data, called the *training sample*, to convert the ordinary objective improper priors into proper posteriors. These are then used to compute Bayes factors for the remaining data. A conceptually similar solution has been proposed by O'Hagan (1995), with the Fractional Bayes Factors (FBF). In this case, improper priors are converted into proper, not by using part of the data, but by using a fraction of the likelihood function.

Finally, the Bayesian Information Criterion (BIC) is an asymptotic method for model selection. It was introduced by Schwarz (1978) with the following form

$$B_{ji} = \frac{f_j(x|\hat{\theta}_j)}{f_i(x|\hat{\theta}_i)} n^{(dim(i)-dim(j))/2},$$

where $\hat{\theta}_j$ and $\hat{\theta}_i$ are, respectively, the maximum likelihood estimates of the parameters θ_j of model M_j and θ_i of model M_i. The BIC is appealing for its simplicity. However, it has been shown that the criterion may lead to issues if any of the models has irregular asymptotics, or it has a likelihood that tends to concentrate at the boundaries of the parameter space.

7.1.2 Model priors

The main point of discussion relevant to this work is on model prior probabilities $P(M_\gamma)$. The set of all possible models (i.e. the model space) is discrete, and therefore suitable to the novel objective approach we present in this thesis.

In objective Bayesian variable selection, an important role is played by the following two ways of assigning prior probability to models. One is intuitive, and it assigns equal probability to each model: $P(M_\gamma) = 1/2^p$. The second way, discussed in Scott and Berger (2010), is based on the idea that the probability of including a covariate in the model, $\omega_j = P(\gamma_j \neq 0)$, can be seen as a Bernoulli trial. Therefore, the prior probability of model M_γ, given ω, is

$$P(M_\gamma|\omega) = \omega^{dim(\gamma)}(1-\omega)^{p-dim(\gamma)}.$$

Integrating out ω, we have

$$P(M_\gamma) = \int_0^1 P(M_\gamma|\omega)\pi(\omega)\, d\omega = \frac{1}{p+1}\binom{p}{dim(\gamma)}^{-1}, \tag{7.2}$$

where it is assumed $\pi(\omega) \sim Be(1,1)$. Prior (7.2) assigns a mass to model M_γ which value depends on $dim(\gamma)$. It is of course possible to assign to ω a Beta prior with different values of the parameters; or to use a different probability distribution, (see George and McCulloch (1997) and Ley and Steel (2009)). However, the above choice seems to be to most appropriate to reflect a *priori* absence of knowledge on the value of ω.

Definition 7.1. *The inclusion prior probability for covariate x_j is given by*

$$\omega_j = \sum_\gamma \Pr(M_\gamma) \cdot 1_{(x_j \text{ } in \text{ } M_\gamma)}, \qquad j = 1, \ldots, p,$$

where $1_{(\cdot)}$ is the indicator function, and the summation is extended to all the 2^p possible regression models.

If we consider the uniform model prior, that is $P(M\gamma) = 1/2^p$, the inclusion prior probabilities ω_j are all equal to 1/2. This result, according to Scott and Berger (2010), leads to an issue (known as *multiplicity*) which affects the use of Bayes factors seen as a "multiplicity" of tests of hypotheses. For model prior (7.2) the prior inclusion probability is $\omega_j = 1/2$ as well; however, as model prior probabilities are different from the uniform case, there is a hidden mechanism in the process that automatically corrects for multiplicity (Scott and Berger, 2010).

Bayesian inference in variable selection problems can be performed in different ways. First, we need to consider that, when the number of covariates is relatively large, model posterior probabilities will most likely be of small value. Therefore, the choice of the most probable model as the estimate can be both not possible and meaningless. A solution is to adopt model averaging techniques (refer to Steel (2012) and the reference therein). The general idea is to estimate the quantity of interest (e.g. a forecast) with each model (or a selection of the models with highest posterior probability) and average the estimates using model prior probabilities as weights.

Another solution is to consider posterior inclusion probabilities.

Definition 7.2. *The inclusion posterior probability for covariate x_j, is given by*

$$\widetilde{\omega}_j = P(\gamma_j \neq 0|y) = \sum_\gamma P(M_\gamma|y) \cdot 1_{(x_j \text{ } in \text{ } M_\gamma)}, \tag{7.3}$$

for $j = 1, \ldots, p$.

It is common to consider the regression model composed by the intercept plus all the covariates with a posterior inclusion probability larger than or equal to

1/2. Barbieri and Berger (2004) show that this model, known as the *median-probability model*, generally has better predictive properties than the model with highest posterior probability.

7.2 The Villa–Walker prior for linear regression

Our approach to variable selection is in line with the general idea we have presented so far. We obtain model prior probabilities, not directly, but by considering the *worth* that each model has in the model space. The *worth* is determined in a similar way to the one we have discussed for model selection in Chapter 6. In addition, model complexity has to be taken into consideration: models with a large number of covariates tend to fit the data better than models with a small number. The "cost" of a better behaviour is, however, a model that is harder to interpret and more demanding, in terms of estimation procedure.

Let us consider the regression model M_γ, with $dim(\gamma)$ covariates. The loss of removing it from the set of all possible 2^p models (which coincides with the utility of keeping it), can be represented as

$$\text{Loss}(M_\gamma) = \text{Loss}(M) + \text{Loss}(Co), \tag{7.4}$$

The loss in (7.4) is a cumulative loss with two components: one representing the *worth* of the model, indicated by Loss(M), and one that takes into account how complex the model is, indicated by Loss(Co). The component $P(M)$ of (7.4) is defined as in Chapter 6, and it represents what do we lose if regression model M_γ is kept in the model space, and it is the true one. This is measured by the expected Kullback–Leibler divergence between the regression model M_γ and the nearest one. Except the full model, each regression model is nested into (at least) another model. Therefore, as discussed in Chapter 6, the minimum Kullback–Leibler divergence is zero. For the full model, in order to determine the first component of the cumulative loss in (7.4), we would need to identify the minimum expected divergence with respect to the p models with $p-1$ covariates. This divergence will obviously depend on the prior for the model-specific parameters. However, for moderate to large values of p, it can be assumed that the expected divergence is very small; therefore, we can approximate (7.4) with the more simple $\text{Loss}(M_\gamma) \approx \text{Loss}(Co)$.

To quantify $\text{Loss}(Co)$, we proceed as follows. If we keep model M_γ in the model space, the loss would be proportional to the number of covariates that have to be considered and measured. Therefore, the loss of keeping a model increases as the dimension of the model increases. Following our approach (refer to Chapter 3), we have that the loss in removing model M_γ is equal to the utility in keeping model M_γ. Considering that the loss in keeping model M_γ is proportional to the "need" of $dim(\gamma)$ covariates, we have

$$l(\text{remove } M_\gamma) = u(\text{keep } M_\gamma) = -c \cdot dim(\gamma),$$

and

$$l(\text{keep } M_\gamma) = c \cdot dim(\gamma),$$

where c is a real constant. By considering the *self-information* loss function, we have $-\log P(M_\gamma) \propto c \cdot dim(\gamma)$; which implies

$$P(M_\gamma) \propto e^{-c \cdot dim(\gamma)}, \qquad dim(\gamma) = 0, 1, \ldots, p. \tag{7.5}$$

In Section 3.3 we have mentioned that the loss functions involved in our prior do not require a constant, as we are equating losses in information. However, the loss associated to model M_γ has two components of different nature: a loss in information and a loss due to the complexity of the model, represented, respectively, by the first and the second term of the right-hand-side of (7.4). As such, as we are no longer equating losses in information only, it is necessary to consider constant c, as shown in (7.5).

The following theorem shows the expression for the prior inclusion probabilities.

Theorem 7.1. *Let us assume that the model prior for a variable selection problem with p covariates is of the form (7.5). Then, the prior inclusion probability for each covariate is given by*

$$\omega_j = (1 + e^c)^{-1}, \qquad j = 1, \ldots, p \quad \textit{and} \quad c \in \mathbb{R}.$$

Proof. The prior probability for each model is $P(M_\gamma) \propto \exp\{-c \cdot dim(\gamma)\}$. The normalising constant is given by

$$K = \sum_{k=0}^{p} \binom{p}{k} e^{-k} = \frac{(1+e^c)^p}{e^{cp}} = (e^{-c}+1)^p.$$

Each covariate is included in half of the 2^p models, and the total number of covariates in the models where a given covariate is included follows Pascal's triangle. For example, if $p = 3$, we have $2^3 = 8$ regression models; each covariate appears in one model with one covariate, three models with, respectively, 2 and 3 covariates, and in the full model. Generalising, the prior inclusion probability for covariate x_j is given by

$$\begin{aligned} \omega_j &= \sum_{k=0}^{p-1} \binom{p-1}{k} e^{-c\cdot(k+1)} \Big/ K \\ &= \frac{(1+e^c)^{p-1}}{e^{cp}} \frac{e^{cp}}{(1+e^c)^p} \\ &= \frac{1}{1+e^c}, \end{aligned} \tag{7.6}$$

which proves the theorem. □

We see from (7.6) that the choice of c has an impact on the prior inclusion probabilities; in particular, ω_j decreases as c increases. In Section 7.4 we will see that, in an orthogonal design, the marginal posterior inclusion probabilities are bounded below by ω_j; as such, the choice of c impacts the model selected as it might determine which covariates will be included. Given that, in principle, the choice of c is arbitrary, the implication on the objectivity of our approach is clear. However, the following considerations are in order. The threshold for the posterior inclusion probability above which we select a covariate is arbitrary as well; therefore, the aim of obtaining a model prior that is "purely" objective is, somehow, rendered pointless.

In Section 7.1.2 we have mentioned that the *median-probability model* can be, for prediction purposes, a sensible choice. Therefore, by setting $c = 1$ we would have $\omega_j \approx 0.27$, and $\widetilde{\omega}_j \in (0.27, 1)$, which comfortably includes the threshold 0.5.

Obviously, a different choice of c would be equally plausible. For example, if we set $c = 2.94$, the prior inclusion probability will be 0.05, allowing $\widetilde{\omega}_j$ to vary in an interval that would include any reasonable threshold.

7.3 Illustration: US crime data analysis

We used the crime data of Liang et al. (2008) as an illustration of our model prior. The dataset consists of 47 observations for $p = 15$ covariates related to crime data in the US. For the model specific parameters, we have used a g-prior with $g = 2^{15}$. Figure 7.1 shows the posterior inclusion probabilities of $x_1, \ldots, x_{15}$ (on log-scale) obtained by adopting as model prior, in turn, the uniform prior, Scott & Berger's and the prior based on our approach with $c = 1$.

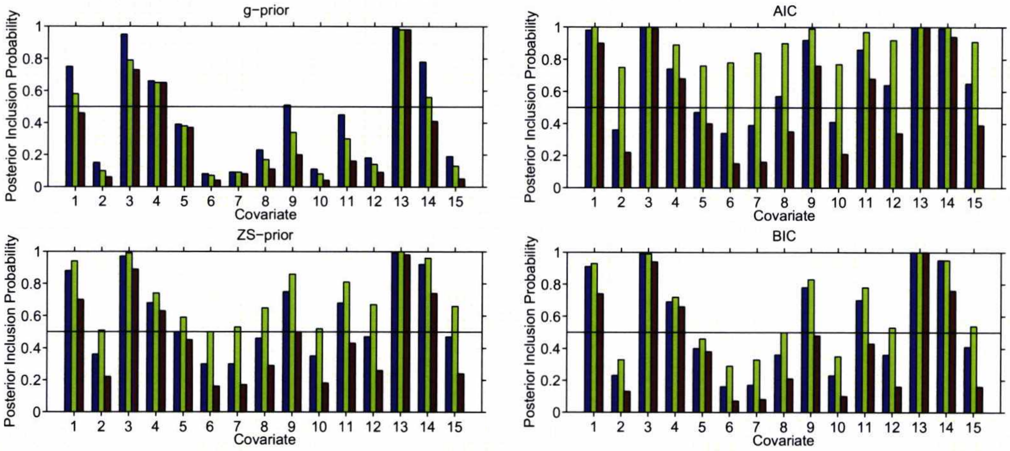

Figure 7.1: Marginal posterior inclusion probabilities for crime data. The prior for the model-specific parameters used are (top-to-bottom, left-to-right) g-prior, Zellner-Siow prior, AIC and BIC. The model priors are Uniform prior (left blue bar), Scott & Berger prior (middle green bar) and Villa-Walker prior (right red bar).

We note that, should we consider the median-probability model, covariates x_1, x_2, x_{13} would be included in all the three cases. Covariate x_4 would appear only when the uniform model prior is considered. Unlike the other priors, our prior does not does not lead to a sufficiently large posterior inclusion probability for covariates x_1 and x_{14}. It seems that, from the largest to the smallest, posterior inclusion probabilities follow Uniform prior, Scott & Bergers' and our.

In Figure 7.1 we have also shown the results when Zellner-Siow prior for model-specific parameters, Akaike information criterion (AIC) and BIC are used. We note that, in any situation, the median-probability model obtained with our model prior is the most parsimonious; in particular, the resulting marginal posterior inclusion probability is consistently not larger than the one obtained by using any of the other two model prior. There is a remarkable result when Scott & Bergers' model prior is adopted: under Zellner-Siow prior and AIC, the median-probability model includes all the covariates. For reasons beyond the scope of this thesis we have not investigated further; however, there may be connections to the results presented in Section 7.4 below.

7.4 Some interesting results

Let us consider the simplest variable selection problem in linear regression, where we compare the null model with the model with one covariate: $p = 1$.

$$M_1 : y_i = \sigma\varepsilon_i \qquad M_2 : y_i = \beta x_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are i.i.d. and σ^2 is the regression variance. Let us assume that we define the model prior in accordance with Scott and Berger (2010): $P(M_1) = P(M_2) = 1/2$ (note that this prior corresponds to the uniform as well). As per (7.3), the posterior inclusion probability for x is given by

$$\begin{aligned}\widetilde{\omega} = P(M_2|y) &= \frac{f(y|M_2)P(M_2)}{f(y)} \\ &= \frac{f(y|M_2)1/2}{\{f(y|M_1) + f(y|M_2)\}1/2} \\ &= (1 + B_{12})^{-1}, \qquad (7.7)\end{aligned}$$

where B_{12} is the Bayes factor comparing model M_1 to model M_2 (refer to Section 6.1). Let us now assume, without loss of generality, that $\sum x_i^2 = 1$. Also, we consider the prior for the precision $\lambda = \sigma^{-2}$ to be proportional to λ^{-1}, and the prior for the coefficient to be the $g-$prior discussed in Section 7.1.1: that is, $N(\beta|0, g/\lambda)$, for $g > 0$. Thus, we have

$$\begin{aligned} B_{12} &= \frac{\int \lambda^{n/2} \exp\{-0.5\lambda \sum y_i^2\} \lambda^{-1}\, d\lambda}{\int \left[\int \lambda^{n/2} \exp\{-0.5\lambda \sum (y_i - x_i\beta)^2\} \lambda^{1/2} \exp\{-0.5g\lambda\beta^2\}\, d\beta\right] \lambda^{-1}\, d\lambda} \\ &= \frac{\int \lambda^{n/2} \exp\{-0.5\lambda \sum y_i^2\} \lambda^{-1}\, d\lambda}{\int \lambda^{n/2-1} \exp\{-0.5\lambda \left[y'y - z^2/(1+g)\right]\}\, d\lambda}, \end{aligned}$$

where $z = \sum y_i x_i$, and where we have not considered the terms that cancel out. Hence, the Bayes factor becomes

$$B_{12} = \frac{\{y'y - z^2/(1+g)\}^{n/2}}{(y'y)^{n/2}} < 1. \tag{7.8}$$

From the relationship in (7.7), the result in (7.8) implies that $\widetilde{\omega} > 1/2$. That is, the posterior inclusion probability for x is larger than the prior inclusion probability $\omega = P(M_2) = 1/2$.

