Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Erik Hide Sæternes

Objective Beliefs and Bayes Estimators

An Approach to Parameter Estimation Using Objective Priors and Distance Between Distributions

Master's thesis in Industrial Mathematics

Supervisor: Gunnar Taraldsen

August 2020

NTNU
Norwegian University of Science and Technology

Erik Hide Sæternes

Objective Beliefs and Bayes Estimators

An Approach to Parameter Estimation Using Objective Priors and Distance Between Distributions

Master's thesis in Industrial Mathematics
Supervisor: Gunnar Taraldsen
August 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

This thesis deals with the problem of parameter estimation in statistical analysis, and in particular Bayesian analysis. Assuming a given model, the main focus is on choosing objective prior distributions for its parameters. The resulting priors are combined with different Bayes estimators in order to derive estimates of the true value. We derive and discuss the flat, Jeffreys and reference priors, in addition to the more novel Penalised Complexity prior. Loss functions for the Bayes estimators include Kullback–Leibler divergence and the Fisher information metric. The priors and Bayes estimators are compared with frequentist and fiducial approaches. The example chosen to illustrate the theory is the bivariate normal distribution with zero means and unit variances. A simulation study is done to see which estimators perform well.

Sammendrag

Denne oppgaven tar for seg problemet med å estimere parametre i statistisk analyse, og i særdeleshet Bayesiansk analyse. Gitt en modell utledes objektive a priori–fordelinger for parametrene. De resulterende fordelingene blir så kombinert med Bayes–estimatorer slik at en kan finne estimater for den sanne verdien av parametrene. Vi utleder og diskuterer flat, Jeffreys og referanse–fordelinger, samt den nyere Penalised Complexity–fordelingen. Tapsfunksjoner som diskuteres inkluderer Kullback–Leibler–divergens og Fisher–informasjons-metrikken. A priori–fordelingene og Bayes–estimatorene sammenlignes med frekventistiske og fiduse tilnærminger. Eksempelet som er valgt for å illustrere teorien er en bivariat normalfordeling med gjennomsnitt null og varians lik én for begge dimensjonene. En simuleringsstudie er inkludert for å belyse hvordan estimatorene presterer.

Preface

This thesis is the result of my work in the subject *TMA4900 – Industrial Mathematics, Master's Thesis* as a student at the Norwegian University of Science and Technology (NTNU). I would like to thank my supervisor, Gunnar Taraldsen, for the help he provided with finding relevant literature and in limiting the scope of the work, and for constructive feedback during meetings and useful comments on the final text. This thesis would surely have contained more errors and inaccuracies without written feedback from Silius Vandeskog, and I would therefore like to use the opportunity to thank him as well.

Erik Hide Sæternes
August 2020
Trondheim

Contents

Chapter 1

Introduction

A central feature of statistical analysis is the use of models that aim to capture the behaviour of some physical process. We have an observable quantity x which is usually a real number, taken either from the whole real line or a subset thereof. It might also be a vector of such values, in which case we say that the model is multivariate; as opposed to the univariate case when x is a scalar. Usually, we make several observations, collecting them in a vector which we name $\boldsymbol{x} = (x_1, ..., x_n)$, where x_i is observation number i for $i = 1, ..., n$.

Formal probability theory views these observed values $\boldsymbol{x}$ as realisations of a random variable X, which is a mapping from an underlying probability space to the domain of interest (that is, the set of values that we can observe). This underlying probability space is made up of a triplet consisting of a set Ω, a set of subsets of Ω satisfying the definition of a sigma algebra, and some measure (which we call a probability measure) satisfying certain conditions. The mapping X has to be a measurable function, and in this case it gives rise to a new probability space that among other things contains a probability measure induced by the probability measure of the underlying space. This probability measure gives the probability of observing different realisations x of the random variable X.

While such an abstract construction is necessary for mathematical rigour, most applications deal with a more restricted set of random variables that allows one to ignore these complications. In much of statistics, when dealing with real valued domains for X, the probability measure (more commonly called a *probability distribution*) induced by the random variable X has a probability density function given by

$$p(x) = \frac{d}{dx} P(X \leq x), \tag{1.1}$$

where we let $P(\cdot)$ denote the probability measure for the probability space induced by X. It might also happen that the domain of X is countable, and in such cases we use the probability mass function defined through $p(x) = P(X = x)$ for all $x \in X$.

Most users of statistical modelling define or choose the probability density function without ever considering the underlying probability space or distribution. Usually, such a function is actually a family of functions with a parameter θ (which might be a vector). We therefore let

$$p(x \mid \theta) \tag{1.2}$$

denote the probability density function, explicitly stating that a realisation x depends on the value of the parameter θ.

The models or probability distributions used in statistics might be quite simple and straightforward to analyse theoretically, or they might be highly complex and out of reach of analytic methods, in which case numerical approaches are needed. Regardless of complexity, a major goal is to determine the actual value of θ through observations $\boldsymbol{x}$, and in that way determine the exact form of the model within its class of densities. Once the exact model has been found, one might perform predictions about future observations, with estimates of the uncertainty related to such predictions. However, before any forecasting can be carried out, inference about the parameter needs to be performed. The determination of the value of the parameter, and the uncertainty related to this, is the task with which this thesis is concerned.

Looking at the parameter θ, the main approach taught in undergraduate statistics courses considers this quantity to be an unknown, unobservable number or vector. In simple terms there exists a correct value for the parameter θ, even though we are unable to observe it directly, and by collecting observations $\boldsymbol{x}$, we can estimate the parameter using the observations together with the model $p(x \mid \theta)$. Frequentist statistics uses this assumption, together with the idea that repeated experiments will asymptotically lead to exact probability estimates. An important and widely used example of a frequentist estimation procedure is the maximum likelihood estimator, which we will discuss later in this thesis.

Observations $\boldsymbol{x}$ are assumed to be described by a probability distribution. We might want to consider what happens if we let θ be described by a probability distribution as well. What we then do, is to approach the problem from a Bayesian perspective. The main reason for the nomenclature has to do with the foundational role played by Bayes' theorem. If we let θ be described by a density $\pi(\theta)$, where we for simplicity assume no

parameters, we can use Bayes' theorem to write

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{p(x)}, \tag{1.3}$$

where $p(x) = \int_{\Omega_\Theta} p(x \mid \theta)\pi(\theta)\, d\theta$, and Ω_Θ is the domain of the parameter.

We call the density $\pi(\theta)$ the prior of θ, and the corresponding density $\pi(\theta \mid x)$ the posterior. Once we have found $\pi(\theta \mid x)$, we can use this to make statements about the value of θ, as well as how certain we are about such a value. There are several ways to do this, and we will look closer at some of them in this thesis. Still, before we can perform inference of any type, we have to specify the density $\pi(\theta)$. This step is by no means a trivial one, and we will look closely at some approaches.

Why bother with the Bayesian approach? One key reason is that we often have some knowledge concerning the value of θ, and we would like to incorporate this knowledge into our analysis. Defining $\pi(\theta)$ in such a way that it captures some of this prior knowledge might improve the subsequent analysis. Another reason is computational. By using a Bayesian approach, use of various Monte Carlo methods for sampling is possible. This, along with powerful computers, lets us consider much more complex models. One other reason has to do with the teaching of introductory statistics. It might be argued (see e.g. J. O. Berger, 2006) that letting Bayesian statistics be the focus of courses aimed at novices would make the theory much more accessible.

Hence, there are reasons for preferring a Bayesian approach regardless of the presence of prior knowledge. If we have nothing on which to base our distribution for θ, could we find some general prior that would still let us use the Bayesian framework? This question will be thorougly discussed in this thesis.

In order to see how the theoretical constructions and results presented in this thesis perform, we will apply them to the bivariate normal distribution with known means and variances. More precisely, we will assume zero means and unit variances, and look at the resulting one-parameter model where the parameter is given by the correlation. This model, while simple, does require a lot of involved calculations, and some formulas do not seem to lend themselves to analytic solution. Consequently, numerical approximations and experiments are needed in order to fully investigate the behaviour of the distributions and estimation procedures.

The bivariate normal model (and indeed the multivariate normal model in general) has received a lot of attention and has been studied in depth. If we limit our focus to

Bayesian approaches, the list of publications is still quite extensive, and we will here give a sample of the available literature.

A thorough investigation into objective priors for the full (that is, all five parameters are unknown) bivariate normal model is given by J. O. Berger and Sun, 2008. Here, recommended priors for the different parameters are listed, with arguments focusing on frequentist matching. Kim, Kang, and Lee, 2009 looks at much the same problem, but make the assumption that the means are identical. Ghosh et al., 2010 chooses to limit the focus to the correlation coefficient, while still assuming all parameters to be unknown. As for J. O. Berger and Sun, 2008, coverage characteristics are central to their approach. Assuming known means equal to zero and unit variances, Fosdick and Raftery, 2012 uses frequentist and Bayesian methods to try and find a good estimator for the correlation coefficient. Criteria include looking at the mean square error between the estimated and true correlation, and using hypothesis tests for values of the correlation drawn uniformly from various intervals. Castro and Vidal, 2019 looks at the problem from a regression perspective and tries to estimate both the variances and the correlation, while assuming the means to be equal to zero.

The remainder of this thesis is organised as follows. Chapter 2 introduces ways to find point and interval estimates, using frequentist, Bayesian or fiducial approaches. Chapter 3 introduces objectivity when using Bayesian statistics, and discusses the notion of an objective prior distribution. After talking briefly about the problem of improper distributions, it goes on to define and investigate three approaches to the construction of default priors. Chapter 4 introduces the bivariate normal model and applies the theory of chapters 2 and 3 to it. Then, in chapter 5 we perform some simulations in order to investigate how well the various approaches to estimation perform on our model. Finally, chapter 6 discusses the content of the previous chapters, and looks at possible future research based on this.

Chapter 2

Point and Interval Estimators

Statistical analysis is usually divided into two subcategories: frequentist and Bayesian. Given an underlying model (i.e. probability distribution), which often has a probability density function $p(x \mid \theta)$, the goal is to observe realisations x of the random variable X in order to say something about the value of θ. A common feature for both frequentist and Bayesian analysis is hence the desire to derive estimates of parameter values, as well as intervals gauging the uncertainty of the parameter estimates, or simply how uncertain we are about the actual value of the parameters. We will here look at some approaches for both categories, and then include a brief discussion of a third approach, namely fiducial inference.

Before we delve into this section, however, we should make absolutely clear what we refer to when talking about the parameter. For a given model, one might refer to the whole vector of parameters when using the word parameter. However, oftentimes only some elements of the vector are of interest, or a transformation of the vector is what we want to look at. We then define the parameter to be $\gamma = \psi(\theta)$. An example might be the normal distribution with $\theta = (\mu, \sigma^2)^\top$; if we are only interested in the mean, $\psi(\cdot)$ is the projection down to the first dimension.

2.1 Frequentist inference

Frequentist methods rely on the basic assumption that the parameter of interest is an unknown number that is not directly observable. The goal, then, is to find estimators for this number given the data, as well as confidence intervals with a certain probability of covering the actual value of the parameter. We will here introduce the much used maximum likelihood estimator, and then discuss how estimators can be used to find confidence intervals through bootstrapping.