For $p = 2$ we have four possible models.

$$\begin{array}{ll} M_1 : y_i = \sigma\varepsilon_i & M_2 : y_i = \beta_1 x_{1i} + \sigma\varepsilon_i \\ M_3 : y_i = \beta_2 x_{2i} + \sigma\varepsilon_i & M_4 : y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma\varepsilon_i \end{array}$$

Let us consider an orthogonal design: $\sum x_{1i}^2 = \sum x_{2i}^2 = 1$ and $\sum x_{1i}x_{2i} = 0$. The g-prior on the coefficients will then have the form

$$\pi(\beta_1, \beta_2, \lambda) \propto \exp\left\{-\frac{1}{2} g\lambda \left(\beta_1^2 + \beta_2^2\right)\right\},$$

with likelihood function

$$f(y|\beta_1, \beta_2, \lambda) \propto \exp\left\{-\frac{1}{2}\lambda \left(y'y + \beta_1^2 + \beta_2^2 - 2\beta_1 z_1 - 2\beta_2 z_2\right)\right\},$$

where $z_1 = \sum y_i x_{1i}$ and $z_2 = \sum y_i x_{21}$. As the design is orthogonal, the term $2\rho\beta_1\beta_2 = 0$, for $\rho = 0$. Under these orthogonality conditions, the marginal likelihood is an increasing function of the model dimension. To see this, let us consider the model prior proposed in Scott and Berger (2010). Therefore, we have $P(M_1) = P(M_4) = 1/3$ and $P(M_2) = P(M_3) = 1/6$. If we adopt the g-prior for the parameters of the model, the marginal likelihoods are

$$\begin{aligned} f(y|M_1) &\propto (y'y)^{-n/2} \\ f(y|M_2) &\propto \left(y'y - z_1^2\right)^{-n/2} \\ f(y|M_3) &\propto \left(y'y - z_2^2\right)^{-n/2} \\ f(y|M_4) &\propto \left(y'y - z_1^2 - z_2^2\right)^{-n/2}, \end{aligned}$$

from which emerges the clear relationship

$$f(y|M_1) < f(y|M_2) \quad \text{and} \quad f(y|M_3) < f(y|M_4).$$

For the model posterior probabilities we have

$$\begin{aligned} P(M_1|y) &\propto 1/3\,(y'y)^{-n/2} \\ P(M_2|y) &\propto 1/6\left(y'y - z_1^2/(1+g)\right)^{-n/2} \\ P(M_3|y) &\propto 1/6\left(y'y - z_2^2/(1+g)\right)^{-n/2} \\ P(M_4|y) &\propto 1/3\left(y'y - (z_1^2 + z_2^2)/(1+g)\right)^{-n/2}, \end{aligned}$$

which results in

$$P(M_1|y) < 2\,P(M_2|y) \quad \text{and} \quad 2\,P(M_3|y) < P(M_4|y). \tag{7.9}$$

The probability that covariate x_1 is in the model can be expressed as

$$\widetilde{\omega}_1 = P(x_1 \text{ in}|x_2 \text{ in}, y)\widetilde{\omega}_2 + P(x_1 \text{ in}|x_2 \text{ out}, y)(1 - \widetilde{\omega}_2).$$

Considering (7.9), we have

$$P(x_1 \text{ in}|x_2 \text{ in}, y) = \frac{P(M_4|y)}{P(M_4|y) + P(M_3|y)} > \frac{1}{2},$$

and

$$P(x_1 \text{ in}|x_2 \text{ out}, y) = \frac{P(M_2|y)}{P(M_2|y) + P(M_1|y)} > \frac{1}{2}.$$

Thus, $\widetilde{\omega}_1 > 0.5\widetilde{\omega}_2 + 0.5(1 - \widetilde{\omega}_2) = 1/2$. This implies that $\widetilde{\omega}_1 > \omega_1$, given that $\omega_1 = 1/2$, as discussed in Section 7.1.2. The result holds for $\widetilde{\omega}_2$ as well. We have $\widetilde{\omega}_2 = P(x_2 \text{ in}|x_1 \text{ in}, y)\widetilde{\omega}_1 + P(x_2 \text{ in}|x_1 \text{ out}, y)(1 - \widetilde{\omega}_1)$. Also

$$P(x_2 \text{ in}|x_1 \text{ in}, y) = \frac{P(M_4|y)}{P(M_4|y) + P(M_2|y)} > \frac{1}{2},$$

and

$$P(x_2 \text{ in}|x_1 \text{ out}, y) = \frac{P(M_3|y)}{P(M_3|y) + P(M_1|y)} > \frac{1}{2}.$$

Then, from $\widetilde{\omega}_2 > 0.5\widetilde{\omega}_1 + 0.5(1 - \widetilde{\omega}_1) = 1/2$, we have $\widetilde{\omega}_2 > \omega_2$.

It is easy to see that the result is valid even if we replace the prior of Scott and Berger (2010) with a uniform prior. That is, by setting $P(M_j) = 1/4$, $j = 1, 2, 3, 4$.

The above results appear to highlight a key issue in objective variable selection: if the design is orthogonal, the posterior inclusion probabilities are bounded below by the prior ones and, if we aim to adopt the *median-probability model*, model priors like the Uniform or the Scott and Berger (2010), which result in $\omega_j = 1/2$, may lead to a model that includes all the p covariates. If, on the other hand, we use the Villa–Walker model prior with, say, $c = 1$, the issue does not appear as $\omega_j = 0.27$.

The mathematical conclusions discussed above are supported by the following simulations. We consider a regression model with $p = 2$ and $\sigma = 1$. For $n = 100$, we draw one million of $n \times p$ orthogonal design matrices of covariates $N(0, 1)$. We uniformly draw β_1 and β_2 from the interval $(-10, 10)$, and the response vector $y \sim N(\beta_1 x_1 + \beta_2 x_2, \sigma)$. Assuming a g-prior ($g = n$) for the parameters of the model, and the model prior proposed by Scott and Berger (2010), we compute marginal likelihoods, model posteriors and posterior marginal inclusion probabilities for x_1 and x_2. We see that in the 0.2% of the simulated scenarios $\widetilde{\omega}_j < 1/2$ ($j = 1, 2$). The reason of the exceptions is related to the fact that, computationally, orthogonal matrices can be obtained with $\sum x_1 x_2 = 0$ up to a certain level of precision only. We have also repeated the simulation considering, in turn, $\beta_1 = 0$ and $\beta_2 = 0$. The results are in line with the above ones. By replacing the model prior with the one we propose in (7.5), and leaving the remaining simulation settings unchanged, we obtain results consistent with Theorem 7.1. In particular, the percentage of cases where the inclusion posterior is less than 0.27 is 0.1%. The percentage is zero for $\beta_1 = 0$ ($\beta_2 = 0$). Thus, in the case of an orthogonal design matrix, the inclusion posterior probability is bounded below by the inclusion prior probability.

7.5 Discussion

In terms of prior probability on the space of models, our proposal based on losses leads to a simple result. With respect to the complexity of the model, expressed by the number of covariates included, our approach assigns more prior mass to the less complex model. The mass then decreases toward zero as we approach the full model. The result on actual data shows that our prior tends to be more parsimonious when compared to the Uniform model prior or the one in Scott and Berger (2010); this is the case when we use the g-prior or the Zellner-Siow prior. Noteworthy the result for Scott & Berger's prior when Zellner-Siow's prior is used for the parameters: the *median-posterior probability model* includes all the covariates. It has to be noted that, as the Villa–Walker prior depends on a constant c, it cannot be considered as "purely" objective; however, this does not constitute an issue as the whole Bayesian variable selection procedure includes subjective arguments. Firstly, the threshold that we use to decide which covariate is included is arbitrary; second, the determination of the degree of complexity of a model, which we assume to be represented by the umber of covariates in the model, is not univocal and other plausible criteria may be considered.

Although it is out of the scope of this thesis, the result in Section 7.4 raises some questions. In an orthogonal design, which is supposed to be the ideal one, posterior inclusion probabilities appear to be bounded below by the corresponding prior inclusion probability. This has been proved and illustrated with a simulation when $p = 2$ and the g-prior is adopted. Whilst the marginal likelihood stops increasing after a certain dimension of the model, creating an automated Occam's razor (Scott and Berger, 2010), this appear to not occur in an orthogonal design. Hence the result. We have not investigated further, but Bayesian approaches for variable selection in linear regression model should be revisited considering the above outcome. We leave this to future work.

Chapter 8

Discussion

The aim of this thesis is to introduce a novel Bayesian approach to derive objective priors for discrete parameters. We show how the idea of measuring the *worth* of each element in the parameter space leads, through the Kullback–Leibler divergence and the *self-information* loss function, to the prior

$$\pi(\theta) \propto \exp\{\min_{\theta' \neq \theta \in \Theta} D_{KL}(f(x|\theta) \| f(x|\theta'))\}.$$

The motivation for developing the prior has to be sought in the fact that no general methods to derive objective priors for discrete parameter spaces have been proposed. The recent publication by Berger et al. (2012) represents an attempt to move in this direction; however, the results lack generality. The application of our approach to various discrete models, illustrated in Chapter 4, shows that working with losses does not require any pre- or post-process analysis, and that the approach is versatile. In Chapter 5, we derive an objective prior for the number of degrees of freedom of a t density. The application of our criterion leads to an important conclusion: an objective prior for this parameter has to be truncated. This is a consequence of the well known property of the t density to converge to a Normal density.

Working with losses instead of probabilities allows us to obtain another important result. Bayes theorem is conceptually problematic when improper prior distributions are used. However, by expressing prior and posterior beliefs as losses, not probabilities, we derive a meaningful representation of Bayes theorem

$$-\log \pi(\theta|x) = K - \log f(x|\theta) - \log \pi(\theta).$$

It is important to reiterate that we do not claim our objective priors are proper, but that being objective in determining losses gives a coherent interpretation of Bayes theorem, as prior and posterior retain the same meaning.

A result from Chapter 6, where we derive objective model prior probabilities, is that the prior on a model has to depend on the model itself; where the model includes the distribution and the prior on the parameters. The result is presented by referring to the well known Jeffreys-Lindley paradox. We show various examples of application of our model prior. In particular, we first show that it can be applied to cases where the model defined on discrete supports and continuous support. We also illustrate how the approach is not affected by the size of the parameter space of the models. We extend the illustrations to model spaces with more than two models and, in addition, include both nested and non-nested models. A noteworthy result is that, if in the set of options we have nested models, the computation is largely simplified as there is no loss in moving from the inner model to the outer one.

The versatility of our approach to discrete spaces is further illustrated in Chapter 7, where we discuss objective Bayesian variable selection. We show that approaching the problem of assigning a prior on a regression model via losses, gives a prior function that is relatively simple: $P(M_\gamma) \propto \exp\{-c \cdot dim(\gamma)\}$. The prior is analysed on real data and compared to the Uniform model prior and the model prior proposed by Scott and Berger (2010). It appears that our prior is more parsimonious that the other two; in addition, unlike the prior of Scott and Berger (2010), does not give singular results when the Zellner-Siow prior is used for model-specific parameters. In the chapter we also briefly discuss a result that, although not directly related to the scope of this thesis, is interesting: in an orthogonal design, when g-priors are used, the marginal posterior inclusion probabilities are bounded below by the respective marginal prior inclusion probability. We believe that the result deserves a thorough investigation, and will be mentioned in the next section, where we present some ideas for future work.

8.1 Future work

What possible represents the main topic of interest for future work, is the extension of the approach to continuous parameter spaces. For a successful outcome would mean the definition of an objective approach capable of dealing with any parameter space. The main challenge is that

$$\min_{\theta' \neq \theta \in \Theta} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta')) = 0,$$

when Θ is continuous. In Appendix A we show how this issue may be solved by considering a discretisation of the parameter space. Leaving aside conceptual concerns that such procedure may raise, we obtain some noteworthy results. By applying it to the parameters of a Normal density, we see that it is always possible to build a discretised structure of the parameter spaces such that the priors are uniform. But, it is not possible to do the same when the target is Jeffreys' prior.

Brown and Walker (2012) show that a prior can be obtained by considering the following result from Blyth (1994)

$$\lim_{\delta \to 0} \frac{1}{\delta^2} D_{KL}(f(\cdot|\theta) \| f(\cdot|\theta + \delta)) = \sum I_{jk}(\theta),$$

where $I_{jk}(\theta)$ is the jk-th element of the Fisher information matrix. By considering the loss of keeping θ equal to $-\log \left\{ \sum I_{jk}(\theta) \right\} / 2$, it can be shown that the approach yields Jeffreys prior when the parameter is a scalar. However, as it seems not necessary to consider the result on the log-scale, nor taking its square root, further work has to be carried out to be able to classify the process as fully objective; that is, depending on the choice of the model only.

Other types of discrete scenarios, not discussed in this thesis, can be considered for future work. One of these is polynomial regression

$$y_i = \sum_{j=0}^{p} \beta_i x_i^j + \sigma \varepsilon_i,$$

where the objective is to estimate the discrete parameter p. More generally

$$y_i = \sum_{j=0}^{p} \beta_j \varphi_j(x),$$

where $\varphi_j(x)$ is a basic function different from the polynomial basic function; for example, splines or wavelets. For mixture models, such as the one of the form $\sum_{j=1}^{p} \omega_j N(y|\theta_j)$, we could be interested in assigning a prior on p, in order to estimate the number of components in the model.

Appendix A

Mathematical Support for Chapter 4

A.1 Proofs

Proof of Lemma 4.1

Proof. We prove the lemma by considering the sign of the difference between $D_{KL}(f_{R_0} \| f_{R_0+1})$ and $D_{KL}(f_{R_0} \| f_{R_0-1})$, which depends on the (relative) values of R_0, N and n. This difference is obtained by replacing in (4.13) R with R_0 and, respectively, R' with $R_0 + 1$ and $R_0 - 1$. Thus

$$\begin{aligned} & D_{KL}(f_{R_0} \| f_{R_0+1}) - D_{KL}(f_{R_0} \| f_{R_0-1}) = \\ & \quad \mathbb{E}_{R_0}\left[\log\left\{\frac{N-R_0}{R_0}\frac{N-R_0+1}{R_0+1}\frac{R_0-r}{N-R_0-(n-r)}\frac{R_0+1-r}{N-R_0+1-(n-r)}\right\}\right]. \end{aligned} \tag{A.1}$$

The proof is performed in two steps. In the first one, we show that the above difference is zero when $R_0 = N/2$. The second step shows that expression (A.1) is non-decreasing in R_0. First, we consider the case where $R_0 = N/2$. If $n < N - n \Rightarrow n < N/2$, we have that $r = 0, \ldots, n$ and, consequently, $n - r = n, \ldots, 0$. Therefore, equation (A.1) becomes

$$
\begin{aligned}
&D_{KL}(f_{N/2}\|f_{N/2+1}) - D_{KL}(f_{N/2}\|f_{N/2-1}) = \\
&\qquad \log\left\{\frac{N-N/2}{N/2}\frac{N-N/2+1}{N/2+1}\right\} + \mathbb{E}_{N/2}\left[\log\left\{\frac{N/2-r}{N/2-(n-r)}\right\}\right] \\
&\qquad\qquad + \mathbb{E}_{N/2}\left[\log\left\{\frac{N/2+1-r}{N/2+1-(n-r)}\right\}\right]
\end{aligned}
$$

By applying the symmetry of the Hypergeometric distribution, as discussed above, we have

$$
\mathbb{E}_{N/2}\left[\log\left\{\frac{N/2-r}{N/2-(n-r)}\right\}\right] = 0,
$$

where we have considered that $N - R_0 = N - N/2 = N/2$ and $n - r = 0, \dots, n$. Similarly, we have

$$
\mathbb{E}_{N/2}\left[\log\left\{\frac{N/2+1-r}{N/2+1-(n-r)}\right\}\right] = 0.
$$

Therefore, $D_{KL}(f_{N/2}\|f_{N/2+1}) - D_{KL}(f_{N/2}\|f_{N/2-1}) = 0$. Implying that the divergence $D_{KL}(f_{R_0}\|f_{R_0+1})$ is equal to the divergence $D_{KL}(f_{R_0}\|f_{R_0-1})$. For $n > N - n \Rightarrow n > N/2$, we have $r = n - N/2, \dots, N/2$ and $n - r = N/2, \dots, n - N/2$: (A.1) becomes

$$
\begin{aligned}
&D_{KL}(f_{N/2}\|f_{N/2+1}) - D_{KL}(f_{N/2}\|f_{N/2-1}) = \\
&\qquad \log\left\{\frac{N-N/2}{N/2}\frac{N-N/2+1}{N/2+1}\right\} + \mathbb{E}_{N/2}\left[\log\left\{\frac{N/2-r}{N/2-(n-r)}\right\}\right] \\
&\qquad\qquad\qquad + \mathbb{E}_{N/2}\left[\log\left\{\frac{N/2+1-r}{N/2+1-(n-r}\right\}\right]
\end{aligned}
$$

By symmetry, and considering that $N - R_0 = N - N/2 = N/2$ and therefore that $n - r = N/2, \dots, n - N/2$, we have

$$D_{KL}(f_{N/2}\|f_{N/2+1}) - D_{KL}(f_{N/2}\|f_{N/2-1}) = \log\{1\} + 0 + 0 = 0.$$

This shows that, when $R_0 = N/2$, for any value of n, the Kullback–Leibler divergence from the central point to the distribution with $R = R_0 + 1$ and to the distribution with $R = R_0 - 1$ are equal.