Given a parameter θ and data $\boldsymbol{x}$, a confidence interval is meant to provide a measure of the certainty we have about the parameter. Given $\alpha \in (0, 1)$, we want to find values

$L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ such that

$$P\left[L(\boldsymbol{x}) < \theta < U(\boldsymbol{x})\right] = 1 - \alpha, \tag{2.1}$$

for all possible θ. In words, we want to find an interval that has a probability $1 - \alpha$ of covering the actual parameter value. We call $1 - \alpha$ the confidence level. Common values for α are 0.05 and 0.025. Note that we can have either $L(\boldsymbol{x}) = -\infty$ or $U(\boldsymbol{x}) = \infty$ (or the left or right limits of the parameter domain), in which case we say that the interval is one-sided. It should be emphasised that, since equation (2.1) should hold for all θ, the random variable in this definition is the interval $(L(\boldsymbol{x}), U(\boldsymbol{x}))$, and not e.g. the parameter θ.

Bootstrapping refers to a broad category of methods in which one uses random sampling with replacement to get an empirical distribution, and through this be able to estimate e.g. mean, variance and confidence intervals.

2.1.1 Maximum likelihood estimate

The maximum likelihood estimate (MLE) is arguably the most used method for estimating parameter values in statistics (see e.g. Miura, 2011). Given observed data $\boldsymbol{x}$, the MLE is the value of θ that maximises the likelihood function $p(\boldsymbol{x} \,|\, \theta)$. A common assumption is that the observations are independent, which implies that the density $p(\boldsymbol{x} \,|\, \theta)$ can be written as $\prod_{i=1}^{n} p(x_i \,|\, \theta)$. Notable cases where we do not make such an assumption are time series analysis, in which past observations are usually assumed to influence future observations, and data from a spatial model, in which observations that are close to each other spatially are assumed to be highly correlated. Regardless, we will assume independent samples throughout this thesis.

From the assumption of independent samples, the likelihood is given by

$$L(\theta \,|\, \boldsymbol{x}) = \prod_{i=1}^{n} p(x_i \,|\, \theta), \tag{2.2}$$

with $\boldsymbol{x} = (x_1, ..., x_n)$.

Using the fact that the logarithm is a monotonically increasing function, it is common to define the log-likelihood as the natural logarithm of the likelihood function, and denote it by $\ell(\theta \,|\, \boldsymbol{x})$. Assuming that the likelihood is unimodal, the MLE can be found by solving

$$\frac{d}{d\theta} \ell(\theta \,|\, \boldsymbol{x}) = 0 \tag{2.3}$$

with respect to θ.

2.1.2 Interval estimation through bootstrapping

Once an estimator for the parameter has been found, we can use it to generate confidence intervals. When then distribution of the parameter is known (such as for the empirical mean of independent and identically distributed normal variables), a $1-\alpha$ confidence interval might be readily available by finding $\hat{\theta}_L$ and $\hat{\theta}_U$ such that

$$P\left[\hat{\theta} \in \left(\hat{\theta}_L, \hat{\theta}_U\right) \,\middle|\, \boldsymbol{x}\right] = 1-\alpha, \tag{2.4}$$

with $\hat{\theta}$, $\hat{\theta}_L$ and $\hat{\theta}_U$ in general depending on the data $\boldsymbol{x}$.

The exact distribution of most estimators are not known, thus creating a need for other methods of interval estimation. One such approach is appropriate when we can draw as many samples as we would like from the distribution of the data. In this way, we can estimate the parameter for as many different samples as we would like, and the result is an empirical distribution approximating the true distribution of the estimator. This empirical distribution can then be used to find an interval. More precisely, we find the α_1 and α_2 percentiles of the empirical distribution, with $\alpha_1 + \alpha_2 = \alpha$, and use these values to define an estimated confidence interval. A common approach is to let $\alpha_1 = \alpha_2 = \alpha/2$, in which case we have an equal-tailed interval. We call this approach *parametric bootstrapping*, and it is this method that we will utilise in this thesis.

As a side note, in cases where we do not know, and hance cannot sample from, the distribution, and only a single sample is observed, a different approach to bootstrapping might be used, in which the sample is resampled with replacement in order to generate "new" samples and find an empirical distribution. Since the example model in this thesis is assumed known, we will not make use of such an approach here.

2.2 Bayesian inference

In a Bayesian world the parameter is *not* simply assumed to be an exact, unknown number. Rather, the parameter itself is given a probability distribution and treated as a random variable. The marginal distribution of the parameter θ is called its prior distribution. Then, Bayes' theorem can be utilised to derive the distribution of θ conditioned upon the observed data – giving what is called the posterior distribution.

The fact that our knowledge of θ given observed data is captured in the posterior distribution means that different procedures for deriving point and interval estimators are needed, than those used in a frequentist setting. We will first introduce a rather simple estimator and argue why it is not likely to work well. Then we will look at Bayes estimators in general and discuss some examples. Lastly, we will present interval estimation using the posterior distribution.

2.2.1 Maximum a posteriori estimation

One extremely simple and straightforward estimate based on the posterior distribution is called maximum a posteriori (MAP). The MAP estimate is quite simply the parameter value for which the posterior distribution reaches its global optimum – obviously within the domain of the parameter. Hence, given data $\boldsymbol{x}$ and a posterior distribution $\pi(\theta \mid \boldsymbol{x})$, the MAP estimate is given by

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\, \pi(\theta \mid \boldsymbol{x}), \tag{2.5}$$

or the *mode* of the distribution.

Note that, since the location of the global optimum of the posterior distribution does not change when multiplying the distribution with anything that is constant with respect to θ, knowing the posterior up to a constant of proportionality is sufficient to be able to derive the MAP estimate. Or, simply put, we have that $\operatorname{argmax}_\theta \pi(\theta \mid \boldsymbol{x}) = \operatorname{argmax}_\theta p(\boldsymbol{x} \mid \theta)\pi(\theta)$.

There are some possible difficulties related to such an estimate. One obvious problem is related to multimodal densities, where the MAP estimate might give a rather biased estimate ignoring a lot of information provided by the posterior. Another problem arises when the posterior density is heavily skewed. The global maximum of the density might be far from the mean, and hence is likely to be biased. Due to potentially quite bad performance, especially for small sample sizes, we will not make use of the MAP estimate in the simulations in chapter 5.

Note that the problems discussed above for the MAP estimate are also true for the MLE of section 2.1.1. Indeed, for a flat prior, the two estimators are the same.

2.2.2 Bayes estimation

Most Bayes estimators differ from the MAP estimate in that they consider the whole posterior distribution, and not only the point of highest density. The starting point of a Bayes estimator is the definition of a loss function (sometimes called cost function) $L(\theta, \hat{\theta})$ which measures the discrepancy between the actual value of the parameter θ and its estimated counterpart $\hat{\theta}$. Here, $\hat{\theta}$ might be any estimator. Given the loss function, our goal is to find the best $\hat{\theta}$ available. Note that $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ in general depends on the observed data $\boldsymbol{x}$.

What we mean by best estimator, is that we would like to find an estimator $\hat{\theta}$ that minimises the expected loss function over both $\boldsymbol{x}$ and θ. That is, we would like to find

$$\hat{\theta}_{\mathrm{BE}} = \underset{\hat{\theta}}{\operatorname{argmin}} \int_{\Omega_\Theta} \int_{\Omega_X^n} L(\theta, \hat{\theta}) p(\boldsymbol{x}, \theta) \; d\boldsymbol{x} d\theta, \tag{2.6}$$

with $\Omega_X^n = \Omega_X \times ... \times \Omega_X$. By writing $p(\boldsymbol{x}, \theta) = \pi(\theta \,|\, \boldsymbol{x}) p(\boldsymbol{x})$ and changing the order of integration, we get

$$\hat{\theta}_{\mathrm{BE}} = \underset{\hat{\theta}}{\operatorname{argmin}} \int_{\Omega_X^n} \left(\int_{\Omega_\Theta} L(\theta, \hat{\theta}) \pi(\theta \,|\, \boldsymbol{x}) \; d\theta \right) p(\boldsymbol{x}) \; d\boldsymbol{x}, \tag{2.7}$$

and since $p(\boldsymbol{x}) \geq 0$ for almost every $\boldsymbol{x}$, we can reduce the problem to

$$\hat{\theta}_{\mathrm{BE}} = \underset{\hat{\theta}}{\operatorname{argmin}} \int_{\Omega_\Theta} L(\theta, \hat{\theta}) \pi(\theta \,|\, \boldsymbol{x}) \; d\theta. \tag{2.8}$$

Common examples of loss functions are the mean square error (MSE) and the absolute error, but other functions might also be used. In particular, functions measuring the distance between the distribution when using θ and $\hat{\theta}$ as parameter are reasonable choices. Next we will look closer at the MSE approach, as well as Bayes estimators with Kullback-Leibler divergence and Fisher information as loss functions.

Mean squared error

Assuming sufficient regularity, we can find the Bayes estimator by solving the equation

$$\frac{d}{d\hat{\theta}} \int_{\Omega_\Theta} (\hat{\theta} - \theta)^2 \, \pi(\theta \,|\, \boldsymbol{x}) \; d\theta = 0, \tag{2.9}$$

by interchanging the order of differentiation and integration to get

$$\begin{aligned}\int_{\Omega_\Theta} \frac{d}{d\hat{\theta}}(\hat{\theta}-\theta)^2\,\pi(\theta\,|\,\boldsymbol{x})\;d\theta &= 0\\ \int_{\Omega_\Theta} 2(\hat{\theta}-\theta)\,\pi(\theta\,|\,\boldsymbol{x})\;d\theta &= 0\\ \hat{\theta}\int_{\Omega_\Theta} \pi(\theta\,|\,\boldsymbol{x})\;d\theta &= \int_{\Omega_\Theta} \theta\pi(\theta\,|\,\boldsymbol{x})\;d\theta.\end{aligned} \tag{2.10}$$

Since we in general assume that we have $\int_{\Omega_\Theta} \pi(\theta\,|\,\boldsymbol{x})\;d\theta = 1$ – that is, the posterior is proper – we get

$$\hat{\theta}_{\mathrm{BE}} = \int_{\Omega_\Theta} \theta\,\pi(\theta\,|\,\boldsymbol{x})\;d\theta, \tag{2.11}$$

which means that the Bayes estimator when using the MSE as loss function, is given by the expected value of the posterior distribution. The Bayes estimator found when using the MSE as loss function is also called the minimum mean square error (MMSE) estimator.