To show that $D_{KL}(f_{R_0}\|f_{R_0+1}) - D_{KL}(f_{R_0}\|f_{R_0-1})$ is non-decreasing, we rearrange the log-term in (A.1) as follow

$$\frac{R_0 - r}{R_0}\frac{R_0 + 1 - r}{R_0 + 1}\frac{N - R_0}{N - R_0 - (n - r)}\frac{N - R_0 + 1}{N - R_0 + 1 - (n - r)}. \tag{A.2}$$

All the terms in the above expression (A.2) are non-decreasing in R_0. Consider the first one. As $(R_0 - r) \le R_0$, we have $(R_0 - r)R_0 \le (R_0 - r) + R_0$. Therefore, $(R_0 - r)(R_0 + 1) \le R_0(R_0 + 1 - r)$, which gives

$$\frac{R_0 - r}{R_0} \le \frac{R_0 + 1 - r}{R_0 + 1}.$$

Similarly, for the second term we have $(R_0 + 1 - r) \le (R_0 + 1)$, and

$$\begin{aligned}(R_0 + 1 - r)(R_0 + 1) + (R_0 + 1 - r) &\le (R_0 + 1 - r)(R_0 + 1) + (R_0 + 1)\\ (R_0 + 1 - r)(R_0 + 1 + 1) &\le (R_0 + 1)(R_0 + 1 - r + 1)\\ \frac{R_0 + 1 - r}{R_0 + 1} &\le \frac{R_0 + 2 - r}{R_0 + 2}.\end{aligned}$$

For the third term, we have $(N - R_0) \ge (N - R_0 - n + r)$. Thus

$$\begin{aligned}(N - R_0)(N - R_0 - n + r) - (N - R_0) &\le (N - R_0)(N - R_0 - n + r)\\ &- (N - R_0 - n + r)\\ (N - R_0)(N - R_0 - n + r - 1) &\le (N - R_0 - n + r)(N - R_0 - 1)\\ \frac{N - R_0}{N - R_0 - (n + r)} &\le \frac{N - (R_0 + 1)}{N - (R_0 + 1) - (n - r)}.\end{aligned}$$

Finally, for the last term, we have $(N - R_0 + 1) \ge (N - R_0 + 1 - n + r)$. Thus

$$
\begin{aligned}
(N-R_0+1)(N-R_0+1-n+r)-(N-R_0+1) \;&\leq \\
&(N-R_0+1)(N-R_0+1-n+r)-(N-R_0+1-n+r) \\
(N-R_0+1)(N-R_0+1-n+r-1) \;&\leq \\
&(N-R_0+1-n+r)(N-R_0+1-1) \\
\frac{N-R_0+1}{N-R_0+1-n+r} \;&\leq \\
&\frac{N-(R_0+1)+1}{N-(R_0+1)+1-n+r},
\end{aligned}
$$

which shows that also the last term is non-decreasing. Note that the above expression is in general strictly increasing, as the equality is met only if and when $r = 0$ or $(n-r) = 0$. From these results, we see that the difference $D_{KL}(f_{R_0}\|f_{R_0+1}) - D_{KL}(f_{R_0}\|f_{R_0-1})$ is increasing in R_0. Given it is zero when $R_0 = N/2$, as shown in the first part of the proof, the statement of the lemma follows. □

Proof of Lemma 4.2

Proof. The proof will be provided for the specific case of a bivariate Hypergeometric distribution, that is where $R = (R_1, R_2, R_3)$. The general case $(d > 3)$ can be derived along the same lines with just more complex notation.

Consider the bivariate Hypergeometric distribution with parameters N, n and $R = (R_1, R_2, R_3)$ and the following form of the probability mass function

$$
f(r|N, R, n) = \frac{\binom{R_1}{r_1}\binom{R_2}{r_2}\binom{R_3}{r_3}}{\binom{N}{n}},
$$

where $R_3 = N - (R_1 + R_2)$ and $r_3 = n - (r_1 + r_2)$. Also, for parameters r_1, r_2 and r_3 we have $\max(0, n-(N-R_1) \leq r_1 \leq \min((n, R_1)))$, $\max(0, n-(N-R_2)) \leq$

$r_2 \leq \min(n, R_2)$ and $\max(0, n-(N-R_3)) \leq r_3 \leq \min(n, N-R_3)$. Given that parameter R_3 depends on R_1 and R_2, once N is fixed, we can consider a two-dimension lattice structure formed by the values of $R_1, R_2 = \{0, 1, \ldots, N\}$ as a representation of the parameter space.

First, we consider the case where $R_1 < N/2$, $R_2 < N/2$. We have seen in 4.2 that n plays an important role; in particular we have to distinguish the case when is below $N/2$ or above it. However, as the following proof is substantially based on the results there obtained, to keep the exposition simple, we do not make the distinction here, being understood that it has to be considered when the prior is actually implemented.

If we allow to vary one of the two parameters R_1, R_2 at a time, the bivariate Hypergeometric can be interpreted as a univariate Hypergeometric, with bins $(R_1, N-R_1)$ and $(R_2, N-R_2)$, respectively. Given the results in 4.2, we have

$$\begin{aligned}
D_{KL}\left(f(R_1,R_2)\|f(R_1+1,R_2)\right) &< D_{KL}\left(f(R_1,R_2)\|f(R_1-1,R_2)\right)\\
D_{KL}\left(f(R_1,R_2)\|f(R_1,R_2+1)\right) &< D_{KL}\left(f(R_1,R_2)\|f(R_1,R_2-1)\right),
\end{aligned}$$

where, for example, the distribution $f_{N,R,n}$ (with $R=(R_1,R_2)$) has been indicated as $f(R_1,R_2)$, as we consider only distributions that vary in the value of these parameters (being N, and n known and common), and this represents a simpler notation. Therefore, at each point (R_1,R_2), the following three divergences (obtained by feeding expression (4.18) with the appropriate parameter values) have to be compared, in order to find the smallest one

$$\begin{aligned}
D_{KL}(f(R_1,R_2)\|f(R_1+1,R_2)) &= \log\left\{\frac{R_3}{R_1+1}\right\} + \mathbb{E}\left[\log\left\{\frac{R_1+1-r_1}{R_3-r_3}\right\}\right]\\
D_{KL}(f(R_1,R_2)\|f(R_1,R_2+1)) &= \log\left\{\frac{R_3}{R_2+1}\right\} + \mathbb{E}\left[\log\left\{\frac{R_2+1-r_2}{R_3-r_3}\right\}\right]\\
D_{KL}(f(R_1,R_2)\|f(R_1+1,R_2+1)) &= \log\left\{\frac{R_3(R_1-1)}{(R_1+1)(R_2+1)}\right\}\\
&\quad + \mathbb{E}\left[\log\left\{\frac{(R_1+1-r_1)(R_2+1-r_2)}{(R_3-r_3)(R_3-r_3-1)}\right\}\right],
\end{aligned}$$

where the expectation are taken with respect to $f(R_1, R_2)$.

Given the lattice structure of the parameter space (R_1, R_2), it is logic to assume that the divergence $D_{KL}(f(R_1, R_2)\|f(R_1+1, R_2+1))$ is not smaller than any of the other two. Therefore, the comparison has to be carried forward between $D_{KL}(f(R_1, R_2)\|f(R_1+1, R_2))$ and $D_{KL}(f(R_1, R_2)\|f(R_1, R_2+1))$. Considering the difference of the two

$$\begin{aligned} D_{KL}(f(R_1, R_2)\|f(R_1+1, R_2)) - D_{KL}(f(R_1, R_2)\|f(R_1, R_2+1)) &= \\ \mathbb{E}\left[\log\left\{\frac{R_2+1}{R_1+1}\frac{R_1+1-r_1}{R_2+1-r_2}\right\}\right] & \end{aligned} \tag{A.3}$$

If $R_1 = R_2$, the difference (A.3) is zero. In fact, replacing R_1 and R_2 with R, r_1 and r_2 with r in (A.3), the right-hand-side becomes

$$\log\left\{\frac{R+1}{R+1}\right\} + \mathbb{E}\left[R+1-r\right] - \mathbb{E}\left[R+1-r\right] = 0$$

If we fix either parameter R_1 or R_2, the log-expression in (A.3) is increasing with respect to the other parameter. Say we fix R_1, then when $R_1 > R_2$ the minimum divergence is $D_{KL}(f(R_1, R_2)\|f(R_1+1, R_2))$ and when $R_1 < R_2$, the minimum divergence is $D_{KL}(f(R_1, R_2)\|f(R_1, R_2+1))$. By combing this result with the one obtained in 4.2, note that $R_2 > N/2$ implies $D_{KL}(f(R_1, R_2)\|f(R_1, R_2+1)) > D_{KL}(f(R_1, R_2)\|f(R_1, R_2-1))$, therefore is the divergence on the right-hand-side that has to be compared with $D_{KL}(f(R_1, R_2)\|f(R_1+1, R_2))$. If we fix R_2 and let R_1 vary, we obtain analogous results. We can then summarise the identification the smallest Kullback–Leibler divergence in the following three cases:

1. $R_1 < N/2$ and $R_2 < N/2$:
    - if $R_1 > R_2 \Rightarrow D_{KL}(f(R_1, R_2)\|f(R_1+1, R_2))$
    - if $R_1 < K_2 \Rightarrow D_{KL}(f(R_1, R_2)\|f(R_1, R_2+1))$
2. $R_1 < N/2$ and $R_2 > N/2$:

- in this case the minimum divergence is $D_{KL}(f(R_1, R_2)\|f(R_1, R_2-1))$, as the distance from R_1 to $N/2$ is always larger than the distance from R_2 to $N/2$, that is $N/2 - R_1 > R_2 - N/2$;
- if $N/2 - R_1 = R_2 - N/2$ the divergences $D_{KL}(f(R_1, R_2)\|f(R_1+1, R_2))$ and $D_{KL}(f(R_1, R_2)\|f(R_1, R_2-1))$ are equal.

3. $R_1 > N/2$ and $R/2 < N/2$:
   - the minimum divergence is $D_{KL}(f(R_1, R_2)\|f(R_1-1, R_2))$ as $R_1 - N/2 < N/2 - R_2$;
   - if $R_1 - N/2 = N/2 - R_2$, then $D_{KL}(f(R_1, R_2)\|f(R_1-1, R_2))$ is equal to $D_{KL}(f(R_1, R_2)\|f(R_1, R_2+1))$

□

Proof of Theorem 4.3

Proof. By applying the standard definition of conditional probability, we have

$$\frac{\pi(n_0|x_1, \ldots, x_k)}{\pi(n \geq n_0|x_1, \ldots, x_k)} \to 1,$$

which is equivalent to

$$\frac{\pi(n \geq n_0|x_1, \ldots, x_k)}{\pi(n_0|x_1, \ldots, x_k)} \to 1. \tag{A.4}$$

The expression (A.4) can be written in the following way

$$\begin{aligned}
&\frac{\sum_{j=0}^{\infty} \left\{ \prod_{i=1}^{k} \binom{n_0+j}{x_i} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(x_i+a)\Gamma(n_0+j-x_i+b)}{\Gamma(n_0+j+a+b)} \times \pi(n_0+j) \right\}}{\prod_{i=1}^{k} \binom{n_0}{x_i} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(x_i+a)\Gamma(n_0-x_i+b)}{\Gamma(n_0+a+b)} \times \pi(n_0)} \\
&= 1 + \sum_{j=1}^{\infty} \left\{ \frac{\binom{n_0+j}{x_i}}{\binom{n_0}{x_i}} \frac{\Gamma(n_0+j-x_i+b)}{\Gamma(n_0+j+a+b)} \frac{\Gamma(n_0+a+b)}{\Gamma(n_0-x_i+b)} \times \frac{\pi(n_0+j)}{\pi(n_0)} \right\}
\end{aligned} \tag{A.5}$$

The ratio of the priors in (A.5), $\pi(n_0+j)/\pi(n_0)$, converges to zero as j tends to infinity. Thus, we need to show that

$$\mathbb{E}_{n_0}\left[\frac{\binom{n_0+j}{x_i}}{\binom{n_0}{x_i}}\frac{\Gamma(n_0+j-x_i+b)}{\Gamma(n_0+j+a+b)}\frac{\Gamma(n_0+a+b)}{\Gamma(n_0-x_i+b)}\right]>1,$$

to prove that the second term in the right-hand-side of (A.5) converges to zero. Note that $\mathbb{E}_{n_0}$ is the expectation with respect to p_{n_0}. As we have

$$\begin{aligned}&\mathbb{E}_{n_0}\left[\frac{\binom{n_0+j}{x_i}}{\binom{n_0}{x_i}}\frac{\Gamma(n_0+j-x_i+b)}{\Gamma(n_0+j+a+b)}\frac{\Gamma(n_0+a+b)}{\Gamma(n_0-x_i+b)}\right]\\ &=\sum_{x=0}^{n_0}\left\{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\Gamma(x+a)\frac{\Gamma(n_0+j-x+b)}{n_0+j+a+b}\frac{(n_0+j)!}{(n_0+j-x)!\,x!}\right\},\end{aligned}$$

we need to show that

$$\lim_{j\to\infty}\sum_{x=0}^{n_0}\left\{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\Gamma(x+a)\frac{\Gamma(n_0+j-x+b)}{\Gamma(n_0+j+a+b)}\frac{(n_0+j)!}{(n_0+j-x)!\,x!}\right\}=0. \tag{A.6}$$

We have

$$\begin{aligned}\frac{\Gamma(n_0+j-x+b)}{\Gamma(n_0+j+a+b)}\frac{(n_0+j)!}{(n_0+j-x)!\,x!} &= \frac{(n_0+j-x+b-1)!}{(n_0+j+a+b-1)!}\frac{(n_0+j)!}{(n_0+j-x)!}\\ &= \frac{(n_0+j-x+(b-1))!}{(n_0+j+(a+b-1))!}\frac{(n_0+j)!}{(n_0+j-x)!},\end{aligned}$$

which, given that we assume $a,b>1$, will always tend to zero for $j\to\infty$, as the power of j at the denominator will always be higher than the one at the numerator. Thus, the limit in (A.6) is proved and, therefore, the theorem proof follows. □

Proof of Lemma 4.3

Proof. Consider the Kullback–Leibler divergence between f_{n_0} and f_{n_0+m}, with $m\geq 1$; to prove that $D_{KL}(f_{n_0}\|f_{n_0+m})$ is minimum when $m=1$, it is sufficient to prove that $D_{KL}(f_{n_0} \| f_{n_0+1}) < D_{KL}(f_{n_0} \| f_{n_0+2})$. Hence, we need to show that

$$-\mathbb{E}\left[\log\binom{n_0+1}{x}\right] - \log(1-p) < -\mathbb{E}\left[\log\binom{n_0+2}{x}\right] - 2\log(1-p),$$

i.e. that

$$\mathbb{E}\left[\log\left(\frac{(n_0+2)!}{(n_0+2-x)!}\right)\right] - \mathbb{E}\left[\log\left(\frac{(n_0+1)!}{(n_0+1-x)!}\right)\right] < -\log(1-p),$$

which holds if

$$\mathbb{E}\left[\log\left(\frac{n_0+2}{n_0+2-x}\right)\right] < -\log(1-p).$$

Since $[(n_0+2)/(n_0+2-x)] < [(n_0+1)/(n_0+1-x)]$, we aim now to show that

$$\mathbb{E}\left[\log\left(\frac{n_0+1}{n_0+1-x}\right)\right] < -\log(1-p).$$

Now, applying Jensen's inequality, we have

$$\mathbb{E}\left[\log\left(\frac{n_0+1}{n_0+1-x}\right)\right] \leq \log\left(\mathbb{E}\left[\frac{n_0+1}{n_0+1-x}\right]\right).$$

Simple algebra gives

$$\mathbb{E}\left[\frac{n_0+1}{n_0+1-x}\right] = \frac{1}{1-p}\sum_{x=0}^{n_0}\binom{n_0+1}{x}p^x(1-p)^{n_0+1-x} \leq \frac{1}{1-p},$$

and hence the result follows. □

The following Lemma A.1 is functional to the proof of Theorem 4.4 in Section 4.5.1. The proof follows.