Maximum a posteriori estimation

It could be noted that MAP estimate in section 2.2.1 can be viewed as the Bayes estimator when using a loss function given by

$$L_{\mathrm{MAP}}(\theta,\hat{\theta}) = \begin{cases} 0 & |\theta-\hat{\theta}| < \delta \\ 1 & |\theta-\hat{\theta}| \geq \delta, \end{cases} \tag{2.12}$$

with δ small and the posterior $\pi(\theta\,|\,\boldsymbol{x})$ smooth. That is, we are simply giving a constant penalty to all deviations larger than some $\delta > 0$. To see why this yields the MAP estimator, we first note that

$$\hat{\theta}_{\mathrm{BE}} = \underset{\hat{\theta}}{\operatorname{argmin}} \left(1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \pi(\theta\,|\,\boldsymbol{x})\;d\theta\right), \tag{2.13}$$

which is simply

$$\hat{\theta}_{\mathrm{BE}} = \underset{\hat{\theta}}{\operatorname{argmax}} \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \pi(\theta\,|\,\boldsymbol{x})\;d\theta. \tag{2.14}$$

Since we assume $\pi(\theta\,|\,\boldsymbol{x})$ to be smooth (and δ to be *small enough*), the maximum is found

at the global maximum of the posterior density, which is the MAP estimator.

Kullback–Leibler divergence

Instead of simply applying measures of distance between the parameter and its estimate, we might want to look at measures of deviation between the model with the true parameter, and the model with the estimated parameter. That is, we want to apply measures of distance between distributions, and not just real numbers.

An important measure in this regard is the Kullback–Leibler divergence. A complete introduction and definition will be given in section 3.4.1, so we will not go into further details here, other than to note that the Kullback–Leibler divergence is non-negative, and hence we do not need to square it, or make any other adjustments, when defining the loss function. Squaring or taking the square root of the Kullback–Leibler divergence might be interesting choices as well, and might lead to different and perhaps better results. When defining the Penalised Complexity prior in section 3.5, we do indeed use the square root of the Kullback–Leibler divergence. We have, however, not investigated this topic further.

Using Kullback–Leibler divergence, the loss function then becomes

$$L_{KL}(\theta, \hat{\theta}) = \kappa\left(p(x \mid \hat{\theta}) \,\middle|\, p(x \mid \theta)\right). \tag{2.15}$$

A possible closed form solution to the problem of finding this Bayes estimator will depend on the model used. In many cases, a closed form solution is unlikely to be readily available, and we will hence resort to the use of numerical optimisation techniques in order to find the estimate given observed data.

Fisher information metric

Continuing with measures of distance between probability distributions, another possibility is to use of the Fisher information, which we define next.

Definition 2.2.1 (Fisher information)**.** *Given a probability density function $p(x \mid \theta)$, the Fisher information is defined to be*

$$I(\theta)_{ij} = \mathbb{E}\left[\left(\frac{d}{d\theta_i}\log\left(p(x \mid \theta)\right)\right)\left(\frac{d}{d\theta_j}\log\left(p(x \mid \theta)\right)\right) \,\middle|\, \theta\right]. \tag{2.16}$$

In the one-parameter case, we get

$$I(\theta) = \mathbb{E}\left[\left(\frac{d}{d\theta}\log\left(p(x\,|\,\theta)\right)\right)^2 \,\middle|\, \theta\right]. \tag{2.17}$$

The Fisher information metric (see Taylor, 2019) considers the parameter as a vector (or a scalar in the one-parameter case) and finds the distance between two distributions by measuring the shortest distance between the parameter vectors. It should be noted that names like Fisher-Rao metric (see Rao, 2009), or Rao's distance measure (see Atkinson and Mitchell, 1981) also seem to be used for the same construction. The exact distance measure used is defined using the Fisher information, and makes use of geodesics from differential geometry (actually, this is part of the field of information geometry, in which statistics is paired with differential geometry – see e.g. Nielsen, 2018). Taylor, 2019 argues that the derivation of a closed form for distributions with an arbitrary number of parameters does in general seem unattainable. However, Taraldsen and Lindqvist, 2013 writes that by viewing the problem as being in L^2 space, the general solution will be the arc length of the unit sphere between the two parameter configurations.

Regardless, for one-parameter distributions the expression for the Fisher information metric is quite simple. A derivation can be found both in Taylor, 2019 and Atkinson and Mitchell, 1981. Our metric, and hence our loss function, becomes

$$L_I(\theta, \hat{\theta}) = \left|\int_\theta^{\hat{\theta}} \sqrt{I(t)}\; dt\right|, \tag{2.18}$$

with $I(\cdot)$ the Fisher information given in equation (2.17). Note that, as with the loss function given by the Kullback–Leibler divergence, we could have squared this or taken the square root. It would be of interest to investigate this further, but we have not done so in this thesis.

As with the Kullback–Leibler divergence, the existence of a closed form solution for the Bayes estimator, as well as the form of such a solution, will depend on the model used. In most cases, numerical optimisation techniques are likely to be necessary.

2.2.3 Interval estimation from posterior distributions

The posterior distribution allows us to derive credible intervals for the value of θ. By finding an interval $C(\boldsymbol{x}) = \big[\theta_L(\boldsymbol{x}), \theta_U(\boldsymbol{x})\big] \subset \Omega_\Theta$ such that

$$P\big[\theta \in C(\boldsymbol{x}) \,|\, \boldsymbol{x}\big] = 1 - \alpha, \tag{2.19}$$

we have a $1-\alpha$ credible interval. There are several strategies for finding $\theta_L(\boldsymbol{x})$ and $\theta_U(\boldsymbol{x})$, but throughout this thesis we will choose these limits such that $P\big[\theta \in \big(-1, \theta_L(\boldsymbol{x})\big) \,|\, \boldsymbol{x}\big] = P\big[\theta \in \big(\theta_U(\boldsymbol{x}), 1\big) \,|\, \boldsymbol{x}\big] = \alpha/2$ – that is, an equal-tailed interval.

Once such a credible interval has been found, the next question is whether or not it actually has a $100(1-\alpha)$ % chance of covering the true value of θ – that is, whether or not it can be treated as a confidence interval. The true probability of covering the parameter, which might not be $1-\alpha$, is called the *frequentist coverage* of the interval, and can be expressed as

$$P\big[\theta \in C(\boldsymbol{x}) \,|\, \theta\big], \tag{2.20}$$

where we condition on the true value of θ and let $\boldsymbol{x}$ be the random variable.

Ideally, we would like the frequentist coverage to be identical to, or at least very close to, the advertised probability of $1-\alpha$, and in these cases we say that the prior has *exact frequentist matching* (see J. O. Berger and Sun, 2008). In practice, however, this might fail spectacularly – especially for small sample sizes. In chapter 5 we will take a closer look at the frequentist coverage of the priors derived and tested in this thesis, and see how the performance changes with the size of the random sample.

2.3 Fiducial inference

Fiducial inference is often overlooked as an alternative to the more common frequentist and Bayesian approaches. As we shall see, a fiducial approach might work well for problems in which either frequentist or Bayesian approaches are normally taken, and hence including it in a statistical analysis might be worthwhile. Since the use of fiducial inference is of secondary interest in this thesis, the introduction will be quite brief, aiming only to cover enough for it to be analysed as an alternative to the other estimators with which this thesis is concerned.

A fiducial model (U, χ), as defined in Taraldsen and Lindqvist, 2018, is given by

$$x = \chi(u, \theta), \tag{2.21}$$

with x being observable, u being distributed according to $p(u \mid \theta)$ (where we have, for simplicity, assumed the distribution of U to have a probability density function), and θ some parameter. Oftentimes, we are interested in estimating the unknown parameter θ. If the density of u is independent of θ, we can simply solve (2.21) with respect to θ to get

$$\theta = \hat{\theta}(u, x). \tag{2.22}$$

If we observe x and simulate a sample u, we can calculate θ using equation (2.22). When doing this, we say that we are sampling from the fiducial distribution. Most often, as is common practice for prior and posterior distributions, we simply omit the word *distribution*, and say that we *sample from the fiducial*.

The fiducial might be thought of as a posterior distribution, but one without a corresponding prior distribution. Hence, we can apply the methods of point and interval estimation introduced in section 2.2 to the fiducial in the same way as we did to the posterior. This last point makes the fiducial approach natural if we are in a situation where we do not have any prior information – that is, we avoid the whole problem of choosing a prior, but can still use the Bayesian approach to estimation by viewing the fiducial as a posterior distribution.

Chapter 3

Objectivity and Priors

An inherent feature in Bayesian statistics, once a suitable model has been selected, is choosing prior distributions for the parameters of the model. Choosing among the myriad distributions available might seem like a daunting task. The realisation that the choice of prior distribution might severely affect the results of the inferences performed, and indeed completely overshadow the data for small sample sizes, makes the problem even more complex and important. Thus, one would ideally like to have clear guidelines as to how one should choose which prior distribution to use in a given setting. The fact of the matter is that such guidelines or heuristics depend heavily on the model in question, the prior knowledge of the problem, and the philosophical inclinations of the statistician making the choice. Further, a trade-off between theoretically sound arguments and practical applicability might help make the answer even less attainable.

In the end, one would ideally like to have default choices of prior distributions that works reasonably well (whatever that may mean) for a whole range of different problems, as this would remove the difficulties encountered when practitioners use less than ideal prior distributions due to lack of better judgement. The question is then, how do we approach this goal of defining a default prior distribution that can be used more or less blindly without having a tremendous impact on the results of the inference?

In some situations a parameter might have a clear and well-understood physical interpretation, and there might exist expert information which can be utilised to choose a prior that performs well for this specific problem. Instances where this might actually work are problems and models that are relatively simple and well understood and where there is a clear procedure for incorporating this knowledge and understanding into a prior distribution. Needless to say, such knowledge and procedures would be highly context dependent, and indeed difficult to translate to other situations. Even ignoring the utter lack of useable, context specific knowledge of the problem at hand, it is a fact that a lot of models contain an enormous amount of parameters having little or no reasonable interpretation that might be of help. Neural networks in machine learning is an easy example

in which the number of parameters might be in the millions, and the user might by no means be able to apply expert information to place prior distributions on the parameters. Further, situations arise where the user has no useful understanding of the problem context, preventing any reasonably accurate information from being utilised. The seeming impossibility of generalising the use of expert information to create workable priors that do in fact perform well, makes it clear that some other approach is needed.

3.1 Objectivity in a Bayesian setting

Creating a framework that provides default prior distributions for the wide range of applications found in statistics is one key goal of objective Bayesian analysis. The problem that needs to be resolved before any such framework can be found and agreed upon, however, is the actual meaning of the word *objective* in the setting of Bayesian analysis. While statistical modelling does include a number of arguably subjective choices, in particular with respect to choice of model, the common way to look at objective Bayesian analysis is by ignoring the subjectivity introduced by the model choices, or at least leave that part out of the discussion, and instead focus on the choice of priors for the parameters of the chosen model. The goal, then, is to set the prior in an objective way.

There are, obviously, some things that need to be adressed and clarified before we can actually find these priors. First of all, are we correct in ignoring choice of model in our definition of objective Bayesian analysis? It is inaccurate to say that the topic is not without controversy, but the general view seems to be that the suitability of the model can be tested through data, which helps give it a different theoretical status than the prior (see J. O. Berger, 2006; Consonni et al., 2018). In the end, statistical and scientific tradition seem to support the choice, and hence we will too.

Next, and more importantly, how do we set a prior in an objective way? For the past decades several different approaches to this question have been taken, with the focus shifting from conditions of invariance, frequentist coverage, and information-theoretic considerations. Before we take a closer look at some of the more notable attempts at developing objective priors, we end this section by stating the overarching goal of objective Bayesian analysis – what Consonni et al., 2018 likens to a "search of the 'philosopher's stone' for the Bayesian community". Ideally, we would like to properly define the idea of *knowing nothing*, and then let the prior reflect this state of ignorance on the part of the statistician. However, this ideal scenario has proved unattainable, and hence the

focus now is primarily to find priors that have minimal impact on the statistical analysis performed. In a way, this means that the prior should introduce no other information than what has already been introduced by the choice of model. It also incorporates the view that the data should be given as much weight as possible in the subsequent analysis, even for small sample sizes. While these goals seem reasonable enough and are mostly agreed upon, there is still no unanimous agreement on the definition and goals of objective Bayesian analysis. Due to the fact that priors, regardless of how they are defined, will always contain information, even the word *objective* is not accepted by the whole community of bayesians (see J. O. Berger, 2006; Consonni et al., 2018).

3.2 Dealing with improper priors

When specifying a prior distribution it might happen (in fact, it often does) that the distribution has infinite measure. What this means, is that it does not integrate to one and hence fails to fulfil one of the axioms of probability. One obvious solution might be to avoid the use of such *improper priors* altogether, and hence remove the problem entirely. Indeed, this seems to be the preferred solution according to Huisman, 2016, which attempts to define proper priors with much the same properties as the improper priors they are meant to replace. However, the fact remains that the usage of such priors is widespread and often useful, and a proper way to deal with them is necessary.

What we mean by a need to *deal with* improper priors, is a need for mathematical arguments defending the usage. After all, an improper prior is not a probability distribution, and it is unclear how prior knowledge is incorporated by such a prior. In most applications, the question of whether or not using improper priors is justifiable is simply ignored, so long as the corresponding posterior is proper. While such an approach, i.e. *if it works then it is fine*, does have a certain appeal, some argue the need for a more rigorous justification for the inclusion of improper distributions (see e.g. Taraldsen, Tufto, and Lindqvist, 2018 for one suggested solution to this issue).

One approach requires the posterior to be a limiting distribution of a sequence of posteriors derived from proper priors. What type of limit we are considering, and how this sequence of proper priors is defined varies. J. O. Berger, Bernardo, and Sun, 2009 defines the proper priors by restricting the improper prior to increasing compact subsets of the parameter space. They then use logarithmic convergence to define the limit. This is the approach followed in the construction of reference priors, which we will look at in

section 3.4. Another approach can be found in Bioche and Druilhet, 2016, where q-vague convergence is used to define the limit of proper priors.

Efforts toward constructing an axiomatic framework for probability theory that allows for improper posteriors are also made. In Taraldsen, Tufto, and Lindqvist, 2018, this is explained thoroughly, and arguments are given for why improper posteriors might capture the information provided by the data.

In the following, when deriving explicit prior and posterior densities, we will only verify whether the prior is proper or improper, and that the posterior is proper, without investigating the issue further.

3.3 Jeffreys prior

Jeffreys prior was first introduced by Harold Jeffreys in Jeffreys, 1997, and has long been in use as a non-informative prior distribution. Its definition is rather simple, and makes use of the Fisher information.

Definition 3.3.1 (Jeffreys prior)**.** *Jeffreys prior is defined to be proportional to the square root of the determinant of the Fisher information. That is, for probability density $p(x \mid \theta)$ we get*

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \tag{3.1}$$

with $I(\theta)$ the Fisher information from definition 2.2.1. In the one-parameter case, we simply have that $\pi(\theta) \propto \sqrt{I(\theta)}$.

For multiparameter distributions, use of the Jeffreys prior is normally avoided, as it may lead to undesirable behaviour. Indeed, J. O. Berger, Bernardo, and Sun, 2015 states that they "know of no multivariable example in which [they] would recommend the Jeffreys-rule prior".

The simplicity of Jeffreys prior is one of its key attributes, and one of the reasons it has been used so extensively. Another important aspect, which was central to Jeffreys' notion of an objective prior, is that the prior is invariant to injective transformations. What this means, is that the the prior density derived by first finding the prior for the original parameter and then using the change of parameter formula to switch to the new parameter, should give the same result as simply computing the prior density for the new parameter directly. Intuitively, this makes sense; the prior density should not depend on whether we parametrise using scale or rate for the gamma distribution, success

probability or odds for the binomial function, or variance, standard deviation or precision for the normal distribution.

Theorem 3.3.1 (Invariance of Jeffreys prior)**.** *Jeffreys prior, as defined in definition 3.3.1, is invariant to injective transformations.*

Proof. We will restrict attention to the one-parameter case here. The multiparameter version can be proved similarly, albeit with a bit more cumbersome notation. Let $\phi = g(\theta)$ be an injective transformation. Jeffreys prior for ϕ becomes

$$\begin{aligned}
&\pi(\phi) \\
&= \sqrt{I(\phi)} \\
&= \mathbb{E}\left[\left(\frac{d}{d\phi}\log\left(p(x \mid \phi)\right)\right)^2 \,\middle|\, \phi\right]^{1/2} \\
&= \mathbb{E}\left[\left(\frac{d\theta}{d\phi}\frac{d}{d\theta}\log\left(p(x \mid \phi)\right)\right)^2 \,\middle|\, \phi\right]^{1/2} \\
&= \mathbb{E}\left[\left(\frac{d}{d\theta}\log\left(p(x \mid \theta)\right)\right)^2 \,\middle|\, \theta\right]^{1/2}\left|\frac{d\theta}{d\phi}\right| \\
&= \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right| \\
&= \pi(\theta)\left|\frac{d\theta}{d\phi}\right|,
\end{aligned} \tag{3.2}$$

where we have used that conditioning on ϕ and θ gives the same result (since $g(\cdot)$ was assumed to be injective). Hence, the prior for ϕ calculated using definition 3.3.1 gives the same result as using the transformation of variables formula on the prior for θ, and the proof is done. □

3.4 Reference prior

The concept of a reference prior was first introduced by José Bernardo in 1979 (see Bernardo, 1979), and has since become one of the preferred frameworks for deriving objective priors. The reference prior offers a complete, rigorous mathematical foundation in support of its definition of objectivity with respect to the prior. Indeed, Simpson et al., 2017 states that "the reference prior framework is the only complete framework for specifying prior distributions". However, we shall see that it nonetheless fails to offer all the desirable features one would hope an objective prior to possess. In fact, the reference

prior is not unique for distributions with more than one parameter, but rather depends on the ordering of the parameters when deriving the prior. Regardless of this last point, the reference prior seems to be one of very few obvious choices to consider when trying to decide on the prior with which to equip the parameter of interest.

The goal of the reference prior is to derive reference posteriors that, in the words of Bernardo, "approximately describe the inferential content of the data without incorporating any other information" (see Bernardo, 1979). In a way, this means that the prior should have as little impact as possible on the results of the inferences performed. Such a goal requires one to specify mathematically what *as little impact as possible* actually means, and hence the theory needed becomes quite involved.

Due to the mathematical nature of the reference prior, some concepts and definitions need to be introduced before the actual definition of a reference prior can be formulated. Since we only deal with one-parameter distributions in this thesis, we will make the simplifying assumption that *the parameter is a scalar*, and define the reference prior with this in mind. This section is, in large parts, based on J. O. Berger, Bernardo, and Sun, 2009.

3.4.1 Kullback–Leibler divergence

Kullback–Leibler divergence is a form of distance measure between probability distributions first introduced in Kullback and Leibler, 1951. Its usage has since become widespread, and it is central to the definition of a reference prior. Hence, we start our tour towards a proper definition of a reference prior by defining this distance measure in definition 3.4.1. Note that although the definition of Kullback–Leibler divergence allows for probability measures that are not absolutely continuous (see Mathematics, n.d.(a)) with respect to the Lebesgue measure, we will here make the simplifying assumption that the probability measures possess the familiar kind of probability densities. More concretely, for a probability measure P, we have that $dP = f(x)dx$ for some positive, integrable (in the Lebesgue sense) function f. This assumption will be valid for the remainder of section 3.4.

Definition 3.4.1 (Kullback–Leibler divergence)**.** *Given two probability distributions with densities f and g satisfying the condition that $f(x) = 0$ whenever $g(x) = 0$, the Kullback–*

Leibler divergence between them is defined as

$$\kappa\left(f \,|\, g\right) = -\int_{\Omega_X} g(x) \ln\left(\frac{f(x)}{g(x)}\right) dx. \tag{3.3}$$

The Kullback–Leibler divergence might also be written as

$$\kappa\left(f \,|\, g\right) = -\mathbb{E}_g\big[\ln(f) - \ln(g)\big], \tag{3.4}$$

where $\mathbb{E}_g[\cdot]$ is the expectation taken with respect to the probability distribution having probability density function g. Hence, what we are really considering here is the expected value of the difference between the logarithms of the densities (taken with respect to one of the densities). Small deviations between f and g will hence lead to smaller values of the Kullback–Leibler divergence, whereas larger deviations will have the opposite effect.

The Kullback–Leibler divergence possesses some properties that help explain why it is so useful. First of all, it is always non-negative.

Theorem 3.4.1 (Non-negativity)**.** *The Kullback–Leibler divergence is non-negative.*

Proof. By assumption $f(x)/g(x)$ will never blow up, since we will never have $g(x) = 0$ and $f(x) \neq 0$. Hence, the fraction is bounded and non-negative, except for the situation in which $g(x) = f(x) = 0$. When this happens, we use the convention that $0 \cdot \ln(0/0) = 0$, and hence everything is well-defined. If for some subset of Ω_X of positive measure we have that $f(x) = 0$ and $g(x) \neq 0$, the Kullback–Leibler divergence becomes equal to positive infinity. When this is not the case, we can use the concavity of the logarithm together with Jensen's inequality (see Mathematics, n.d.(d)) to show that

$$\begin{aligned}
-\kappa\left(f \,|\, g\right) &= \int_{\Omega_X} g(x) \ln\left(\frac{f(x)}{g(x)}\right) \, dx \\
&\leq \ln\left(\int_{\Omega_X} g(x) \frac{f(x)}{g(x)} \, dx\right) \\
&= \ln\left(\int_{\Omega_X} f(x) \, dx\right) \\
&= \ln(1) \\
&= 0,
\end{aligned}$$

which completes the proof. □

Further, the Kullback–Leibler divergence is zero if and only if the distributions f and g are equal almost everywhere.