Lemma A.1. *For $n = 1, 2, \ldots$, $x = 0, 1, \ldots, n$ and $p \in (0, 1)$, we have*

$$\frac{1}{n+1} \prod_{x=0}^{n} \left\{ (n+1-x)^{\binom{n}{x} p^x (1-p)^{n-x}} \right\} < 1. \tag{A.7}$$

Proof. Re-arrange (A.7) and take the logarithm of both sides

$$\sum_{x=0}^{n} \left\{ \log(n+1-x) \times \binom{n}{x} p^x (1-p)^{n-x} \right\} \leq \log(n+1). \tag{A.8}$$

As for $x = n$ we have $\log(n+1-x) = 0$, the left-hand-side of (A.8) can be written as

$$\sum_{x=0}^{n-1} \left\{ \log(n+1-x) \times \binom{n}{x} p^x (1-p)^{n-x} \right\}.$$

By replacing the n terms $\log(n+1-x)$ above with $\log(n+1)$, we have

$$\sum_{x=0}^{n} \left\{ \log(n+1-x) \times \binom{n}{x} p^x (1-p)^{n-x} \right\} \leq \log(n+1) \times \sum_{x=0}^{n-1} \left\{ \binom{n}{x} p^x (1-p)^{n-x} \right\},$$

and because

$$\begin{aligned} \log(n+1) \times \sum_{x=0}^{n-1} \left\{ \binom{n}{x} p^x (1-p)^{n-x} \right\} &< \log(n+1) \\ \sum_{x=0}^{n-1} \left\{ \binom{n}{x} p^x (1-p)^{n-x} \right\} &< 1, \end{aligned} \tag{A.9}$$

which is always true as the left-hand-side of (A.9) is the sum of the probabilities of a binomial distribution except the last one (i.e. when $x = n$). Therefore, the relation in (A.8) is satisfied. □

Proof of Theorem 4.5

Proof. By applying the standard definition of conditional probability, the (4.5) becomes

$$\frac{\pi(n_0|x_1,\ldots,x_k)}{\pi(n \geq n_0|x_1,\ldots,x_k)} \to 1,$$

which is equivalent to

$$\frac{\pi(n \geq n_0|x_1,\ldots,x_k)}{\pi(n_0|x_1,\ldots,x_k)} \to 1. \tag{A.10}$$

Now, the expression in (A.10) can be written as

$$\begin{aligned}&\frac{\sum_{j=0}^{\infty}\left\{\prod_{i=1}^{k}\binom{n_0+j}{x_i}(1-p)^{n_0+j} \times \pi(n_0+j)\right\}}{\prod_{i=1}^{k}\binom{n_0}{x_i}(1-p)^{n_0} \times \pi(n_0)} \\ &= 1 + \sum_{j=1}^{\infty}\left\{\prod_{i=1}^{k}\frac{\binom{n_0+j}{x_i}}{\binom{n_0}{x_i}}(1-p)^j \times \frac{\pi(n_0+j)}{\pi(n_0)}\right\}.\end{aligned} \tag{A.11}$$

To show that the second term on the right-hand-side of (A.11) converges to zero, we consider its expected value with respect to x and, as the ratio $\pi(n_0+j)/\pi(n_0)$ converges to zero as $j \to \infty$. The following relation has to hold

$$\mathbb{E}\left[\frac{\binom{n_0+j}{x}}{\binom{n_0}{x}}(1-p)^j\right] < 1,$$

with

$$\begin{aligned}\mathbb{E}\left[\frac{\binom{n_0+j}{x}}{\binom{n_0}{x}}(1-p)^j\right] &= \sum_{x=0}^{n_0}\left\{\binom{n_0}{x}p^x(1-p)^{n_0-x}\frac{\binom{n_0+j}{x}}{\binom{n_0}{x}}(1-p)^j\right\} \\ &= \sum_{x=0}^{n_0}\left\{p^x(1-p)^{n_0-x}\binom{n_0+j}{x}(1-p)^j\right\} \\ &= \sum_{x=0}^{n_0}\left\{p^x(1-p)^{n_0-x}\frac{(n_0+j)!}{x!\,(n_0-x)!}(1-p)^j\right\}.\end{aligned}$$

We have

$$\lim_{j\to\infty}\sum_{x=0}^{n_0}\left\{\binom{n_0+j}{x}p^x(1-p)^{n_0+j-x}\right\} = \lim_{j\to\infty}\sum_{x=0}^{n_0}\left\{p^x(1 - p)^{n_0-x}\frac{1}{x!}(1-p)^j\frac{(n_0+j)!}{(n_0+j-x)!}\right\}.$$

As

$$\frac{(n_0+j)!}{(n_0+j-x)!} \leq \frac{(n_0+j)!}{j!} = j \times (j+1) \times \cdots \times (n_0+j) \leq (n_0+j)^{n_0},$$

we have

$$\sum_{x=0}^{n_0}\left\{p^x(1-p)^{n_0-x}\frac{1}{x!}(1-p)^j\frac{(n_0+j)!}{(n_0+j-x)!}\right\} \leq \sum_{x=0}^{n_0}\left\{p^x(1-p)^{n_0-x}\frac{1}{x!}(1-p)^j(n_0+j)^{n_0}\right\}.$$

And

$$\begin{aligned}&\lim_{j\to\infty}\left\{(1-p)^j(n_0+j)^{n_0}\right\}\\ &= \lim_{j\to\infty}\exp\left\{-j[-\log(1-p)] + n_0 \times \log(n_0+j)\right\}\\ &= 0,\end{aligned}$$

given that, for $j \to \infty$, j dominates $\log j$. We can then conclude that

$$\lim_{j\to\infty}\sum_{x=0}^{n_0}\left\{\binom{n_0+j}{x}p^x(1-p)^{n_0+j-x}\right\} = 0,$$

which implies that the second term of the right-hand-side of (A.11) converges to zero and, therefore, (A.10) holds, proving that the posterior is consistent. □

Appendix B

Mathematical Support for Chapter 6

In this appendix, we include the details of the derivation of the Kullback–Leibler divergences for the models discussed in Chapter 6. We also analytically prove how the minimum is attained, where applicable.

B.1 Theorems and Proofs

Theorem B.1 (Poisson–Geometric Kullback–Leibler minimum). *Consider a Poisson distribution with rate parameter θ, $f_1(x|\theta) = \theta^x e^{-\theta}/x!$, and a Geometric distribution with parameter ϕ, $f_2(x|\phi) = \phi(1-\phi)^x$. The Kullback–Leibler divergence between the Poisson and the Geometric, indicated by $D_{KL}(f_1(x|\theta)\|f_2(x|\phi))$, attains its minimum, with respect to ϕ, for $\phi = 1/(1+\theta)$. The Kullback–Leibler divergence between the Geometric and the Poisson, $D_{KL}(f_2(x|\phi)\|f_1(x|\theta))$, attains its minimum, with respect to θ, for $\theta = (1-\phi)/\phi$.*

Proof. To prove the first result, we have

$$D_{KL}(f_1(x|\theta)\|f_2(x|\phi)) \quad = \quad \sum_{x=0}^{\infty} \frac{\theta^x}{x!} e^{-\theta} \left\{ x \log\theta - \log x! - \theta - \log\phi - x\log(1-\phi) \right\}$$

$$\begin{aligned} = \quad & \theta \log \theta - \sum_{x=0}^{\infty} \left\{ \log x! \frac{\theta^x}{x!} e^{-\theta} \right\} - \theta - \log \phi \\ & -\theta \log(1-\phi). \end{aligned}$$

Differentiating with respect to ϕ, we have $\partial/\partial\phi\{D_{KL}(f_1(x|\theta)\|f_2(x|\phi))\} = -1/\phi + \theta/(1-\phi)$. By equating to zero, we have the result stated. For the second result, we consider

$$\begin{aligned} D_{KL}(f_2(x|\phi)\|f_1(x|\theta)) &= \sum_{x=0}^{\infty} \phi(1-\phi)^x \left\{ \log \phi + x \log(1-\phi) - x \log \theta + \log x! + \theta \right\} \\ &= \log \phi + \frac{1-\phi}{\phi} \log(1-\phi) - \frac{1-\phi}{\phi} \log \theta + \theta + \sum_{x=0}^{\infty} \left\{ \phi(1-\phi)^x \log x! \right\}. \end{aligned}$$

To find the minimum with respect to θ: $\partial/\partial\theta\{D_{KL}(f_2(x|\phi)\|f_1(x|\theta))\} = 1 - (1-\phi)/(\phi\theta)$, and by setting equal to zero we obtain the second statement of the theorem. □

Remark B.1. *Assume random variable x has a Weibull distribution with parameters λ and κ. Then, we have the following results (Johnson and Kotz, 1970)*

$$\begin{aligned} &\mathbb{E}(x) = \lambda\Gamma(1+1/\kappa) \\ &\mathbb{E}(\log x) = \log \lambda - \gamma/\kappa \\ &\mathbb{E}(\log^2 x) = \pi^2/(6\kappa^2) + (\log \lambda - \gamma/\kappa)^2 \\ &\mathbb{E}(x^\kappa) = \lambda^\kappa \\ &Var(\log x) = \pi^2/(6\kappa^2). \end{aligned}$$

Remark B.2. *If random variable x has a Log-normal distribution with parameters μ and τ, then the following results are true (Johnson and Kotz, 1970)*

$$\mathbb{E}(x) = \exp\{\mu + 1/(2\tau)\}$$

$$\begin{aligned}
&\mathbb{E}(\log x) = \mu\\
&\mathbb{E}(\log^2 x) = 1/\tau + \mu^2\\
&\mathbb{E}(x^\kappa) = \exp\{\kappa^2/(2\tau) + \mu\kappa\}.
\end{aligned}$$

Theorem B.2 (Weibull–Log-normal Kullback–Leibler minimum). *Consider the Weibull density function* $f_1(x|\lambda,\kappa) = \kappa/\lambda(x/\lambda)^{k-1}\exp(-x/\lambda)^\kappa$ *and the Log-normal density function* $f_2(x|\mu,\tau) = x^{-1}\{\tau/(2\pi)\}^{1/2}\exp\{-\tau(\log x - \mu)^2/2\}$. *The Kullback–Leibler divergence* $D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau))$ *is minimised for* $\mu = \mathbb{E}(\log x) = \log\lambda - \gamma/\kappa$ *and* $\tau = 1/Var(\log x) = 6(\kappa/\pi)^2$, *where the expectation is taken with respect to* $f_1(x|\lambda,\kappa)$. *The Kullback–Leibler divergence* $D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa))$ *attains its minimum at* $\lambda = \mathbb{E}(x^\kappa)^{1/\kappa} = \exp\{1/(2\sqrt{\tau}) + \mu\}$ *and* $\kappa = \sqrt{\tau}$.

Proof. First we see that

$$\begin{aligned}
D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau)) &= \int_0^\infty f_1(x|\lambda,\kappa)\Big\{\log\kappa\\
&- \log\lambda + (\kappa-1)\log x - (\kappa-1)\log\lambda - \frac{x^\kappa}{\lambda^\kappa} + \log x -\\
&\frac{1}{2}\log\tau + \frac{1}{2}\log(2\pi) + \frac{1}{2}\tau(\log^2 x - 2\mu\log\ +\mu^2)\Big\}\ dx\\
&= \log\kappa + \kappa\,\mathbb{E}(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa) - \frac{1}{2}\log\tau + \frac{1}{2}\log(2\pi) + \frac{1}{2}\tau\,\mathbb{E}(\log^2 x)\\
&- \tau\mu\,\mathbb{E}(\log x) + \frac{1}{2}\tau\mu^2.
\end{aligned}$$

To find the minimum: $\partial/\partial\mu\{D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau))\} = -\tau\,\mathbb{E}(\log x) + \tau\mu$, which is solved for $\mu = \mathbb{E}(\log x) = \log\lambda - \gamma/\kappa$; $\partial/\partial\tau\{D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau))\} = -1/(2\tau) + 1/\{2\,\mathbb{E}(\log^2 x)\} - \mu\,\mathbb{E}(\log x) + \mu^2/2$, which, by setting $\mu = \mathbb{E}(\log x)$, is solved for $\tau = 1/Var(\log x) = 6(\kappa/\pi)^2$. This proves the first statement of the theorem. For the second result, we have

$$D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa)) = \int_0^\infty f_2(x|\mu,\tau)\Bigg\{-\log x$$
$$+\frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\log^2 x + \tau\mu\log x - \frac{1}{2}\tau\mu^2 -$$
$$\log\kappa + \log\lambda - (\kappa-1)\log x + (\kappa-1)\log\lambda + \frac{x^\kappa}{\lambda^\kappa}\Bigg\}\,dx$$
$$=\frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\,\mathbb{E}(\log^2 x) + \tau\mu\,\mathbb{E}(\log x) - \frac{1}{2}\tau\mu^2 - \log\kappa - \kappa\,\mathbb{E}(\log x)$$
$$+\kappa\log\lambda + \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa).$$

To minimise: $\partial/\partial\lambda\{D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa))\} = \kappa/\lambda - \kappa\,\mathbb{E}(x^\kappa)/\lambda^{\kappa+1}$, has solution $\lambda = \mathbb{E}(x^\kappa)^{1/\kappa} = \exp\{1/(2\sqrt{\tau})+\mu\}$; then $\partial/\partial\kappa\{D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa))\} = -1/\kappa + \kappa/\tau$, where we have considered $\mathbb{E}(\log x) = \mu$ and $\mathbb{E}(x^\kappa) = 1/\lambda^\kappa$, has solution $\kappa = \sqrt{\tau}$, proving the second statement of the theorem. □

Remark B.3. *If random variable x has a Gamma distribution with parameters α and β, that if $f(x|\alpha,\beta) = \beta^\alpha x^{\alpha-1}\exp(-\beta x)/\Gamma(\alpha)$, then*

$$\mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$$
$$\mathbb{E}(x^\kappa) = \beta^{-\kappa}\Gamma(\kappa+\alpha)/\Gamma(\alpha),$$

refer to Johnson and Kotz (1970).

Theorem B.3 (Weibull–Gamma Kullback–Leibler minimum). *Consider the density $f_1(x|\lambda,\kappa)$, which has a Weibull distribution with λ as the scale parameter and κ as the shape parameter. Consider also the Gamma density $f_3(x|\alpha,\beta)$ where α and β are, respectively, the shape and the rate parameter. The Kullback-Leibler divergence $D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta))$ attains its minimum for $\mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$ and $\mathbb{E}(x) = \alpha/\beta$. The divergence $D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa))$ is minimised for $\lambda = \mathbb{E}(x^\kappa)^{1/\kappa}$ and $\Psi(\kappa+\alpha) - 1/\kappa = \Psi(\alpha)$.*

Proof. For the first statement, we have

$$D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta)) = \int_0^\infty f_1(x|\lambda,\kappa)\Bigg\{\log\kappa - \log\lambda + (\kappa-1)\log x$$

$$
\begin{aligned}
&- (\kappa - 1)\log\lambda - \frac{x^\kappa}{\lambda^\kappa} - \alpha\log\beta + \log\Gamma(\alpha) - (\alpha - 1)\log x + \beta x\Bigg\}\,dx \\
&\qquad = \log\kappa + \kappa\,\mathbb{E}(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa) - \alpha\log\beta \\
&\qquad\quad + \log\Gamma(\alpha) - \alpha\,\mathbb{E}(\log x) + \beta\,\mathbb{E}(x).
\end{aligned}
$$

By differentiating, we find $\partial/\partial\alpha\{D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta))\} = -\log\beta + \Psi(\alpha) - \mathbb{E}(\log x)$ and $\partial/\partial\beta\{D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta))\} = -\alpha/\beta + \mathbb{E}(x)$. The resulting system formed by equations $\mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$ and $\mathbb{E}(x) = \alpha/\beta$ has to be solved numerically (i.e. Newton-Raphson method). For the second statement, we have

$$
\begin{aligned}
D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa)) &= \int_0^\infty f_3(x|\alpha,\beta)\Bigg\{\alpha\log\beta - \log\Gamma(\alpha) + (\alpha - 1)\log x \\
&\quad - \beta x - \log\kappa + \log\lambda - (\kappa - 1)\log x + (\kappa - 1)\log\lambda + \frac{x^\kappa}{\lambda^\kappa}\Bigg\}\,dx \\
&= \alpha\log\beta - \log\Gamma(\alpha) + \alpha\,\mathbb{E}(\log x) - \beta\,\mathbb{E}(x) - \log\kappa \\
&\quad - \kappa\,\mathbb{E}(\log x) + \kappa\log\lambda + \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa),
\end{aligned}
$$

which, considering $\partial/\partial\lambda\{D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa))\} = -\kappa/\lambda + \kappa/\lambda^{\kappa+1}\mathbb{E}(x^\kappa)$, and κ, $\partial/\partial\kappa\{D_{KL}(f_3(x|\alpha,\beta)\|f_1(x|\lambda,\kappa))\} = -1/\kappa - \Psi(\alpha) - \Psi(\kappa+\alpha)$, results in a system with equations $\lambda = \mathbb{E}(x^\kappa)$ and $\Psi(\kappa + \alpha) - 1/\kappa = \Psi(\alpha)$. The system has to be solved with numerical methods.