Theorem 3.4.2 (Vanishing Kullback–Leibler divergence)**.** *The Kullback–Leibler divergence vanishes if and only if $f = g$ almost everywhere.*

Proof. Let us first prove the forward implication. Assume the Kullback–Leibler divergence equals zero. For all $x \geq 0$ we have that $e^x \leq 1 + x$. Substituting x with $\ln(x)$, we get $\ln(x) \leq x - 1$, with obvious equality if $x = 1$. Moving $\ln(x)$ over to the right-hand side and differentiating the resulting expression, we see that the derivative is strictly negative for all $x < 1$ and strictly positive for all $x > 1$. This implies that $x = 1$ is the only point for which we get equality. Since both f and g are positive functions, we get

$$-\int_{\Omega_X} g(x) \ln\left(\frac{f(x)}{g(x)}\right)\, dx \geq -\int_{\Omega_X} g(x)\left(\frac{f(x)}{g(x)} - 1\right)\, dx = 0,$$

with equality if and only if

$$\ln\left(\frac{f(x)}{g(x)}\right) = \frac{f(x)}{g(x)} - 1$$

for almost all x. From the above, we therefore get (for almost all x)

$$\frac{f(x)}{g(x)} = 1 \implies f(x) = g(x).$$

Now for the backward implication. Assume $f = g$ almost everywhere. It follows that $\ln(f/g)$ is equal to zero almost everywhere. Since g is integrable, it follows that the integral is equal to zero as well. □

These properties make it possible to use the Kullback–Leibler divergence as a measure of distance between distributions. It should, however, be noted that the we do not have symmetry, and hence the Kullback–Leibler divergence does not define a metric.

Further, the Kullback–Leibler divergence is invariant to injective transformations. As discussed for the Jeffreys prior in section 3.3, this property is highly desirable if we want to define a prior based on this distance measure.

Theorem 3.4.3 (Invariance to injective transformations)**.** *The Kullback–Leibler divergence is invariant to injective transformations.*

Proof. Due to a desire for brevity, we will omit the proof. Note, however, that it can be done quite easily by applying the transformation of variables formula to the definition of the Kullback–Leibler divergence given by equation (3.3). □

3.4.2 Expected convergence

Using the Kullback–Leibler divergence, we can further define a concept of convergence of a sequence of probability distributions. We call this concept logarithmic convergence, and define it concisely in definition 3.4.2. While the name *logarithmic convergence* might not seem especially obvious or reasonable, it is the one used by J. O. Berger, Bernardo, and Sun, 2009, and hence we will not do otherwise here.

Definition 3.4.2. *A sequence of probability distributions with probability density functions given by p_i, $i \in \mathbb{N}$, converges logarithmically to a probability distribution with probability density function p if and only if we have that*

$$\lim_{i \to \infty} \kappa\left(p_i \,|\, p\right) = 0. \tag{3.5}$$

When dealing with prior distributions that are improper, it is argued in J. O. Berger, Bernardo, and Sun, 2009 that one needs some form of justification before using Bayes' theorem to derive the posterior, even though the posterior turns out to be a proper probability distribution. Thus, next, we turn to the notion of an approximating sequence of probability distributions. What we would like, is for our prior to be the limit of such an approximating sequence, and then use this as a justification for using Bayes' theorem.

Definition 3.4.3 (Approximating sequence of distributions)**.** *For some probability distribution over the space Ω_Θ with probability density function $p(\theta)$, $\theta \in \Omega_\Theta$, let $\{\Omega_{\Theta_i}\}_{i \in \mathbb{N}}$ be such that $\Omega_{\Theta_i} \subset \Omega_{\Theta_j}$ for $i < j$, and $\cup_{i=1}^{\infty} \Omega_{\Theta_i} = \Omega_\Theta$. We call $\{\Omega_{\Theta_i}\}_{i \in \mathbb{N}}$ an approximating compact sequence. Further, let $p_i(\theta)$ be equal to $p(\theta)$ restricted to Ω_{Θ_i}, that is*

$$p_i(\theta) = \frac{p(\theta)}{\int_{\Omega_{\Theta_i}} p(t)\ dt}. \tag{3.6}$$

Then, $\{p_i(\theta)\}_{i \in \mathbb{N}}$ is the approximating sequence of distributions to $p(\theta)$.

Suppose we have a prior density $\pi(\theta)$ with an approximating sequence of priors $\{\pi_i(\theta)\}_{i \in \mathbb{N}}$ defined as in definition 3.4.3, and suppose the prior satisfies $\int_{\Omega_\Theta} p(x \,|\, \theta)\pi(\theta)\ d\theta < \infty$. Then, the corresponding approximate sequence of posteriors given by $\pi_i(\theta \,|\, x) \propto p(x \,|\, \theta)\pi_i(\theta)$ converges logarithmically to the posterior $\pi(\theta \,|\, x)$. A proof of this point can be found in e.g. J. O. Berger, Bernardo, and Sun, 2009.

It turns out that this last result is insufficient, as the pointwise convergence with respect to the data x might cause difficulties in some situations (see Fraser, Monette, and Ng, 1985). Hence, we take expectations in order to gain a stronger, global form of convergence of posteriors.

Definition 3.4.4 (Expected logarithmic convergence of posteriors)**.** *Given a prior $\pi(\theta)$ and an approximating sequence $\{\pi_i(\theta)\}_{i\in\mathbb{N}}$, we say that the corresponding sequence of posteriors $\{\pi_i(\theta \mid x)\}_{i\in\mathbb{N}}$ converges logarithmically to the posterior $\pi(\theta \mid x)$ in expectation if*

$$\lim_{i\to\infty} \int_{\Omega_X} \kappa\left(\pi_i(\cdot \mid x) \mid \pi(\cdot \mid x)\right) p_i(x)\ dx = 0, \tag{3.7}$$

with $p_i(x) = \int_{\Omega_{\Theta_i}} p(x \mid \theta)\pi_i(\theta)\ d\theta$, and Ω_X the space of the data.

3.4.3 Permissible priors

When choosing a prior distribution, we need to know from which subset of the whole set of existing distributions we are choosing. Some distributions should not even be worth considering, and hence placing some limitations on the range of possible priors seems reasonable. First of all it makes sense for the prior to be strictly positive on the whole domain, since otherwise it would give zero weight to certain parameter values. [continuity]

Definition 3.4.5 (Permissible prior)**.** *If a function $\pi(\theta)$ satisfies the conditions*

1. *$\pi(\theta)$ is continuous,*
2. *$\pi(\theta) > 0$ for all $\theta \in \Omega_\Theta$,*
3. *$\int_{\Omega_\Theta} p(x \mid \theta)\pi(\theta)\ d\theta < \infty$, and*
4. *there exists some approximating sequence $\{\pi_i(\theta)\}_{i\in\mathbb{N}}$ such that the corresponding posteriors $\pi_i(\theta \mid x)$ converge logarithmically in expectation to the posterior $\pi(\theta \mid x)$, as defined in definition 3.4.4,*

we say that it is a permissible prior function for our model.

3.4.4 Expected information

How can we quantify the information gain about the parameter θ from observing data x? Using some measure, the distance between the prior $\pi(\theta)$ and the posterior $p(\theta \mid x)$

should contain the added information from conditioning on the oberved x. In order to make this independent of the actual x observed, we take the expected value with respect to x. Perhaps not surprisingly, the distance measure chosen will be the Kullback–Leibler divergence.

Definition 3.4.6 (Expected information)**.** *The expected information obtained from making one observation of x from the model $p(x \mid \theta)$, using a prior $\pi(\theta)$ is*

$$I\big(\pi \mid p\big) = \int_{\Omega_X} \kappa\left(\pi(\theta \mid x) \mid \pi(\theta)\right) p(x)\ dx, \tag{3.8}$$

with $p(x) = \int_{\Omega_\Theta} p(x \mid \theta)\pi(\theta)\, d\theta$.

With definition 3.4.6 giving us a way to measure the amount of information gained from making one observation of x, we would now like to look at what happens when we make more than one observation, and in particular what happens when the number of observations grows large. For $\boldsymbol{x} = (x_1, ..., x_n)$, denoting n independent and identically distributed observations, we let

$$I(\pi \mid p^n) = \int_{\Omega_X^n} \kappa\left(\pi(\theta \mid \boldsymbol{x}) \mid \pi(\theta)\right) p(\boldsymbol{x})\ d\boldsymbol{x}, \tag{3.9}$$

with $\Omega_X^n = \Omega_X \times ... \times \Omega_X$ n times, be the expected information from the whole sample. We would like our prior to be the one that gives the largest increase in information when we make an observation x. Hence, what we want is to find the prior $\pi(\theta)$ that maximises the expected information for large n. A prior that does this, is said to have the Maximising Missing Information (MMI) property.

Some care must be taken, since for continuous parameter spaces an infinite amount of information would be required to determine the parameter value with absolute certainty, and hence the quantity $I(\pi \mid p^n)$ should be expected to diverge. The following definition gives precise meaning, in the one-parameter case, to what maximising the missing information entails.

Definition 3.4.7 (Maximising Missing Information (MMI) Property)**.** *Assume we are given a model $p(x \mid \theta)$ with one continuous parameter, and a class of priors $\mathcal{P}$ yielding proper posteriors. For any compact set $\Omega_{\Theta_0} \subset \Omega_\Theta$ and any $q \in \mathcal{P}$, if a function $\pi(\theta)$ is such that*

$$\lim_{n\to\infty} \left(I(\pi_0 \mid p^n) - I(q_0 \mid p^n)\right) \geq 0, \tag{3.10}$$

where π_0 and p_0 are the renormalised restrictions (as seen in definition 3.4.3) of π and p to the compact subset of Ω_Θ, we say that the function π has the MMI property for the given model and class of priors.

3.4.5 Defining a reference prior

The definition of a reference prior for a one-parameter model can now be formulated.

Definition 3.4.8. *If a prior distribution $\pi(\theta)$ satisfies the MMI property of definition 3.4.7, as well as being a permissible prior as stated in definition 3.4.5, it is a reference prior.*

Next, we turn to a method of deriving an explicit form for the reference prior for a given model, provided that we are in the one-parameter case.

3.4.6 Explicit solution for one-parameter distributions

In the case of one-parameter distributions, including multi-parameter distributions for which all but one parameter are assumed known, there exists an explicit method for deriving the reference prior.

Theorem 3.4.4 (Explicit form of reference prior)**.** *Assume that we have a random sample $\boldsymbol{x} = (x_1, ..., x_n)$, and some asymptotically sufficient statistic $t_n = t_n(\boldsymbol{x})$ with $t_n \in T_n$. Let $\pi^*(\theta)$ be a continuous, strictly positive function such that the posterior $\pi^*(\theta \mid t_n)$ is proper. Let $\theta_0 \in \Omega_\Theta$ be any interior point, and define*

$$f_n(\theta) = \exp\left\{ \int_{T_n} p(t_n \mid \theta) \ln\left(\pi^*(\theta \mid t_n)\right)\, dt_n \right\} \tag{3.11}$$

with

$$\pi^*(\theta \mid t_n) = \frac{p(t_n \mid \theta)\pi^*(\theta)}{\int_{\Omega_\Theta} p(t_n \mid \theta)\pi^*(\theta)\, d\theta}, \tag{3.12}$$

and

$$f(\theta) = \lim_{n\to\infty} \frac{f_n(\theta)}{f_n(\theta_0)} \tag{3.13}$$

If the function $f(\theta)$ is a permissible prior, then $cf(\theta)$ is a reference prior for any $c > 0$.

Proof. A proof of this theorem is outside the scope of this thesis, but can be found in the appendices of J. O. Berger, Bernardo, and Sun, 2009. □

Applying theorem 3.4.4 would seem to make finding the reference priors for one-parameter distributions an easy task. This is not necessarily the case. The actual calculation might be quite tedious, or even intractable, and hence we introduce a numerical scheme for estimating the reference prior.

3.4.7 Numerical computation of one-parameter reference prior

Algorithm 1 is based on the numerical scheme presented in J. O. Berger, Bernardo, and Sun, 2009, setting $\pi^*(\theta) = 1$, which is based on the method presented in theorem 3.4.4.

Algorithm 1 Computing reference priors numerically

Input: n_θ = number of values of θ (assumed here to be evenly spaced in the domain – but this does not have to be the case). n_i = number of iterations per value of θ. n_d = number of data points per iteration.

for i in 1 to n_θ **do**

Set θ equal to value number i in list of parameter values

for j in 1 to n_i **do**

Sample n_d realisations from the model $p(x \mid \theta)$ giving the data $\boldsymbol{x} = (x_1, ..., x_{n_d})$

Compute $c_j = \int_{\Omega_\theta} \prod_{k=1}^{n_d} p(x_k \mid \theta)\ d\theta$

Evaluate $r_j(\theta) = \log\left[\prod_{k=1}^{n_d} p(x_k \mid \theta)/c_j\right]$

end for

$\pi(\theta) = \exp(n_i^{-1} \sum_{j=1}^{n} r_j(\theta))$

Add $(\theta, \pi(\theta))$ to list of pairs

end for

return List of pairs $(\theta, \pi(\theta))$

3.5 Penalised Complexity prior

A more recent addition to the list of frameworks for choosing default priors is the Penalised Complexity priors (shortened to PC priors). The PC prior concept was introduced in Simpson et al., 2017, and have since been incorporated into the R-INLA software package (see source code in Rue, n.d.). PC priors are meant to provide general default choices of prior distributions, and its main usage is with complex, hierarchical models. In order to achieve this, mathematical rigour has been sacrificed for computational tractability, and hence it stands in some contrast with e.g. reference priors, which rely on a more solid

theoretical foundation, but remains difficult to use in practise.

3.5.1 General idea

Although this thesis concerns itself with objective priors in a Bayesian setting, the PC priors do not conform to this standard. In fact, the PC priors are not going to be objective, but rather contain some weak information based on a few underlying principles deemed reasonable by the authors of Simpson et al., 2017. The main reasoning behind the choice to inject or allow some deviation from a pure non-informative prior, goes as follows: An overparametrised model will likely yield over-fitting when using an objective prior. This, in addition to the fact that objective priors perform badly from a computational point of view, leads to the conclusion that the injection of some information is desirable.

At the heart of the PC prior lies a base model, which is in general model specific. In a way, the base model is the simplest possible model in the class of possible priors, and the introduction of such a concept helps to avoid the unwanted over-fitting often seen when using objective priors. An example to illustrate the base model concept might be given by the a univariate normal distribution with zero mean – that is, the parameter of interest is the variance. For the base model, the variance is set to zero, consequently placing all mass at zero. (This is, obviously, a degenerate distribution, and it shows that the base model might often be regarded as the limit of some sequence of distributions – in this case, $\lim_{i\to\infty} \mathrm{N}(0, \sigma_i^2)$ with $\sigma_i \to 0$ when $i \to \infty$.) As a further example, for a Student's t-distribution with degrees of freedom as its parameter of interest, the base model is typically given by setting degrees of freedom equal to infinity. It thus becomes even more apparent that the base model is, oftentimes, given by an asymptotic, perhaps idealised, case.

The desire to avoid over-fitting can be seen as a preference for parsimonious models, often called Occam's (or Ockham's) razor principle (see Jefferys and J. O. Berger, 1992). Without a clear indication to the contrary, the prior should lean towards the simpler models. The reason for needing a base model hence becomes clearer, as it supplies a means of measuring the deviation from a parsimonious model. If the evidence does not suggest a large deviation from the base model, the PC prior will prefer to stay close to it. If, on the other hand, the data strongly indicate a deviation from the simpler model, the complexity of the resulting model will be higher.

3.5.2 One-parameter distributions

Given a model and corresponding base model, we need to calculate the complexity of the model with respect to the base model using some measure of distance. For the PC priors, the choice of measure is the Kullback–Leibler divergence. Since we would like the prior to have less weight for higher complexity models (i.e. models farther from the base model), the measure of complexity has to be interpretable as a distance, and Simpson et al., 2017 argues that the proper way to do this is by letting the distance measure used for the PC prior, for given probability densities f and g, be given by

$$d(f \mid g) = \sqrt{2\kappa\,(f \mid g)}. \tag{3.14}$$

Since f and g are going to be models from the same family, but with the parameter of g set to $\xi = 0$ (or some other value defining the base case), we will in the following just write $d(\xi)$. We will then provide $d(\xi)$ with a prior distribution, which in turn leads to a distribution for the parameter of interest ξ.

It should be noted that the distance measure given by equation (3.14) is not symmetric, since the Kullback–Leibler divergence is not symmetric, and hence is not a distance in the normal way.

In order to derive the prior distribution, we need one further assumption. It will be assumed that the penalisation of the distance $d > 0$ from the base model is given by r^d for $r \in (0, 1)$. This implies that increasing the distance from d to $d + \delta$ for some $\delta > 0$ is independent of d, which can be written as

$$\frac{\pi_d(d + \delta)}{\pi_d(d)} = r^\delta. \tag{3.15}$$

The implication of such an assumption is that the prior distribution for d is exponentially distributed (see Simpson et al., 2017), which we prove in the following theorem.

Theorem 3.5.1. *The prior distribution for the distance d, defined by equations (3.14) and (3.15), follows an exponential distribution.*

Proof. We first note, using equation (3.15), that

$$\int_s^\infty \pi(x)\; dx = r^s \int_0^\infty \pi(x)\; dx = r^s. \tag{3.16}$$

Further, we have that

$$\int_{s+t}^{\infty} \pi(x)\ dx = r^s \int_t^{\infty} \pi(x)\ dx = \int_s^{\infty} \pi(x)\ dx \cdot \int_t^{\infty} \pi(x)\ dx. \tag{3.17}$$

Letting $f(s) = \int_s^{\infty} \pi(x)\ dx$, we get

$$f(s+t) = f(s)f(t), \tag{3.18}$$

and we have that $f(a) = f(a/2)^2 = f(a/4)^4 = ... = f(1)^a$. The only continuous function satisfying this equation is

$$f(a) = e^{-\ln(f(1))a} = e^{-\ln(r)a}. \tag{3.19}$$

We now note that, letting $\lambda = -\ln(r)$, we have the equality

$$\int_0^a \pi(x)\ dx = 1 - e^{-\lambda a}, \tag{3.20}$$

which is the cumulative distribution function for an exponential distribution. Hence, the prior $\pi(d)$ has to be exponentially distributed, and we are done. □

For the parameter ξ of interest, of which $d = d(\xi)$ is a function, we hence get

$$\pi(\xi) = \lambda e^{-\lambda d(\xi)} \left| \frac{\partial d(\xi)}{\partial \xi} \right|, \tag{3.21}$$

by using the transformation of variables formula, assuming a one-to-one correspondence.

One last item remains before the univariate PC priors are fully specified; the value of the parameter λ must be chosen. This is, in some sense, where the prior becomes weakly informative, as the choice of value for this parameter might be used to control the mass in the tail of the prior, and hence decide how much prior information should be inserted.

The value of λ might be chosen in the following way. Suppose we define a transformation $Q(\cdot)$ of ξ (this might be the identity function) and a threshold U such that $Q(\xi) > U$ is considered a tail event. Let α denote the probability of being in the tail. Then we need to choose the value of λ for the PC prior in such a way that

$$P\left(Q(\xi) > U\right) = \alpha. \tag{3.22}$$

Note that λ will be a function of U and α, given the choice of Q, and hence what we

need to decide is the value of these two parameters. That is, we need to establish what we mean by a tail event, and how much weight we would like to place on such an event.

Another approach, which is not a part of the definition of PC priors given by Simpson et al., 2017, would be to place a prior distribution on λ. Since λ is just the rate of an exponential distribution, one natural choice is the prior $\pi(\lambda) = 1/\lambda$. While investigating this further does seem worthwhile, we have not done so in this thesis.

Chapter 4

Bivariate Normal with Unknown Correlation

This chapter will introduce the bivariate normal distribution, assuming only the correlation coefficient to be unknown. First we will supply a section giving an overview of the content. Then, we take a look at the frequentist approach to the problem of estimating the correlation. After this, we turn to the main theme of this thesis, namely the derivation and subsequent discussion of possible prior distributions for the unknown parameter. Then, we derive expressions for the Bayes estimators when using Kullback–Leibler divergence and the Fisher information metric as loss functions. Lastly, we include some other approaches to estimating the correlation.

4.1 Overview

This section is meant to give an overview of the example model and the results of this chapter.

4.1.1 The model

Assume that we are given a bivariate normal distribution with mean equal to $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top = (0, 0)^\top$ and covariance matrix given by

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{4.1}$$

From Σ we see that the variances are both equal to 1, that is $\sigma_1 = \sigma_2 = 1$, and we are left with a single parameter ρ which denotes the correlation between the two components of the random vector. Note that we assume $\rho \in (-1, 1)$, since $\rho = \pm 1$ would make this a degenerate distribution (see Mathematics, n.d.(c)) and essentially reduce it to the univariate case. This is the same approach as that taken in J. O. Berger and Sun, 2008.

The probability density function is given by

$$p(x \mid \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)}\right). \tag{4.2}$$

When talking about a sample of size n, we will in the remainder of this section let $\boldsymbol{x} = (x_1, ..., x_n)$ denote the data, with each $x_i = (x_{i1}, x_{i2})$ being a single realisation from the bivariate normal distribution, for $i \in \{1, ..., n\}$, with some specified correlation ρ. Note that when deriving prior distributions, we will use the model $p(x \mid \rho)$, in which $x = (x_1, x_2)$ is just one single realisation of the bivariate normal. The distinction should be clear both from context, and from the bold (i.e. $\boldsymbol{x}$) vs. normal (i.e. x) notation for the data.

4.1.2 Frequentist results

In section 4.2 we show that there exists a two-dimensional minimal sufficient statistic for ρ given by equation (4.5). Further, the maximum likelihood estimator is shown to be the solution to the cubic equation given by equation (4.13). In addition to the MLE, the empirical correlation is introduced in equation (4.8). Since we assume the means and variances to be known, variants of the empirical correlation incorporating some or all of this information are given in equations (4.9) and (4.10). Both the MLE and (the variants of) the empirical correlation might be used as an estimator for the correlation coefficient. In order to limit the number of estimators used in the simulation studies, we will only use the MLE and the version of the empirical correlation given by (4.9) in chapter 5.