□

Theorem B.4 (Log-normal–Gamma Kullback–Leibler minimum). *If we consider densities $f_2(x|\mu,\tau)$ and $f_3(x|\alpha,\beta)$, $D_{KL}(f_2(x|\mu,\tau)\|f_3(x|\alpha,\beta))$ it is minimised when, simultaneously, we have $\mathbb{E}(x) = \alpha/\beta$ and $\mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$. The divergence $D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau))$, attains its minimum for $\mu = \mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$ and $\Psi(\alpha) - \log\beta$ and $\tau = 1/Var(\log x) = 1/\Psi'(\alpha)$.*

Proof. The Kullback–Leibler divergence between the Log-normal and the Gamma densities is given by

$$D_{KL}(f_2(x|\mu,\tau)\|f_3(x|\alpha,\beta)) = \int_0^\infty f_2(x|\mu,\tau)\Big\{-\log x + \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi)$$
$$-\frac{1}{2}\tau(\log^2 x - 2\mu\log x + \mu^2) - \alpha\log\beta + \log\Gamma(\alpha) - (\alpha-1)\log x + \beta x\Big\}\,dx$$
$$= \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\,\mathbb{E}(\log^2 x) - \frac{1}{2}\tau\mu\,\mathbb{E}(\log x)$$
$$-\frac{1}{2}\tau\mu^2 - \alpha\log\beta + \log\Gamma(\alpha) - \alpha\,\mathbb{E}(\log x) + \beta\,\mathbb{E}(x).$$

Differentiating with respect to α, we have $\partial/\partial\alpha\{D_{KL}(f_2(x|\mu,\tau)\|f_3(x|\alpha,\beta))\} = -\log\beta + \Psi(\alpha) - \mathbb{E}(\log x)$, which is solved for $\mathbb{E}(\log x) = \Psi(\alpha) - \log\beta = \mu$. Differentiating with respect to β, we have $\partial/\partial\beta\{D_{KL}(f_2(x|\mu,\tau)\|f_3(x|\alpha,\beta))\} = -\alpha/\beta + \mathbb{E}(x)$, with solution $\mathbb{E}(x) = \alpha/\beta = \exp\{\mu + 1/(2\tau)\}$. The minimum is obtained by solving the system of equations with numerical methods (i.e. Newton-Raphson). We now consider the Kullback–Leibler divergence between the Gamma and the Log-normal density

$$D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau)) = \int_0^\infty f_3(x|\alpha,\beta)\Big\{\alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log x$$
$$-\beta x + \log x - \frac{1}{2}\log\tau + \frac{1}{2}\log(2\pi) + \frac{1}{2}\tau(\log^2 x 2\mu\log x + \mu^2)\Big\}\,dx$$
$$= \alpha\log\beta - \log\Gamma(\alpha) + \alpha\,\mathbb{E}[\log x] - \beta\,\mathbb{E}[x] - \frac{1}{2}\log\tau$$
$$+\frac{1}{2}\log(2\pi) + \frac{1}{2}\tau\,\mathbb{E}[\log^2 x] - \tau\mu\,\mathbb{E}[\log x] + \frac{1}{2}\tau\mu^2.$$

The divergence is minimised by considering $\partial/\partial\mu\{D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau))\} = -\tau\,\mathbb{E}(\log x) + \tau\mu$, and $\partial/\partial\tau\{D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau))\} = -1/(2\tau) + \mathbb{E}(\log^2 x)/2 - \mu\,\mathbb{E}(\log x) + \mu^2/2$. Numerically, the system that has to be solved to find the minimum, is formed by equation $\mu = \mathbb{E}(\log x) = \Psi(\alpha) - \log\beta$ and $\tau = 1/Var(\log x) = 1/\Psi'(\alpha)$.

□

Theorem B.5 (t–Normal Kullback–Leibler minimum). *The Kullback–Leibler divergence between a t density with location parameter θ, scale parameter λ and* *number of degrees of freedom ν, denoted by $f_2(x|\theta,\lambda,\nu)$, and a Normal density with mean μ and precision τ, denote by $f_1(x|\mu,\tau)$, is minimised for $\mu=\theta$ and $\tau=\lambda$.*

Proof. Consider the divergence between the two densities

$$\begin{aligned}
D_{KL}(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau)) &= \int_{-\infty}^{\infty} f_2(x|\theta,\lambda,\nu)\Big\{\log\Gamma\left(\frac{\nu+1}{2}\right)-\log\Gamma\left(\frac{\nu}{2}\right)\\
&\quad+\frac{1}{2}\log\lambda-\frac{1}{2}\log(\nu\pi)-\frac{\nu+1}{2}\log\left(1+\frac{\lambda}{\nu}(x-\theta)^2\right)-\frac{1}{2}\log\tau\\
&\qquad+\frac{1}{2}\log(2\pi)+\frac{\tau}{2}(x-\mu)^2\Big\}\,dx\\
&=\log\Gamma\left(\frac{\nu+1}{2}\right)-\log\Gamma\left(\frac{\nu}{2}\right)+\frac{1}{2}\log\lambda-\frac{1}{2}\log\nu-\frac{\nu+1}{2}\\
&\quad\mathbb{E}\left\{\log\left(1+\frac{\lambda}{\nu}(x-\theta)^2\right)\right\}-\frac{1}{2}\log\tau+\frac{1}{2}\log 2+\frac{1}{2}\tau\,\mathbb{E}(x^2)\\
&\quad-\tau\mu\,\mathbb{E}(x)+\frac{1}{2}\tau\mu^2.
\end{aligned}$$

Differentiating, we have $\partial/\partial\mu\{D_{KL}(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau))\} = -\tau\,\mathbb{E}(x)+\tau\mu$, which is solved for $\mu=\mathbb{E}(x)=\theta$; and $\partial/\partial\tau\{D_{KL}(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau))\} = -1/(2\tau)+\mathbb{E}(x^2)/2-\mu\,\mathbb{E}(x)+\mu^2/2$, which is solved for $\tau=1/Var(x)=\lambda$. □

Theorem B.6 (Weibull–Log-normal–Gamma–Exponential Kullback–Leibler minimum)**.** *If we consider the Kullback–Leibler divergence between either a Weibull, a Log-normal and a Gamma distribution, and an Exponential distribution with rate parameter θ, each divergence is minimised for $\theta=1/\mathbb{E}(x)$.*

Proof. If we consider the three divergences, we have

$$\begin{aligned}
D_{KL}(f_1(x|\lambda,\kappa)\|f_4(x|\theta)) &= \int_0^{\infty} f_1(x|\lambda,\kappa)\Big\{\log\kappa-\log\kappa+(\kappa-1)\log x\\
&\quad-(\kappa-1)\log\lambda-\frac{x^\kappa}{\lambda^\kappa}-\log\theta+\theta x\Big\}\,dx
\end{aligned}$$

$$= \log\kappa + \kappa\,\mathbb{E}(\log x) - \mathbb{E}(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}(x^\kappa) - \log\theta + \theta\,\mathbb{E}(x),$$

and

$$\begin{aligned} D_{KL}(f_2(x|\mu,\tau)\|f_4(x|\theta)) = &\int_0^\infty f_2(x|\mu,\tau)\Big\{-\log x + \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) \\ &- \frac{1}{2}\tau(\log x-\mu)^2 - \log\theta + \theta x\Big\}\,dx \\ &= -\mathbb{E}(\log x) + \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\,\mathbb{E}(\log^2 x) \\ &+ \tau\mu\,\mathbb{E}(\log x) - \frac{1}{2}\tau\mu^2 - \log\theta + \theta\,\mathbb{E}(x), \end{aligned}$$

and

$$\begin{aligned} D_{KL}(f_3(x|\alpha,\beta)\|f_4(x|\theta)) = &\int_0^\infty f_3(x|\alpha,\beta)\Big\{\alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log x \\ &- \beta x - \log\theta + \theta x\Big\}\,dx \\ &= \alpha\log\beta - \log\Gamma(\alpha) + \alpha\,\mathbb{E}(\log x) - \mathbb{E}(\log x) - \beta\,\mathbb{E}(x) \\ &- \log\theta + \theta\,\mathbb{E}(x). \end{aligned}$$

The derivative of each divergence with respect to θ has the result $-1/\theta + \mathbb{E}(x)$, which has solution $\theta = 1/\mathbb{E}(x)$. In particular, for the Weibull distribution, the divergence between the two densities is attained for $\theta = 1/\{\lambda\Gamma(1+1/\kappa)\}$; for the Log-normal, the minimum is at $\theta = \exp\{-\mu - 1/(2\tau)\}$. And, for the Gamma, the minimum is attained at $\theta = \beta/\alpha$.

□

Remark B.4. *If the random variable x has an Exponential distribution with parameter θ, then*

$$\mathbb{E}(\log^2 x) = (\log\theta + \gamma)^2 + \pi^2/6$$

$$\mathbb{E}(\log x) = \pi^2/6.$$

refer to Johnson and Kotz (1970).

Theorem B.7 (Exponential–Log-normal Kullback–Leibler minimum). *Consider the Exponential density with parameter θ and the Log-normal density with parameters μ and τ. The Kullback–Leibler divergence $D_{KL}(f_4(x|\theta)\|f_2(x|\mu,\tau))$ is minimised for $\mu = \mathbb{E}(\log x) = -\log\theta - \gamma$ and $\tau = 1/Var(\log x) = 6/\pi^2$.*

Proof. The divergence between the two densities is given by

$$\begin{aligned} D_{KL}(f_4(x|\theta)\|f_2(x|\mu,\tau)) &= \int_0^\infty f_4(x|\theta)\left\{\log\theta - \theta x + \log x - \frac{1}{2}\log\tau\right. \\ &\quad \left. +\frac{1}{2}\log(2\pi) + \frac{1}{2}\tau(\log x - \mu)^2\right\}\, dx \\ &= \log\theta - \theta\,\mathbb{E}(x) + \mathbb{E}(\log x) - \frac{1}{2}\log\tau + \frac{1}{2}\log(2\pi) \\ &\quad +\frac{1}{2}\tau\,\mathbb{E}(\log^2 x) - \tau\mu\,\mathbb{E}(\log x) + \frac{1}{2}\tau\mu^2. \end{aligned}$$

The derivative with respect to μ is $\partial/\partial\mu\{D_{KL}(f_4(x|\theta)\|f_2(x|\mu,\tau))\} = -\tau\,\mathbb{E}(\log x) + \tau\mu$, which is solved for $\mu = \mathbb{E}(\log x) = -\log\theta - \gamma$. When we differentiate with respect to τ, we obtain the partial derivative $\partial/\partial\tau\{D_{KL}(f_4(x|\theta)\|f_2(x|\mu,\tau))\} = -1/(2\tau) + \mathbb{E}(\log^2 x)/2 - \mu\,\mathbb{E}(\log x) + \mu^2/2$, which is solved for $\tau = 1/Var(\log x) = 6/\pi^2$.

□

Appendix C

Discretisation of the Parameters of a Normal Density

A possible way to derive objective priors for continuous parameter spaces through our idea, is by discretising the spaces. Here we present the approach for the Normal distribution and a generalisation to the exponential family.

Let $f_{j,k}(x|\mu_j, \sigma^2_{j,k})$ be a density of the Normal family, with mean $\mu_j \in (-\infty, \infty)$ and variance $\sigma^2_{j,k} > 0$. We also assume that the parameter are increasingly ordered, that is $\mu_{j-1} < \mu_j$ and $\sigma^2_{j,k-1} < \sigma^2_{j,k}$. Note that it is possible to obtain the same results by considering a decreasing order, and in general, there is no substantial difference in considering an increasing or decreasing order, as moving from one or the other is simply a "mirroring" exercise. As such, we discuss in detail the case where the discretised parameters are increasingly ordered and, where necessary, we reserve a special treatment for when they are decreasingly ordered.

C.1 Unknown mean and known variance

Let us consider the case where the variance is known, and we need to put a prior on μ. We have then, $\sigma^2_{j,k} = \sigma^2$, $\forall j, k$. We can then simplify the notation for this case, by setting $f_j = N(\mu_j, \sigma^2)$. Thus, applying our approach (as discussed in Chapter 3), the prior mass to be put on μ_j will be proportional to the minimum between $D_{KL}(f_j \| f_{j-1})$ and $D_{KL}(f_j \| f_{j+1})$.

For convenience, we repeat here the expression of the Kullback–Leibler divergence between two univariate Normal distribution, that is

$$D_{KL}(f(x|\mu_1,\sigma_1^2)\|f(x|\mu_2,\sigma_2^2)) = \frac{1}{2}\left\{\frac{(\mu_1-\mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - 1\right\}. \quad \text{(C.1)}$$

Therefore, from (C.1) we see that if the nearest model to f_j is f_{j-1}, meaning that $(\mu_j - \mu_{j-1}) < (\mu_j - \mu_{j+1})$, and the prior will be given by

$$\pi(\mu_j) \propto \exp\left\{(\mu_j - \mu_{j-1})^2\right\}.$$

If the nearest model to f_j is f_{j+1}, that is $(\mu_j - \mu_{j-1}) > (\mu_j - \mu_{j+1})$, we have

$$\pi(\mu_j) \propto \exp\left\{(\mu_j - \mu_{j+1})^2\right\}.$$

In other words, in the case where the variance is known, the minimum Kullback–Leibler divergence is determined by the density which has the value of the mean closer to μ_j. The prior would then be proportional to the square of the difference between the two means.

C.2 Known mean and unknown variance

If the mean is known, and we need to put a prior on the variance, we will have $\mu_j = \mu, \forall j$ and $f_k = N(\mu, \sigma_k^2)$. To find the minimum divergence between $D_{KL}(f_k\|f_{k-1})$ and $D_{KL}(f_k\|f_{k+1})$, we note that (C.1) has the form $x - \log x$ when we consider two Normal densities with same mean and different variance. The variable x, in this case, represents the ratio of the variances.

In Figure C.1 we have plotted function $x - \log x$ in the interval $(0, 5)$. From this, we can infer the behaviour of the Kullback–Leibler divergence computed between to Normal distributions with same mean and different variance (where, as assumed here, the variance is discretised and the densities considered have different and consecutive value of the parameter). When we compare consecutive variances, we can have either an increasing scenario $(\sigma_k^2 > \sigma_{k+1}^2)$ or a decreasing

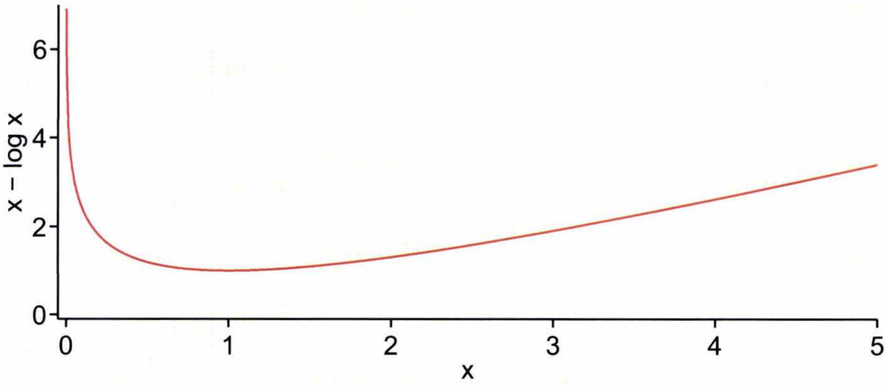

Figure C.1: Plot of the function $x - \log x$, where x represents the ratio between two (consecutive) variances of a Normal density.

scenario ($\sigma_k^2 < \sigma_{k+1}^2$). In the former case, the the ratio of consecutive variances is larger than one; whilst in the latter, this ratio is smaller than one. In either case, the Kullback–Leibler divergence is an increasing function of the difference between the two variances.

Let us us now define $\lambda_k = \sigma_{k+1}^2/\sigma_k^2$. Thus, (C.1) can be written as

$$(\lambda_{k-1} - \log \lambda_{k-1} - 1)/2 \propto (\lambda_{k-1} - \log \lambda_{k-1} - 1).$$

Given that $\lambda_{k-1} - \log \lambda_{k-1} < 1/\lambda_k - \log 1/\lambda_k$ implies that $\sigma_k^2/\sigma_{k-1}^2 - \log(\sigma_k^2/\sigma_{k-1}^2) < \sigma_k^2/\sigma_{k+1}^2 - \log(\sigma_k^2/\sigma_{k+1}^2)$, the smaller Kullback–Leibler divergence is $D_{KL}(f_k \| f_{k-1})$. Thus, the prior is

$$\pi(\sigma_k^2) \propto \exp\left\{\frac{\sigma_k^2}{\sigma_{k-1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k-1}^2}\right)\right\}.$$

If $\lambda_{k-1} - \log \lambda_{k-1} < 1/\lambda_k - \log 1/\lambda_k$, which implies that $\sigma_k^2/\sigma_{k-1}^2 - \log(\sigma_k^2/\sigma_{k-1}^2) > \sigma_k^2/\sigma_{k+1}^2 - \log(\sigma_k^2/\sigma_{k+1}^2)$, then the minimum divergence is $D_{KL}(f_k \| f_{k+1})$, and the prior

$$\pi(\sigma_k^2) \propto \exp\left\{\frac{\sigma_k^2}{\sigma_{k+1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k+1}^2}\right)\right\}.$$

C.3 Unknown mean and variance

Let us now consider the case where both parameters are unknown. Thus, $f_{j,k} = N(\mu_j, \sigma^2_{j,k})$. The discretised parameter spaces will now form a lattice structure, and from each point in the structure $(f_{j,k})$ we can compute eight divergences. That is, for $f_{j,k+1}$, $f_{j+1,k+1}$, $f_{j+1,k}$, $f_{j+1,k-1}$, $f_{j,k-1}$, $f_{j-1,k-1}$, $f_{j-1,k}$ and $f_{j-1,k+1}$.