4.1.3 The prior distributions

Table 4.1 lists the priors that will be derived and tested in sections 4.3 (the flat prior), 4.4 (the Jeffreys prior), 4.5 (the reference prior) and 4.6 (the PC prior). Note that the flat prior and the PC prior are proper priors, and have hence been normalised. Figure 4.1 shows the five priors that will be derived and tested in this thesis, including three versions of the PC prior, having value of λ equal to 0.0818, 1.249 and 5.360. The reasoning being using these seemingly quite arbitrary values will be provided in section 4.6.3.

Table 4.1: Summary of the priors that will be derived in this section. The proper priors have been normalised, as indicated by the inclusion of an equality sign, whereas the improper priors (for obvious reasons) have not.

Symbol	Name	Formula
π_{flat}	Flat prior	$= 1/2$
π_J	Jeffreys prior	$\propto \sqrt{1+\rho^2}/(1-\rho^2)$
π_R	Reference prior	$\propto \pi_J$
π_{PC}	PC prior	$= \frac{\lambda\rho\cdot\text{sgn}(\rho)}{2(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} e^{-\lambda\sqrt{-\ln(1-\rho^2)}}$

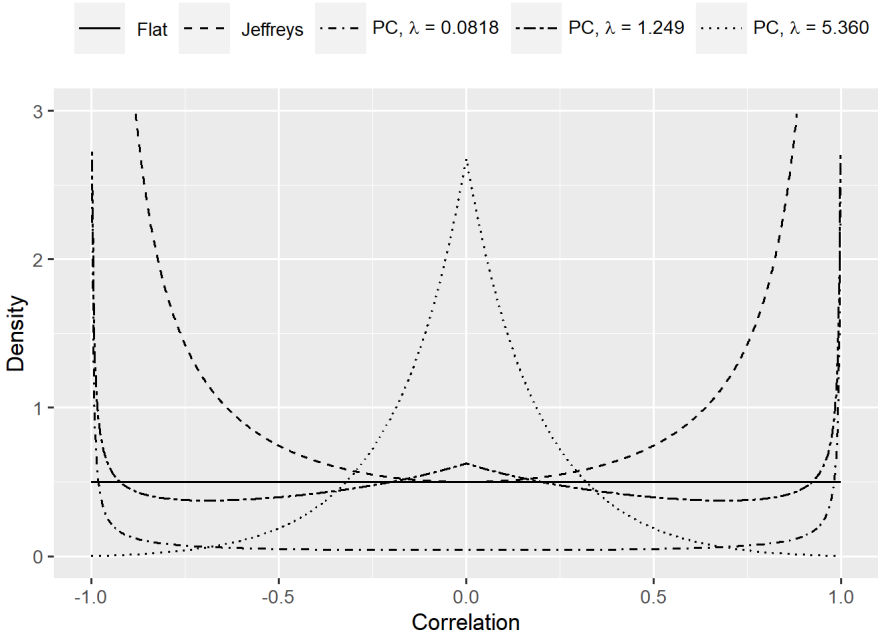


Figure 4.1: The densities of the prior distributions from table 4.1 is shown. The PC prior is shown for values of λ equal to 0.0818, 1.249 and 5.360. Note that the Jeffreys prior has been scaled by a factor 1/2.

In order to get an indication of how the priors behave, we also include a plot of the posteriors using one common sample of size 3 for all five priors, and with $\rho = 0.5$. The results are shown in figure 4.2. We see that the PC prior with $\lambda = 5.360$ gives a posterior that behaves quite differently from the other. More detailed analysis of the posteriors and estimators derived therefrom will be given later in this section, and in section 5.

4.1.4 Bayes estimators

Different variants of the Bayes estimator will be used together with the prior distributions listed in table 4.1 and displayed in figure 4.1 to estimate the value of the correlation

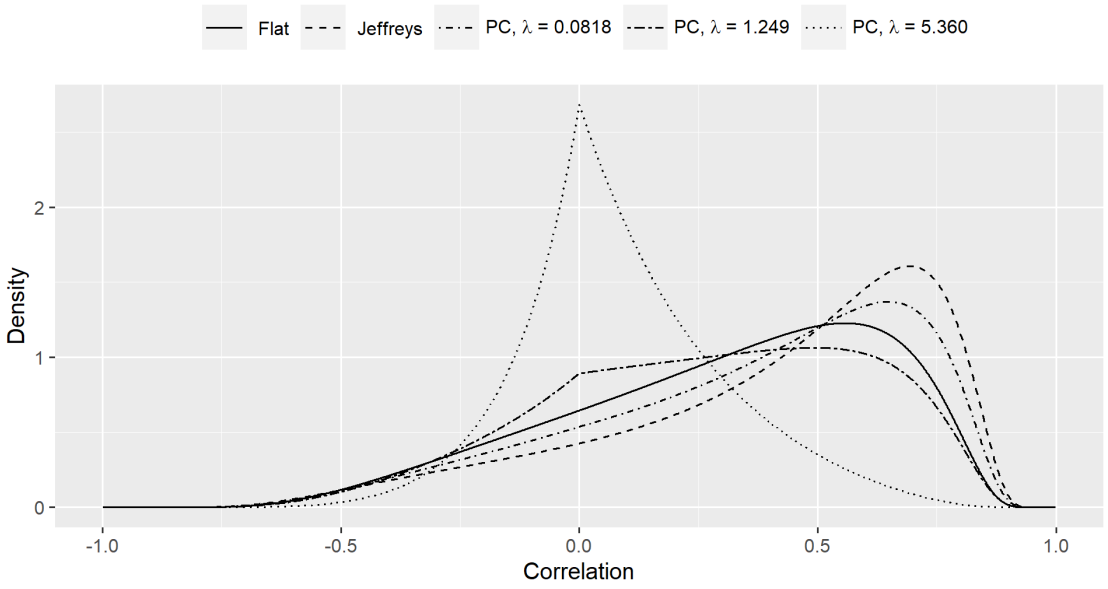


Figure 4.2: The densities of the posterior distributions with corresponding priors given in table 4.1 is shown. The posterior has been calculated using one random sample of size 3 from the bivariate normal distribution using $\rho = 0.5$.

coefficient. The concept of a Bayes estimator was introduced in section 2.2.2, and we will here make use of three variants in order to estimate the correlation coefficient.

- Using mean square error as loss function, the estimator becomes

$$\hat{\rho}_{\text{MSE}} = \int_{-1}^{1} \rho \, \pi(\rho \,|\, x) \; d\rho, \tag{4.3}$$

  which is simply the expected value of the posterior distribution.

- Using Kullback–Leibler divergence as loss function, the estimator is given by equation (4.42).

- Using the Fisher information metric as loss function, the estimator is given by equation (4.47).

4.1.5 Other approaches to parameter estimation

In order to get a better understanding of how the Bayes estimators perform, we will test a fiducial approach, in which we approximate the fiducial (see section 4.8) and use the resulting distribution together with the Bayes estimators – that is, with the fiducial taking the place of the posterior.

4.2 Frequentist approach

In order to thoroughly compare and evaluate the prior distributions derived in this chapter, a look at the frequentist approach to parameter inference is appropriate.

4.2.1 Sufficient statistic

In order to derive a sufficient statistic for ρ, we must find $T(\boldsymbol{x})$ such that $p(\boldsymbol{x} \,|\, \rho) = h(\boldsymbol{x})g(T(\boldsymbol{x}) \,|\, \rho)$ with h and g non-negative functions (see theorem 6.2.6 in Casella and R. L. Berger, 2002, p. 276). Looking at the likelihood, we get

$$\begin{aligned} L(\rho) &= \prod_{i=1}^{n} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_{i1}^2 + x_{i2}^2 - 2\rho x_{i1}x_{i2}}{2(1-\rho^2)}\right) \\ &= \frac{1}{(2\pi)^n (1-\rho^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^{n}\left(x_{i1}^2 + x_{i2}^2\right)}{2(1-\rho^2)} + \frac{\rho\sum_{i=1}^{n} x_{i1}x_{i2}}{(1-\rho^2)}\right), \end{aligned} \tag{4.4}$$

and, setting $h(\boldsymbol{x}) = 1$, we have a two-dimensional sufficient statistic given by

$$T(\boldsymbol{x}) = \big(T_1(\boldsymbol{x}), T_2(\boldsymbol{x})\big) = \left(\sum_{i=1}^{n}\left(x_{i1}^2 + x_{i2}^2\right),\ \sum_{i=1}^{n} x_{i1}x_{i2}\right). \tag{4.5}$$

In the following, we will usually refer to the components of the statistic given by equation (4.5) leaving $\boldsymbol{x}$ implicit, as long as the data in question is clear from the context. That is, we will simply write $T_i = T_i(\boldsymbol{x})$ for $i \in \{1, 2\}$.

Theorem 4.2.1. *The sufficient statistic for ρ given by equation (4.5) is minimal.*

Proof. In order to show that the sufficient statistic $T(\boldsymbol{x})$ is minimal, we need to prove that for random samples $\boldsymbol{x}$ and $\boldsymbol{y}$, we have that $T(\boldsymbol{x}) = T(\boldsymbol{y})$ if and only if $p(\boldsymbol{x} \,|\, \rho)$ / $p(\boldsymbol{y} \,|\, \rho)$ is independent of ρ (see theorem 6.2.13 in Casella and R. L. Berger, 2002, p. 281). If $T(\boldsymbol{x}) = T(\boldsymbol{y})$, we immediately see that the fraction is independent of ρ, by noting that

$$\begin{aligned} \frac{p(\boldsymbol{x} \,|\, \rho)}{p(\boldsymbol{y} \,|\, \rho)} = \exp\Bigg(&\frac{\sum_{i=1}^{n}\left(y_{i1}^2 + y_{i2}^2\right) - \sum_{j=1}^{n}\left(x_{j1}^2 + x_{j2}^2\right)}{2(1-\rho^2)} + \\ &\frac{\rho\left(\sum_{i=1}^{n} x_{i1}x_{i2} - \sum_{j=1}^{n} y_{j1}y_{j2}\right)}{(1-\rho^2)}\Bigg) \end{aligned} \tag{4.6}$$

becomes equal to 1. If we assume that the fraction is independent of ρ, we need both

$\sum_{i=1}^{n}\left(y_{i1}^2+y_{i2}^2\right)-\sum_{j=1}^{n}\left(x_{j1}^2+x_{j2}^2\right)=0$ and $\sum_{i=1}^{n}x_{i1}x_{i2}-\sum_{j=1}^{n}y_{j1}y_{j2}=0$, which implies $T(\boldsymbol{x})=T(\boldsymbol{y})$, and the statistic is minimal sufficient. □

Note that, since we have

$$\begin{aligned} 0 \leq \sum_{i=1}^{n}(x_{i1}-x_{i2})^2 &= \sum_{i=1}^{n}\left(x_{i1}^2-2x_{i1}x_{i2}+x_{i2}^2\right) \\ \implies \sum_{i=1}^{n}\left(x_{i1}^2+x_{i2}^2\right) &\geq 2\sum_{i=1}^{n}x_{i1}x_{i2}, \end{aligned} \tag{4.7}$$

the equality $T_1 \geq 2T_2$ always holds.