Theorem C.1 and Theorem C.2 below, show that the minimum Kullback–Leibler divergence, with respect to $f_{j,k}$, is attained when only one of the two parameter varies.

Theorem C.1. *Let $f_{j,k} = N(\mu_j, \sigma^2_{j,k})$, where μ_j and $\sigma^2_{j,k}$ are the discretised version of, respectively, the mean and the variance of the Normal density. Then, $D_{KL}(f_{j,k} \| f_{j,k+1}) < D_{KL}(f_{j,k} \| f_{j-1,k+1})$ and $D_{KL}(f_{j,k} \| f_{j,k+1}) < D_{KL}(f_{j,k} \| f_{j+1,k+1})$.*

Proof. By considering the expression of the Kullback–Leibler divergence between two Normal densities in (C.1), we have

$$\begin{aligned} D_{KL}(f_{j,k} \| f_{j,k+1}) &< D_{KL}(f_{j,k} \| f_{j-1,k+1}) \\ \frac{1}{2}\left\{ \frac{(\mu_j - \mu_j)^2}{\sigma^2_{j,k+1}} + \frac{\sigma^2_{j,k}}{\sigma^2_{j,k+1}} \log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k+1}}\right) - 1 \right\} &< \\ &\frac{1}{2}\left\{ \frac{(\mu_j - \mu_{j-1})^2}{\sigma^2_{j,k+1}} + \frac{\sigma^2_{j,k}}{\sigma^2_{k+1}} \log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k+1}}\right) - 1 \right\} \\ 0 &< \frac{(\mu_j - \mu_{j-1})^2}{\sigma^2_{j,k+1}}. \end{aligned}$$

This is true for any μ_j, μ_{j-1} and $\sigma^2_{j,k+1}$, proving the first part of the statement.

Similarly, we have

$$\begin{aligned} D_{KL}(f_{j,k} \| f_{j,k+1}) &< D_{KL}(f_{j,k} \| f_{j+1,k+1}) \\ \frac{1}{2}\left\{ \frac{(\mu_j - \mu_j)^2}{\sigma^2_{j,k+1}} + \frac{\sigma^2_{j,k}}{\sigma^2_{j,k+1}} \log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k+1}}\right) - 1 \right\} &< \end{aligned}$$

$$\frac{1}{2}\left\{\frac{(\mu_j-\mu_{j+1})^2}{\sigma^2_{j,k+1}}+\frac{\sigma^2_{j,k}}{\sigma^2_{k+1}}\log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k+1}}\right)-1\right\}$$
$$0<\frac{(\mu_j-\mu_{j+1})^2}{\sigma^2_{j,k+1}}.$$

This is also true for any μ_j, μ_{j+1} and $\sigma^2_{j,k+1}$, proving the second part of the theorem statement. □

Theorem C.2. *Let $f_{j,k}=N(\mu_j,\sigma^2_{j,k})$, where μ_j and $\sigma^2_{j,k}$ are the discretised version of, respectively, the mean and the variance of the Normal density. Then, $D_{KL}(f_{j,k}\|f_{j,k-1})<D_{KL}(f_{j,k}\|f_{j-1,k-1})$ and $D_{KL}(f_{j,k}\|f_{j,k-1})<D_{KL}(f_{j,k}\|f_{j+1,k-1})$.*

Proof. By considering the expression of the Kullback–Leibler divergence between two Normal densities in (C.1), we have

$$D_{KL}(f_{j,k}\|f_{j,k-1})<D_{KL}(f_{j,k}\|f_{j-1,k-1})$$
$$\frac{1}{2}\left\{\frac{(\mu_j-\mu_j)^2}{\sigma^2_{j,k-1}}+\frac{\sigma^2_{j,k}}{\sigma^2_{j,k-1}}\log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k-1}}\right)-1\right\}<$$
$$\frac{1}{2}\left\{\frac{(\mu_j-\mu_{j-1})^2}{\sigma^2_{j,k-1}}+\frac{\sigma^2_{j,k}}{\sigma^2_{k-1}}\log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k-1}}\right)-1\right\}$$
$$0<\frac{(\mu_j-\mu_{j-1})^2}{\sigma^2_{j,k-1}}.$$

This is true for any μ_j, μ_{j-1} and $\sigma^2_{j,k-1}$, proving the first part of the statement.

Similarly, we have

$$D_{KL}(f_{j,k}\|f_{j,k-1})<D_{KL}(f_{j,k}\|f_{j+1,k-1})$$
$$\frac{1}{2}\left\{\frac{(\mu_j-\mu_j)^2}{\sigma^2_{j,k-1}}+\frac{\sigma^2_{j,k}}{\sigma^2_{j,k-1}}\log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k-1}}\right)-1\right\}<$$
$$\frac{1}{2}\left\{\frac{(\mu_j-\mu_{j+1})^2}{\sigma^2_{j,k-1}}+\frac{\sigma^2_{j,k}}{\sigma^2_{k-1}}\log\left(\frac{\sigma^2_{j,k}}{\sigma^2_{k-1}}\right)-1\right\}$$

$$0 < \frac{(\mu_j - \mu_{j+1})^2}{\sigma_{j,k-1}^2}.$$

This is also true for any μ_j, μ_{j+1} and $\sigma_{j,k-1}^2$, proving the second part of the theorem statement. □

The prior mass to be put on $(\mu_j, \sigma_{j,k}^2)$, depends on the smallest divergence among $D_{KL}(f_{j,k}\|f_{j,k-1})$, $D_{KL}(f_{j,k}\|f_{j,k+1})$, $D_{KL}(f_{j,k}\|f_{j-1,k})$ or $D_{KL}(f_{j,k}\|f_{j+1,k})$.

C.4 Discretisation of the parameters - Special cases

With the results discussed above, we can see that the prior depends on the discretisation scheme. For example, in the case where the unknown parameter is the mean, the mass will be proportional for the Kullback–Leibler divergence $D_{KL}(f_{j,k}\|f_{j+1,k})$ if the distances between the consecutive means are decreasing. It is then interesting to analyse some particular structures of the discretisation to retrieve well known priors.

In this section we show that, while it is possible to construct the discretised structures in a way that the prior to be assigned to the parameters is uniform, it is not possible (except for the prior on μ) to define a discretisation of the parameters such that the prior is Jeffreys'.

Uniform prior

If the parameter to estimate is the mean, that is the variance is known, the uniform prior of μ can be obtained by simply considering consecutive means separated by equal intervals. In fact, if we set $\mu_j - \mu_{j-1} = \delta$, for every j, the prior will be

$$\pi(\mu_j) \propto \exp\left\{(\mu_j - \mu_{j-1})^2\right\} = 1.$$

It is clear that, in this case, we have $D_{KL}(f_j\|f_{j-1}) = D_{KL}(f_j\|f_{j+1})$. In fact, this is a situation where the Kullback–Leibler divergence is symmetrical.

A noteworthy aspect is that this prior can be designed by setting up the sequence of means such that every point is always the arithmetic mean of the two contiguous ones. That is, if $\mu_j = (\mu_{j-1} + \mu_{j+1})/2$.

If the unknown parameter is the variance, we can put an uniform prior on it by considering the ratio between consecutive variances constant. The general result assumes that the variances are increasingly ordered and it is obtained as follows.

First, we note that to obtain a uniform prior for σ_k^2, we need $D_{KL}(f_{k-1}\|f_k) = D_{KL}(f_k\|f_{k+1})$. In fact, from (C.1) we see that the Kullback–Leibler divergence between to Normal distribution with common mean but difference variance is not symmetrical. Therefore, in order to have a uniform prior, the minimum divergences have to be all in the same "direction". To obtain this construction, and recalling that we set $\lambda_k = \sigma_{k+1}^2/\sigma_k^2$, we have

$$\begin{aligned} D_{KL}(f_{k-1}\|f_k) &= D_{KL}(f_k\|f_{k+1}) \\ \frac{\sigma_{k-1}^2}{\sigma_k^2} - \log\left(\frac{\sigma_{k-1}^2}{\sigma_k^2}\right) &= \frac{\sigma_k^2}{\sigma_{k+1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k+1}^2}\right) \\ \frac{1}{\lambda_{k-1}} - \log\left(\frac{1}{\lambda_{k-1}}\right) &= \frac{1}{\lambda_k} - \log\left(\frac{1}{\lambda_k}\right). \end{aligned} \tag{C.2}$$

The solution of (C.2) is $\lambda_{k-1} = \lambda_k$. Thus, in order to obtain equal "right" Kullback–Leibler divergences, we need to have

$$\frac{\sigma_{k-1}^2}{\sigma_k^2} = \frac{\sigma_k^2}{\sigma_{k+1}^2}, \tag{C.3}$$

which means that the ratio between consecutive variances has to be constant: $\sigma_k^2/\sigma_{k-1}^2 = \lambda$. However, to have the desired uniform prior on the variance, it is not sufficient that the "right" divergences are all equal, but also that they are always smaller that the ones computed in the opposite direction. In fact, our approach requires a prior mass proportional to the minimum Kullback–Leibler divergence from the point. To show that $D_{KL}(f_k\|f_{k-1}) > D_{KL}(f_k\|f_{k+1})$, given that $D_{KL}(f_k\|f_{k+1}) = D_{KL}(f_{k-1}\|f_k)$, we see that we need to have

$$\begin{aligned}
\lambda_{k-1} - \log \lambda_{k-1} &> \frac{1}{\lambda_k} - \log\left(\frac{1}{\lambda_k}\right) \\
\lambda_{k-1} - \log \lambda_{k-1} &> \frac{1}{\lambda_{k-1}} - \log\left(\frac{1}{\lambda_{k-1}}\right) \\
\lambda_{k-1} - \frac{1}{\lambda_{k-1}} &> 2\log \lambda_{k-1},
\end{aligned} \tag{C.4}$$

where the second inequality in (C.4) holds as the variances are assumed to be increasing. We note that the last inequality in (C.4) has the form of $x - 1/x > 2\log x$, with $x > 1$. In Figure C.2 we have plotted function $x - 1/x$, represented by the continuous red curve, and function $2\log x$, represented by the dashed black curve. We can easily see that, for $x > 1$, which corresponds to our initial assumptions that the variances are increasing, the inequality in (C.4) holds.

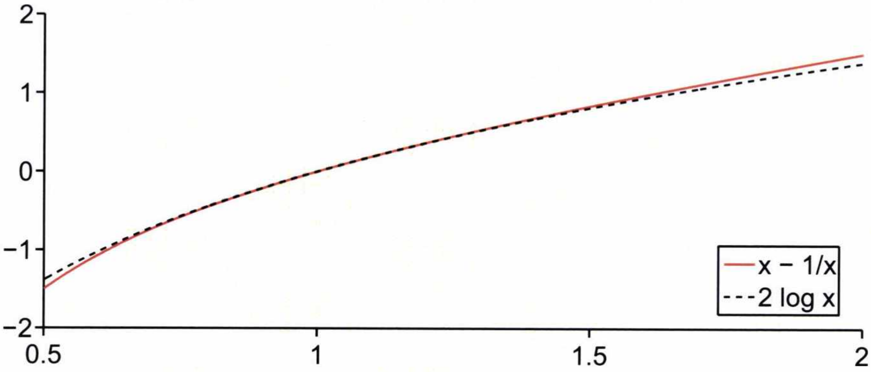

Figure C.2: Plot of function $x - 1/x$ (continuous red line) and function $2\log x$ (dashed black line). The variable x represents the ratio of two consecutive variances $\lambda_{k-1} = \sigma_k^2/\sigma_{k-1}^2$.

Should we assume that the variances are decreasing, we would have $\lambda_k = \sigma_{k+1}^2/\sigma_k^2 < 1$. In this case, we would have $D_{KL}(f_{k+1}\|f_k) = D_{KL}(f_k\|f_{k-1})$. By inspecting in Figure C.2 the region where $x < 1$ (i.e. $\lambda < 1$) we see that the equal divergences are always smaller than the one computed in the opposite direction.

For the uniform prior for the mean, we have seen that this can be designed by setting each point as the arithmetic mean of the two contiguous points. Similarly, a sensible result can be obtained for the prior on the variance. In particular, as $\lambda = \sigma_k^2/\sigma_{k-1}^2$ has to be constant, from (C.3), this is possible when $\sigma_k^2 = \sqrt{\sigma_{k-1}^2 \cdot \sigma_{k+1}^2}$. In other words, when each point is the geometric mean of the contiguous points.

To find the uniform distribution when both parameters μ and σ^2 are unknown, an additional condition has to be included. This is determined as follows.

Let us consider the lattice structure defined by μ_j and $\sigma_{j,k}^2$, where the mean points define the columns of the lattice and the variance points the rows on the lattice. To clarify: if we fix the variance, changes in the mean are represented by horizontal movements. If we fix the mean, changes in the variance are represented by vertical movements.

In order to define a uniform prior on $(\mu_j, \sigma_{j,k}^2)$, we need to design the lattice in a way that, at each point, the minimum Kullback–Leibler divergence (measured to a contiguous point) is constant. From the above discussions, we know that the minimum divergence from a point of the lattice occurs when only one of the two parameter varies. That is, it is either a "horizontal" or a "vertical" distance. However, from (C.1), we see that it is not possible to keep any of the "horizontal" divergences constant. In fact, its value does not depend on the means only, by it is inversely proportional to the variance. For example, if $\sigma_k^2 > \sigma_{k-1}^2$, then $D_{KL}(f_{j,k} \| f_{j+1,k}) < D_{KL}(f_{j,k-1} \| f_{j+1,k-1})$. Therefore, the only way to design a lattice structure that leads to a uniform prior, is done by forcing one of the "vertical" divergences to be constant. In fact, the Kullback–Leibler divergence between two Normal distributions with the same mean (refer to (C.1)) depend only on the ratio of the variances.

Let us now start by constructing a lattice structure in agreement with the conditions we have discussed above to obtain a uniform prior on each parameter independently. That is, we put $\delta = \mu_j - \mu_{j-1}$ and $\lambda = \sigma_k^2/\sigma_{k-1}^2$. The first condition ensures that, in each row, the "horizontal" distances are constant; the second condition ensures that, in each column, the "vertical" distances are constant. Then, by construction, the minimum Kullback–Leibler divergence measure from $f_{j,k}$ can be either $D_{KL}(f_{j,k} \| f_{j-1,k})$ or $D_{KL}(f_{j,k} \| f_{j,k+1})$. Note that the first divergence can be replaced by $D_{KL}(f_{j,k}\|f_{j+1,k})$, as the Kullback–Leibler divergence is symmetrical in this circumstances. However, on the basis of what said above, the smallest divergence has to be the "vertical" one, that is $D_{KL}(f_{j,k}\|f_{j,k+1})$, for the "horizontal" divergence depends on the row (i.e. on k). To obtain this, we need to set the following extra condition

$$\begin{aligned} D_{KL}(f_{j,k}\|f_{j,k+1}) &< D_{KL}(f_{j,k}\|f_{j-1,k}) \\ \frac{\sigma_{j,k}^2}{\sigma_{j,k+1}^2} - \log\left(\frac{\sigma_{j,k}^2}{\sigma_{j,k+1}^2}\right) &< \frac{(\mu_j - \mu_{j-1})^2}{\sigma_{j,k}^2} + 1 \\ \frac{1}{\lambda}\log\left(\frac{1}{\lambda}\right) &> \frac{\delta^2}{\sigma_{j,k}^2} + 1. \end{aligned} \tag{C.5}$$

Equation (C.5) is solve with respect to the variance, and we have

$$\sigma_{j,k}^2 < \frac{\delta^2}{1/\lambda - \log(1/\lambda) - 1}. \tag{C.6}$$

We see from (C.6) that the additional condition is an upper bound for the variance. Thus, when both parameters are unknown, and we wish to define a discretised parameter space such that the prior mass on each point is uniform, we need to set three conditions. Namely, the arithmetic mean for μ, the geometric mean for σ_2 and fix an upper bound for the variances, where this bound depends on the first two conditions. We note that this limit can be controlled by the distance between the means (δ) and the ratio between the variances (λ).

Jeffreys' prior

We now discuss the possibility of discretising the parameter space of a Normal density and, by applying our approach, retrieve Jeffreys' prior.

The Jeffreys prior for the mean of a Normal, when the variance is known, is a uniform: $\pi(\mu) \propto 1$. As such, we can conclude that it is possible to design a discretisation of the mean such that the prior mass is a uniform and, in particular, this is the result we have discussed in Section C.4 above.

We now show that, on the basis of the proposed prior, it is not possible to construct a discrete structure such that Jeffreys' prior can be obtained, in both cases when the parameter is the variance or the pair mean and variance.

The Jeffreys' prior for the variance of a Normal density, when the mean is known, is $\pi(\sigma^2) \propto 1/\sigma^2 = \exp(-\log\sigma^2)$. Thus, because our approach assigns a mass which is proportional to the exponential of the Kullback–Leibler divergence, we nee to have $-\log\sigma^2 \geq 0$. Therefore, we need to consider discrete structure of variances between zero and one.

If the minimum divergence from f_k is $D_{KL}(f_k \| f_{k+1})$, we would have

$$\frac{1}{2}\left\{\frac{\sigma_k^2}{\sigma_{k+1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k+1}^2}\right) - 1\right\} = -\log\sigma_k^2,$$

which has solution

$$\sigma_k^2 = \sigma_{k+1}^2\left\{1 - \log(\sigma_k^2\sigma_{k+1}^2)\right\}. \tag{C.7}$$

Given that the variances have to be smaller than one, $-\log(\sigma_k^2\sigma_{k+1}^2) > 0$, and $\sigma_k^2/\sigma_{k+1}^2 > 1$. Thus, under the assumption that the minimum divergence is $D_{KL}(f_k \| f_{k+1})$, it is not possible to have increasing variances. Then, we consider the case where the variances are decreasing. We should than have

$$\frac{1}{2}\left\{\frac{\sigma_k^2}{\sigma_{k-1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k-1}^2}\right) - 1\right\} > \frac{1}{2}\left\{\frac{\sigma_k^2}{\sigma_{k+1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k+1}^2}\right) - 1\right\} \tag{C.8}$$

By replacing (C.7) into the right-hand-side of (C.8), we obtain that $\sigma_k^2/\sigma_{k-1}^2 + \log(\sigma_k^2\sigma_{k-1}^2) > 1$. However, as we assume decreasing variances, we have $\sigma_k^2/\sigma_{k-1}^2$ and $\log(\sigma_k^2\sigma_{k-1}^2) < 0$, inequality (C.8) does not hold. Thus, it is not possible to simultaneously have the condition for Jeffreys' prior and $D_{KL}(f_k \| f_{k+1}) < D_{KL}(f_k \| f_{k-1})$ satisfied.

Let us now assume that the minimum divergence is $D_{KL}(f_k \| f_{k-1})$. In this case, Jeffreys' prior would be possible if, for every variance value smaller than one, we have

$$\sigma_k^2 = \sigma_{k-1}^2 \left\{1 - \log(\sigma_k^2 \sigma_{k-1}^2).\right\} \tag{C.9}$$

If we assume decreasing variances, equation (C.9) never holds. In fact, this equation implies that $\sigma_k^2/\sigma_{k-1}^2 > 1$, which is not compatible with decreasing variances. If we assume increasing variances, we have

$$\frac{1}{2}\left\{\frac{\sigma_k^2}{\sigma_{k-1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k-1}^2}\right) - 1\right\} < \frac{1}{2}\left\{\frac{\sigma_k^2}{\sigma_{k+1}^2} - \log\left(\frac{\sigma_k^2}{\sigma_{k+1}^2}\right) - 1\right\}, \tag{C.10}$$

which, by substituting (C.9) into the left-hand-side, gives $\sigma_k^2/\sigma_{k+1}^2 + \log(\sigma_k^2\sigma_{k+1}^2) > 1$. Given that in this case $\sigma_k^2/\sigma_{k+1}^2 < 1$ and that $\log(\sigma_k^2\sigma_{k+1}^2) < 0$, inequality (C.10) never holds.

We can then conclude that it is not possible to design a discretised structure of the parameter space of the variance of a Normal density such that, by applying our objective prior, it is possible to obtain Jeffreys' prior.

Considering the case both the parameters of the Normal distribution are unknown, we need to distinguish between Jeffreys' rule prior and Jeffreys' independent prior. We recall that the first one is obtained by applying Jeffreys' invariance method, whilst the second assumes independence of the parameters. Jeffreys' rule prior is $\pi(\mu, \sigma^2) \propto 1/\sigma^4$. Jeffreys' independent prior, which coincides with the reference prior, is $\pi(\mu, \sigma^2) = \pi(\mu)\pi(\sigma^2) = 1/\sigma^2$. As the prior obtained by applying Jeffreys' rule gives unacceptable results (i.e. the posterior distribution would be a chi-square with n degrees of freedom, where n is the sample size, which does not take into account the loss of a degree of freedom in estimating the mean), we focus our discussion on Jeffreys' independent prior only.

We have seen that it is not possible to design the lattice structure in such a way that Jeffreys' prior can be obtained by considering the minimum divergence one of the "vertical" ones, namely $D_{KL}(f_{j,k}\|f_{j,k+1})$ and $D_{KL}(f_{j,k}\|f_{j,k-1})$. Therefore, we show that it is not possible to work out an appropriate structure by aiming to have as minimum distance one of the "horizontals".

We begin by noticing that we can assume the distance between the means constant, $\delta = \mu_j - \mu_{j-1}$. In fact, should we find an appropriate distance such that any of the "horizontal" divergences is the minimum and it leads to Jeffreys' prior, this automatically applies to any column of the lattice. Thus, we consider only the divergence $D_{KL}(f_{j,k} \| f_{j-1,k})$, being the one from j to $j+1$ identical. Also, we recall that the variances have to be confined in the interval $(0, 1)$.

Let us assume that the variances are increasing. If $D_{KL}(f_{j,k} \| f_{j-1,k})$ is smaller than $D_{KL}(f_{j,k} \| f_{j,k+1})$, then

$$\frac{\delta^2}{\sigma_{j,k}^2} + 1 = \frac{\sigma_{j,k}^2}{\sigma_{j,k+1}^2} - \log\left(\frac{\sigma_{j,k}^2}{\sigma_{j,k+1}^2}\right). \tag{C.11}$$

If $D_{KL}(f_{j,k} \| f_{j-1,k})$ is the minimum divergence, than it has to be $D_{KL}(f_{j,k} \| f_{j-1,k}) = \log \sigma_{j,k}^2$, which implies $\delta^2 = -2\sigma_{j,k}^2 \log \sigma_{j,k}^2$. By replacing this result into (C.11), we have $\sigma_{j,k}^2/\sigma_{j,k+1}^2 + \log(\sigma_{j,k}^2 \sigma_{j,k+1}^2) > 1$. This result is not possible because, by assuming increasing variances with value smaller than one, we have $\sigma_{j,k}^2/\sigma_{j,k+1}^2 < 1$ and $\log(\sigma_{j,k}^2 \sigma_{j,k+1}^2) < 0$. As such, any of the "horizontal" divergences, under the above assumptions, is never smaller than $D_{KL}(f_{j,k} \| f_{j,k+1})$.

If we assume decreasing variances, we show that $D_{KL}(f_{j,k} \| f_{j-1,k})$ is never smaller than $D_{KL}(f_{j,k} \| f_{j,k-1})$. In fact

$$\frac{\delta^2}{\sigma_{j,k}^2} + 1 = \frac{\sigma_{j,k}^2}{\sigma_{j,k-1}^2} - \log\left(\frac{\sigma_{j,k}^2}{\sigma_{j,k-1}^2}\right). \tag{C.12}$$

Setting $\delta^2 = -2\sigma_{j,k}^2 \log \sigma_{j,k}^2$ in (C.12) above, we have $\sigma_{j,k}^2/\sigma_{j,k-1}^2 + \log(\sigma_{j,k}^2)\sigma_{j,k-1}^2 > 1$. As the variances are decreasing and smaller than one, $\sigma_{j,k}^2/\sigma_{j,k-1}^2 < 1$ and $\log(\sigma_{j,k}^2 \sigma_{j,k-1}^2) < 0$, also inequality (C.12) does not hold. Again, given that $D_{KL}(f_{j,k} \| f_{j-1,k})$ can never be the smallest one under these assumption, we conclude that Jeffreys' prior cannot be obtained.

C.5 Extension to the exponential family

The general results discussed above can be generalised to the exponential family of distributions.

Let us consider a distribution $f(x|\theta)$ belonging to the exponential family. The general form is

$$f(x|\theta) = c(x) \exp\{h(x)a(\theta) - m(\theta)\}, \tag{C.13}$$

where $c(x)$ and $h(x)$ are functions of x, and $a(\theta)$ and $m(\theta)$ are functions of the parameter θ only. If $a(\theta) = \theta$, the exponential family is said to be in *canonical form*. If we consider the exponential family in canonical form with the simplest $h(x)$, that is $h(x) = x$, we have

$$f(x|\theta) = c(x) \exp\{x\theta - m(\theta)\}.$$

Let us consider density $f(x|\theta_1)$ and $g(x|\theta_2)$, both belonging to the exponential family. Thus, the Kullback–Leibler divergence between $f(x|\theta_1)$ and $g(x|\theta_2)$ is given by

$$\begin{aligned} D_{KL}(f(x|\theta_1)\|g(x|\theta_2)) &= \int f(x|\theta_1) \log\left\{\frac{f(x|\theta_1)}{g(x|\theta_2)}\right\} dx \\ &= \mathbb{E}\{h(x)\}\{a(\theta_1) - a(\theta_2)\} + \{m(\theta_2) - m(\theta_1)\} \end{aligned} \tag{C.14}$$

If the distributions are in the canonical form and $h(x) = x$, expression (C.14) becomes

$$D_{KL}(f(x|\theta_1)\|g(x|\theta_2)) = \mathbb{E}(x)(\theta_1 - \theta_2) + \{m(\theta_2) - m(\theta_1)\}.$$

As the moment of a distribution belonging to the exponential family are obtained by differentiating $n(\theta)$, we have $\overline{x} = m'(\theta)$. In this case, the Kullback–Leibler divergence between two densities of the exponential family, in their general form, becomes

$$D_{KL}(f(x|\theta_1)\|g(x|\theta_2)) = m'(\theta_1)\{h(x)\}\{a(\theta_1) - a(\theta_2)\} + \{m(\theta_2) - m(\theta_1)\},$$

whilst the one in canonical form, with $h(x) = x$ becomes

$$D_{KL}(f(x|\theta_1)\|g(x|\theta_2)) = m'(\theta_1)(\theta_1 - \theta_2) + \{m(\theta_2) - m(\theta_1)\}\,.$$

Let us now consider a discretisation of the parameter space of θ for a distribution of the exponential family in the canonical form, $f(x|\theta)$. To obtain the mass to be assigned on θ_j, we need to be able to asses which divergence between $D_{KL}(f(x|\theta_j)\|f(x|\theta_{j-1}))$ and $D_{KL}(f(x|\theta_j)\|f(x|\theta_{j+1}))$ is the smallest. We see that

$$\begin{aligned} D_{KL}(f(x|\theta_j)\|f(x|\theta_{j-1})) &< D_{KL}(f(x|\theta_j)\|f(x|\theta_{j+1})) \\ m'(\theta_j)(\theta_j - \theta_{j-1}) + \{m(\theta_{j-1}) - m(\theta_j)\} &< m'(\theta_j)(\theta_j - \theta_{j+1}) + \{m(\theta_{j+1}) - m(\theta_j)\} \\ m'(\theta_j) &< \frac{m(\theta_{j+1}) - m(\theta_{j-1})}{\theta_{j+1} - \theta_{j-1}}. \qquad \text{(C.15)} \end{aligned}$$

As $m'(\theta_j)$ is the expectation of the distribution, we see that expression (C.15) suggests the condition for the mean of the distribution for which the prior mass on θ_j is proportional to the divergence from $f(x|\theta_j)$ to $f(x|\theta_{j-1})$. By inverting the inequality sign in (C.15), we have that the prior mass will be proportional to $D_{KL}(f(x|\theta_j)\|f(x|\theta_{j+1}))$ when

$$m'(\theta_j) > \frac{m(\theta_{j+1}) - m(\theta_{j-1})}{\theta_{j+1} - \theta_{j-1}}. \qquad \text{(C.16)}$$

The following Example C.1 applies the above results to the special case of a Normal density with known variance, expressed in the exponential family form. It also shows how the result is consistent with the one obtained in Section C.4.

Example C.1. *Let us consider a Normal distribution with unknown mean μ and known variance σ^2. To express this distribution in canonical form* (C.13)*, we set $c(x) = \exp(-x^2/2\sigma^2)/\sqrt{2\pi\sigma^2}$, $h(x) = x/\sigma^2$, $a(\theta) = \mu$ and $m(\theta) = \mu^2/2\sigma^2$.*

By considering either (C.15) *or* (C.16)*, and that $m'(\mu) = \mu/\sigma^2$, we have*

$$\frac{\mu}{\sigma^2} = \frac{\mu_{j+1}^2/2\sigma^2 - \mu_{j-1}^2/2\sigma^2}{\mu_{j+1} - \mu_{j-1}},$$

which solved for μ_j gives the result

$$\mu_j = \frac{\mu_{j-1} + \mu_{j+1}}{2}. \tag{C.17}$$

The result in (C.17) *reconciles with the result we have obtained in Section C.4. In fact, what* (C.17) *states is that, if we chose each point in the sample space f zmu equal to the arithmetic mean of the two contiguous points, then the divergences* $D_{KL}(f(x|\mu_j)\|f(x|\mu_{j-1}))$ *and* $D_{KL}(f(x|\mu_j)\|f(x|\mu_{j+1}))$ *will be equal. Therefore, we would obtain the uniform prior on* μ.

We see that, if we choose $\mu_j < (\mu_{j-1}+\mu_{j+1})/2$ *at each point, then the minimum divergence will be always* $D_{KL}(f(x|\mu_j)\|f(x|\mu_{j-1}))$. *In fact, this means that the inequality* $|\mu_j - \mu_{j-1}| < |\mu_j - \mu_{j+1}|$ *holds. And this inequality has solutions* $\mu_j > \mu_{j+1} + |\mu_j - \mu_{j-1}|$ *and* $\mu_j < \mu_{j+1} - |\mu_j - \mu_{j-1}|$. *Given that we assume* $\mu_{j-1} < \mu_{j+1}$, *the only possible solution is the former one, giving*

$$\mu_j < \frac{\mu_{j-1} - \mu_{j+1}}{2},$$

which agrees with the result previously obtained. With a similar process, it is straightforward to see that when we set each point larger than the arithmetic mean of the two contiguous points, that is $\mu_j > (\mu_{j-1}-\mu_{j+1})/2$, *the minimum divergence is always* $D_{KL}(f(x|\mu_j)\|f(x|\mu_{j+1}))$.

In the next example, we consider a Normal density with known mean $\mu = 0$ and unknown variance. In this case as well, we obtain results consistent with the previous.

Example C.2. *Let us consider a Normal density with known mean* $\mu = 0$ *and unknown variance* σ^2. *We express it in the form of* (C.13) *by setting* $c(x) = 1/\sqrt{2\pi}$, $h(x) = x^2$, $a(\theta) = -1/2\sigma^2$ *and* $m(\theta) = \log\sigma^2$. *By applying either* (C.15) *or* (C.16), *and considering that* $m'(\theta) = 1/\sigma^2$, *we have*

$$\sigma^2_{j,k} = \frac{\log\sigma^2_{j,k+1} - \log\sigma_{j,k-1}}{1/(2\sigma^2_{j,k-1}) - 1/(2\sigma^2_{j,k+1})}. \tag{C.18}$$

By setting $\lambda_k = \sigma^2_{j,k+1}/\sigma^2_{j,k}$, *then the right-hand-side of equation* (C.18) *becomes* $-\sigma^2_{j,k}\left\{[\log\lambda_k + \log\lambda_{k-1}/[1/\lambda_k - \lambda_{k-1}\right\}$. *Therefore, from* (C.17) *we have*

$$\frac{-\log\lambda_k - \log\lambda_{k-1}}{1/\lambda_k - \lambda_{k-1}} = 1,$$

which can also be written as

$$\lambda_{k-1} - \log\lambda_{k-1} = \frac{1}{\lambda_k} - \log\left(\frac{1}{\lambda_k}\right).$$

We can than reconcile the previous general results. In fact, if we set $\lambda_{k-1} - \log\lambda_{k-1} < 1/\lambda_k - \log(1/\lambda_k)$*, then the minimum divergence will be* $D_{KL}(f_{j,k}\|f_{j,k-1})$*. On the contrary, if we set* $\lambda_{k-1} - \log\lambda_{k-1} > 1/\lambda_k - \log(1/\lambda_k)$*, then the minimum divergence will be* $D_{KL}(f_{j,k}\|f_{j,k+1})$*.*

References

F.J. Anscomber. Topics in the Investigation of Linear Relations Fitted by the Method of Least Squares. *Journal of the Royal Statistical Society, Series B*, 29: 1–52, 1967.

M. Barbieri and J.O. Berger. Optimal Predictive Model Selection. *The Annals of Statistics*, 32:870–897, 2004.

K. Barger and J. Bunge. Bayesian Estimation of the Number of Species Using Noninformative Priors. *Biometrical Journal*, 50:1064–1076, 2008.

M.J. Bayarri and J. Berger. The Interplay Between Bayesian and Frequentist Analysis. *Statistical Science*, 19:58–80, 2004.

T. Bayes. Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

J.O. Berger. *Statistical Decision Theory and Related Topics*. Springer, New York, 2nd edition, 1985.

J.O. Berger. The Case for Objective Bayesian Analysis. *Bayesian Analysis*, 1: 385–402, 2006. (with discussion).

J.O. Berger and J.M. Bernardo. Estimating a Product of Means: Bayesian Analysis with Reference Priors. *Journal of the American Statistical Association*, 84: 200–207, 1989.

J.O. Berger and J.M. Bernardo. On the Development of the Reference Priors. *Bayesian Statistics*, 4:35–60, 1992a.

J.O. Berger and J.M. Bernardo. Ordered Group Reference Priors with Application to the Multinomial Problem. *Biometrika*, 79:25–37, 1992b.

J.O. Berger and L.R. Pericchi. The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, 91:109–122, 1996.

J.O. Berger and L.R. Pericchi. Objective Bayesian Methods for Model Selection: Introduction and Comparison. *IMS Lecture Notes - Monograph Series*, 38:135–193, 2001.

J.O. Berger and D. Sun. Objective Priors for the Bivariate Normal Model. *The Annals of Statistics*, 36:963–982, 2006.

J.O. Berger, B. Liseo, and R.L. Wolpert. Integrated Likelihood Methods for Eliminating Nuisance Parameters. *Statistical Science*, 18:1–28, 1999.

J.O. Berger, J.M. Bernardo, and D. Sun. The Formal Definition of Reference Priors. *The Annals of Statistics*, 37:905–938, 2009.

J.O. Berger, J.M. Bernardo, and D. Sun. Objective Priors for Discrete Parameter Spaces. *Journal of the American Statistical Association*, 107:636–648, 2012.

R.H. Berk. Limiting Behaviour of Posterior Distributions When the Model is Incorrect. *Annals of Mathematical Statistics*, 37:51–58, 1966.

R.H. Berk. Consistency a Posteriori. *Annals of Mathematical Statistics*, 41:894–906, 1970.

J.M. Bernardo. Reference Posterior Distributions for Bayesian Reference. *Journal of the Royal Statistical Society, Series B*, 41:113–147, 1979.

J.M. Bernardo. Noninformative Priors do not Exist: A Discussion. *Journal of Statistical Planning and Inference*, 65:159–189, 1997.

J.M. Bernardo. Nested Hypothesis Testing: the Bayesian Reference Criterion. *Bayesian Statistics*, 6:101–130, 1999.

J.M. Bernardo. Reference Analysis. In *Handbook of Statistics*, (D.K.Day and C.R. Rao eds.), pages 17–90. Amsterdam: Elsevier, 2005.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* Chichester: Wiley, 1994.

A. Bhattacharyya. On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distirbutions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

S. Blyth. Local Divergence and Association. *Biometrika*, 81:579–584, 1994.

W.M. Briggs and R. Zaretzki. A New Loook at Inference for the Hypergeometric Distribution. Unpublished, February 2009.

P.J. Brown and S.G. Walker. Bayesian Priors from Loss Matching. *International Statistical Review*, 80:60–82, 2012. (with discussion).

G. Casella and E.L. Lehmann. *Theory of Point Estimation.* Springer, 2nd edition, 1998.

J. Casellas, N. Ibánez-Escriche, García-Cortéz, and L. Verona. Bayes Factor Between Student t anf Gaussian Mixed Models Within an Animal Breeding Context. *Genetics Selection Evolution*, 40:395–413, 2008.

H. Chipman, E.I. George, and R.E. McCulloch. The Practical Implementation of Bayesian Model Selection. *IMS Lecture Notes - Monograph Series*, 38:65–116, 2001.

N. Chopin, C.P. Robert, and J. Rousseau. Harold Jeffreys' Theory of Probability Revisited. *Statistial Science*, 24:141–172, 2009.

J.T. Chu. Errors in Normal Approximation to the y, τ, and Similar Types of Distributions. *Annals of Mathematical Statistics*, 27:780–789, 1956.

B. Clarke and A. Barron. Information-theoretic Asymptotics of Bayes Methods. *IEEE Transactions on Information Theory*, 36:453–473, 1990.

B. Clarke and A. Barron. Jeffreys' Prior is Asymptotically Least Favorable Under Entropy Risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

B. Clarke and D. Sun. Reference Priors Under the Chi-square Distance. *Sankhya A*, 59:215–231, 1997.

B. Clarke and D. Sun. Asymptotics of the Expected Posterior. *Annals of the Institute of Statistical Mathematics*, 51:163–185, 1999.

W. Cui and E.I. George. Empirical Bayes vs. Fully Bayes Variable Selection. *Journal of Statistical Planning and Inference*, 138:888–900, 2008.

G.S. Datta. On Priors Providing Frequentis Validity of Bayesian Inference for Multiple Parametric Functions. *Biometrika*, 83:287–298, 1996.

G.S. Datta and M. Ghosh. On the Invariance of Noninformative Priors. *Annals of Statistics*, 24:141–159, 1996.

G.S. Datta and R. Mukerjee. *Probability Matching Priors: Higher Order Asymptotics*. Lecture Notes in Statistics. Springer, New York, 2004.

G.S. Datta and T.J. Sweeting. Probability Matching Priors. Research Report 252, Department of Statistical Science, University College London, 2005.

G.S. Datta, M. Ghosh, and R. Mukerjee. Some New Results on Probability Matching Priors. *Calcutta Statistical Association Bulletin*, 50:179–192, 2000a.

G.S. Datta, R. Mukerjee, M. Ghosh, and T.J. Sweeting. Bayesian Prediction with Approximate Frequentist Validity. *Annals of Statistics*, 28:1414–1426, 2000b.

R.R. Davidson and B.R. Johson. Interchanging Parameters of the Hypergeometric Distribution. *Mathematics Magazine*, 66:328–329, 1993.

A.P. Dawid. *Invariant Prior Distributions*, volume 4. (S. Kotz, N. L. Johnson and C. B. Read, eds.) New York: Wiley, 1983.

A.P. Dawid, M. Stone, and J.V. Zidek. Marginalisation Paradoxes in Bayesian and Structural Inference. *Journal of the Royal Statistical Society, Series B*, 35: 189–233, 1973. (with discussion).

B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937.

N. Draper and I. Guttman. Bayesian Estimation of the Binomial Parameter. *Technometrics*, 13:667–673, 1971.

R. Dumonceaux and C.E. Antle. Discrimination Between the Log-normal and the Weibull Distributions. *Technometrics*, 15:923–926, 1973.

R. Dumonceaux, C.E. Antle, and G. Haas. Likelihood Ratio Test for Discrimination Between Two Models with Unknown Location and Scale Parameters. *Technometrics*, 15:19–31, 1973.

M.L. Eaton and D.A. Freedman. Dutch Book Against Some "Objective" Priors. *Bernoulli*, 10:861–872, 2004.

F.Y. Edgeworth. On the Probable Errors of Frequency-Constants. *Journal of the Royal Statistical Society, Series B*, 71:499–512, 1908.

F.J. Fabozzi, S.M. Focardi, M. Höchstötter, and S.T. Rachev. *Probability and Statistics for Finance.* Wiley, 2010.

C. Fernández and M. Steel. Reference Priors for the General Location-scale Model. *Statistics and Probability Letters*, 43:377–384, 1999a.

C. Fernández and M. Steel. Multivariate Student-t Regression Models: Pitfalls and Inference. *Biometrika*, 86:153–167, 1999b.

E. Fienberg. When Did Bayesian Inference Become"Bayesian"? *Bayesian Analysis*, 1:1–40, 2006.

T.C.O. Fonseca, M.A.R. Ferreira, and H.S. Migon. Objective Bayesian Analysis for the Student-t Regression Model. *Biometrika*, 95:325–333, 2008.

S. French. On the Axiomatization of Subjective Probabilities. *Theory and Decision*, 14:19–33, 1982.

S. French and D. Rios Insua. *Statistical Decision Theory*. Dendal's Library of Statistics 9. Arnold, London, 2000.

E.I. George and R.E. McCulloch. Approaches for Bayesian Variable Selection. *Statistica Sinica*, 7:339–373, 1997.

J. Geweke. Bayesian Treatment of the Independent Student-t Linear Model. *Journal of Applied Econometrics*, 8:S19–S40, 1993.

M. Ghosh and R. Liu. Moment Matching Priors. *Sankhyā: The Indian Journal of Statistics*, 73-A:185–201, 2011.

M. Ghosh, V. Mergel, and R. Liu. A General Divergence Criterion for Prior Selection. *Annals of the Institute of Statistical Mathematics*, 63:43–58, 2011.

M. Goldstein. Subjective Bayesian Analysis: Principles and Practice. *Bayesian Analysis*, 1:403–420, 2006. (with discussion).

I.B.J. Goudie and C.M. Glodie. Initial Size Estimation for the Pure Death Process. *Biometrika*, 68:543–550, 1981.

E. Hellinger. Neue Bergründung der Theorie Quadratischen Formen von Unendlichen Vielen Veränderlichen. *Journal für Reine und Angewandte Mathematik*, 136:210–271, 1909.

J. Hirshleifer and J.G. Riley. *The Analytics of Uncertainty and Information*. Cambridge University Press, Cambridge, New York, 1992.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14:382–417, 1999.

E. Jacquier, N.G. Polson, and P.E. Rossi. Bayesian Analysis of Stochastic Volatility Models with Fat-tails and Correlated Errors. *Journal of Econometrics*, 122: 185–212, 2004.

E.T. Jaynes. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4:227–241, 1968.

H. Jeffreys. *Theory of Probability*. University Press, Oxford, 1961.

N.L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions*. Houghton Mifflin, Boston, 1970.

M.A. Juárez and M.F.J. Steel. Model-based of non-Gaussian Panel Data Based on Skew-t Distributions. *Journal of Business and Economic Statistics*, 28:52–66, 2010.

R.E. Kass and L. Wasserman. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, 91:1343–1370, 1996.

J.P. Klein and M.L. Moeschberger. *Survival Analysis*. Springer, 1997.

S. Kullback and R.A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

K.L. Lange, R.J.A. Little, and J.M.G. Taylor. Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association*, 84:881–896, 1989.

P.S. Laplace. Mémoire sur la Probabilité deas Causes par les Événements. *Mémoires de Mathématique et de Physique Presentés é a l'Académie Royale des Sciences, Par Divers Savans, & Lûs dans ses Assemblées*, 6:621–656, 1774.

E. Ley and M.F.J. Steel. On the Effect of Prior Assumptions in Bayesian Model Averaging with applications to growth regression. *Journal of Applied Econometrics*, 24:651–674, 2009.

F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g-priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103:410–423, 2008.

J.G. Lin, J. Chen, and Y. Lin. Bayesian Analysis of Student t Linear Regression with Unknown Change-Point and Application to Stock Data Analysis. *Computational Economics*, 40:203–217, 2012.

D.V. Lindley. On the Measure of Information Provided by an Experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

D.V. Lindley. A Statistical Paradox. *Biometrika*, 44:187–192, 1957.

D.V. Lindley. Fiducial Distributions and Bayes' Theorem. *Journal of the Royal Statistical Society, Series B*, 20:102–107, 1958.

D.V. Lindley. *Bayesian Statistics: A Review*. Philadelphia, PA: SIAM, 1972.

B.G. Lindsay and K. Roeder. A Unified Treatment of Integer Parameter Models. *Journal of the American Statistical Association*, 82:758–764, 1987.

B. Liseo. Elimination of Nuisance Parameters with Reference Priors. *Biometrika*, 80:295–304, 1993.

B. Liseo. *The Elimination of Nuisance Parameters*. Number 25 in Handbook of Statistics: Bayesian Thinking, Modeling and Computation. Gulf Professional Publishing, 2005.

R.A. Maronna. Robust m-estimators of Multivariate Location and Scatter. *Annals of Statistics*, 4:51–67, 1976.

N. Merhav and M. Feder. Universal Prediction. *IEEE Transactions on Information Theory*, 44:2124—-2147, 1998.

R. Mukerjee and M. Ghosh. Second-order Probability Matching Priors. *Biometrika*, 84:970–975, 1997.

A. O'Hagan. On Outlier Rejection Phenomena in Bayes Inference. *Journal of the Royal Statistical Society, Series B*, 41:358–367, 1979.

A. O'Hagan. Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society, Series B*, 57:99–138, 1995.

H. Peers. On Confidence Sets and Bayesian Probability Points in the Case of Several Parameters. *Journal of the Royal Statistical Society, Series B*, 27:9–16, 1965.

H. Peers. Confidence Properties of Bayesian Interval Estimates. *Journal of the Royal Statistical Society, Series B*, 30:535–544, 1968.

J.M. Pérez and J.O. Berger. Expected-Posterior Prior Distributions for Model Selection. *Biometrika*, 89:491–511, 2002.

A.E. Raftery. Inference for the Binomial n Parameter: A Hierarchical Bayes Approach. *Biometrika*, 75:223–228, 1988.

F.P. Ramsey. *Studies in Subjective Probability*, chapter Truth and Probability. Wiley, New York, H.E. Kyburg and H.E. Smokler edition, 1964.

J. Rissanen. A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals of Statistics*, 11:416–431, 1983.

G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464, 1978.

J.G. Scott and J.O. Berger. Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *Annals of Statistics*, 38:2587–2619, 2010.

T. Sellke, M.J. Bayarri, and J.O. Berger. Calibration of p-values for Testing Precise Null Hypotheses. *The American Statistician*, 55:62–71, 2001.

T.A. Severini. On the Relationship Between Bayesian and non-Bayesian Interval Estimates. *Journal of the Royal Statistical Society, Series B*, 53:611–618, 1991.

G. Shafer. Lindley's Paradox. *Journal of the American Statistical Association*, 77: 325–334, 1982.

C.E. Shannon. A Mathematical Theory of Communication. *Bell Systems Technical Journal*, 27:379–423, 1948.

M.F.J. Steel. Bayesian Model Averaging and Forecasting. Unpublished, 2012.

R.W. Stracham and H.K. van Dijk. Bayesian Model Selection with Uninformative Prior. *Oxford Bulletin of Economics and Statistics*, 65:863–876, 2003.

T.J. Sweeting. On Predictive Probability Matching Priors. *Institute of Mathematical Statistics Collections*, 3:46–59, 2008.

T.J. Sweeting. Objective Priors: Criticism and Challenges. Seminar at University of Kent, School of Mathematics, Statistics and Actuarial Sciences, November 2011.

C. Villa and S.G. Walker. An Objective Approach to Prior Mass Functions for Discrete Parameter Spaces. 2013a. Submitted Paper.

C. Villa and S.G. Walker. An Objective Bayesian Criterion to Determine Model Prior Probabilities. 2013b. Submitted Paper.

C. Villa and S.G. Walker. Objective Prior for the Number of Degrees of Freedom of a t Distribution. *Bayesian Analysis*, 2013c. Accepted for Pubblication.

B. Welch and H. Peers. On Formulae for Confidence Points Based on Integrals of Weighted Likelihood. *Journal of the Royal Statistical Society, Series B*, 25: 318–329, 1963.

M. West. Outliers Models and Prior Distributions in Bayesian Linear Regression. *Journal of the Royal Statistical Society, Series B*, 46:431–439, 1984.

A. Zellner. On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. Goel and A. Zellner, eds.), pages 233–243, 1986. North-Holland, Amsterdam.

A. Zellner and A. Siow. Posterior Odds Ratios for Selected Regression Hypotheses. Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain), pages 585–603. University Press, Valencia, 1980.