4.2.2 Empirical correlation

The empirical correlation coefficient between samples of size n from two random variables Y and Z, denoted $\boldsymbol{y}=(y_1,...,y_n)$ and $\boldsymbol{z}=(z_1,...,z_n)$, is in general given by

$$\hat{\rho}_1=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(z_i-\bar{z})}{\sqrt{\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)\left(\sum_{i=1}^{n}(z_i-\bar{z})^2\right)}}, \tag{4.8}$$

with $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ and $\bar{z}=\frac{1}{n}\sum_{i=1}^{n}z_i$ the empirical means.

In our example with the bivariate normal distribution, the means and variances are known, and hence we can find alternative versions based on this extra knowledge. Firstly, we assume only the means to be known, for which we get

$$\hat{\rho}_2=\frac{\sum_{i=1}^{n}y_iz_i}{\sqrt{\left(\sum_{i=1}^{n}y_i^2\right)\left(\sum_{i=1}^{n}z_i^2\right)}}. \tag{4.9}$$

The denominator of the empirical correlation coefficient is the product of the empirical standard deviations for the two random variables multiplied by n. Hence, we can replace the denominator of equation (4.9) by n times the known standard deviations (which are just equal to 1), and we get

$$\hat{\rho}_3=\frac{1}{n}\sum_{i=1}^{n}y_iz_i. \tag{4.10}$$

Note that this last version will not give values restricted to the interval $[-1,1]$, and hence a truncated version is sometimes used (see e.g. Fosdick and Raftery, 2012), in which values above 1 is set to 1, and values below -1 is set to -1. For a sample $\boldsymbol{x}$ from

the bivariate normal distribution, equation (4.10) will be equal to T_2/n, with T_2 being the second component of the sufficient statistic given in equation (4.5). Note that neither equation (4.8) nor equation (4.9) can be written only in terms of the sufficient statistic.

Some simple tests shows that $\hat{\rho}_1$ performs very poorly for small sample sizes, and that using $\hat{\rho}_2$ leads to markedly better results. We also find that while $\hat{\rho}_3$ does perform better than $\hat{\rho}_1$, a lot of estimates fall outside of the interval $(-1, 1)$. Hence, when using the truncated version of the estimator, the density of $\hat{\rho}_3$ places a lot of weight on the extreme cases -1 and 1, regardless of the true value of ρ. From these considerations, and in order to limit the number of estimators used in the simulation studies of chapter 5, we will only use the estimator $\hat{\rho}_2$.

4.2.3 Maximum likelihood estimator

The likelihood, given by equation (4.4), can be used to derive the maximum likelihood estimator (MLE) for ρ. Taking logarithms, and using the statistic given by equation (4.5) to simplify the expression, we get the log-likelihood

$$\ell(\rho) = -n\log(2\pi) - \frac{n}{2}\log\left(1-\rho^2\right) - \frac{T_1}{2(1-\rho^2)} + \frac{\rho T_2}{1-\rho^2}. \tag{4.11}$$

Further, differentiating with respect to ρ gives

$$\frac{d}{d\rho}\ell(\rho) = \frac{n\rho}{1-\rho^2} - \frac{\rho T_1}{(1-\rho^2)^2} + \frac{T_2}{1-\rho^2} + \frac{2\rho^2 T_2}{(1-\rho^2)^2}. \tag{4.12}$$

Finally, setting the resulting expression equal to zero yields the qubic equation

$$\begin{aligned} &\rho(1-\rho^2) + (1+\rho^2)\frac{T_2}{n} - \rho\frac{T_1}{n} = 0 \\ \implies\ &\rho^3 - \frac{T_2}{n}\rho^2 + \left(\frac{T_1}{n} - 1\right)\rho - \frac{T_2}{n} = 0. \end{aligned} \tag{4.13}$$

The explicit form for the solutions to equation (4.13) can be found in both Fosdick and Raftery, 2012 and MathWorld, n.d.(a). The formulas are quite involved, but may nontheless be used to locate the MLE. It can be shown (see references in Fosdick and Raftery, 2012) that at least one solution to this cubic equation falls in the interval $(-1, 1)$, guaranteeing that we always have an MLE. However, there might be three real solutions in the aforementioned interval, two of which might be local maxima. In this case, the largest of these gives us the MLE.

4.3 Flat prior

A flat prior for the correlation coefficient is simply given by

$$\pi_{\text{flat}} = 1/2. \tag{4.14}$$

This is a proper prior, and hence the posterior will also be proper. For a parameter space like the one we are dealing with, a flat prior does intuitively seem like a good choice; giving equal weight to all possible parameter values, without getting a prior with infinite measure, seems reasonable. Whether or not this prior does work well will be investigated in chapter 5.

4.3.1 Corresponding posterior

The posterior distribution when using a flat prior is given by

$$\begin{aligned} &\pi_{\text{flat}}(\rho \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \rho)\pi_{\text{flat}}(\rho) \\ &= \frac{1}{2^{n+1}\pi^n(1-\rho^2)^{n/2}} \exp\left(-\frac{T_1}{2(1-\rho^2)} + \frac{\rho T_2}{(1-\rho^2)}\right). \end{aligned} \tag{4.15}$$

By plotting the posterior distribution, we can see how it behaves for different sample sizes. In order to get an indication both of how it performs for small samples, and of how increasing the sample size affects the posterior, we use samples of size 3, 10 and 100. (Note that this will also be the preferred sample sizes in chapter 5.) The posterior is calculated for 1000 different samples with $\rho = 0.5$, and the average of these are displayed in figure 4.3.

For a sample size of 3, the posterior density still retains some of the features from the prior, in that it puts a relatively high weight on values close to -1 and 1. Increasing the sample size to 10 and 100 gradually focuses the densities around the true value of ρ, as we would expect. Indeed, for a sample size of 100 the density seems to be close to that of a univariate normal distribution with mean close to 0.5.

4.4 Jeffreys prior

In this section we will first find the Fisher information and use it to derive the Jeffreys prior, and then look at the properties of the prior and corresponding posterior densities.

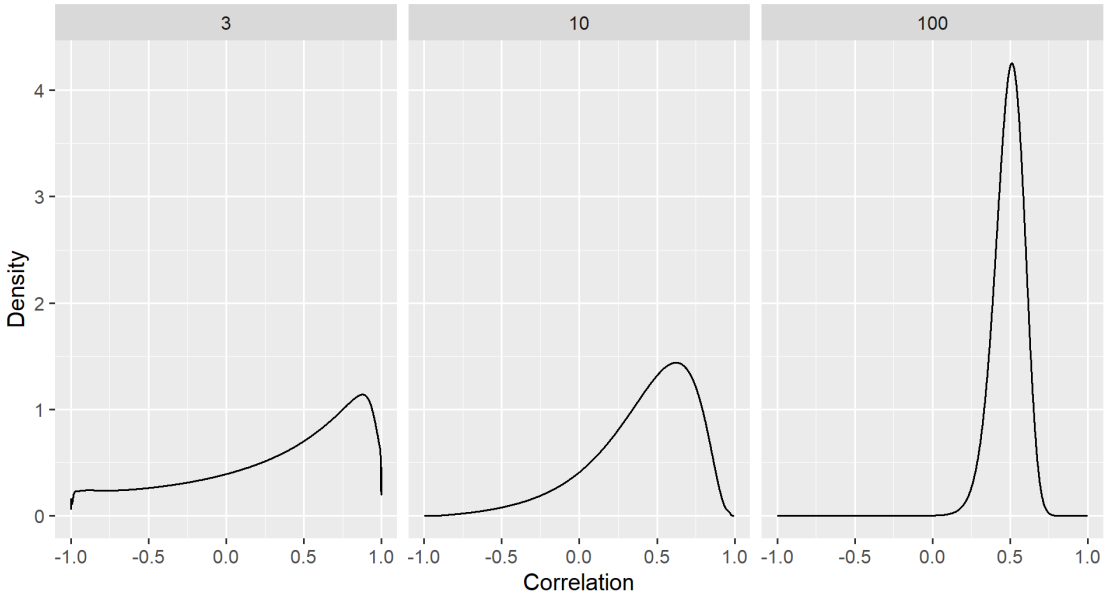


Figure 4.3: Using the **flat prior**, an average over 1000 posterior densities using equation (4.15) is shown. The densities have been calculated for sample sizes 3, 10 and 100, and with $\rho = 0.5$. Note that the number above each plot indicates the sample size for that averaged posterior.

4.4.1 Finding the Fisher information

In order to derive the Jeffreys prior, we first need to find the Fisher information. The square of the score becomes

$$\begin{aligned}
&\left(\frac{d}{d\rho}\log\left(p(x\,|\,\rho)\right)\right)^2 \\
&= \left(\frac{d}{d\rho}\left(-\log(2\pi) - \frac{1}{2}\log\left(1-\rho^2\right) - \frac{x_1^2+x_2^2}{2(1-\rho^2)} + \frac{\rho x_1 x_2}{(1-\rho^2)}\right)\right)^2 \\
&= \left(\frac{\rho}{1-\rho^2} - \frac{(x_1^2+x_2^2)\rho}{(1-\rho^2)^2} + \frac{x_1x_2(1-\rho^2)+2\rho^2x_1x_2}{(1-\rho^2)^2}\right)^2 \\
&= \left(\frac{\rho-\rho^3-\rho x_1^2-\rho x_2^2+\rho^2x_1x_2+x_1x_2}{(1-\rho^2)^2}\right)^2 \\
&= \frac{1}{(1-\rho^2)^4}\Bigg(\rho^2-2\rho^4+\rho^6+(-2\rho^2+2\rho^4)x_1^2+(-2\rho^2+2\rho^4)x_2^2 \\
&\quad + (2\rho-2\rho^5)x_1x_2+\rho^2x_1^4+\rho^2x_2^4+(1+4\rho^2+\rho^4)x_1^2x_2^2 \\
&\quad + (-2\rho-2\rho^3)x_1^3x_2+(-2\rho-2\rho^3)x_1x_2^3\Bigg).
\end{aligned} \tag{4.16}$$

We proceed by taking the expectation of the expression derived in equation (4.16), as-

suming ρ to be known. Due to the linearity of integration and the fact that we integrate over x_1 and x_2 while keeping ρ constant, we need solutions to the integrals

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1^i x_2^j \exp\left(-\frac{x_1^2+x_2^2-2\rho x_1 x_2}{2(1-\rho^2)}\right)\, dx_1 dx_2 \tag{4.17}$$

for $i,j \in \{0,1,2,3\}$ satisfying either $i+j=2$ or $i+j=4$. Calculating these integrals is admittedly quite tedious. In order to provide an indication of the necessary steps, we solve the integral for the combination $(i,j)=(4,0)$, which gives

$$\begin{aligned}
&\int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2(1-\rho^2)}} \int_{-\infty}^{\infty} x_1^4 e^{-\frac{x_1^2}{2(1-\rho^2)}+\frac{\rho x_2}{1-\rho^2}x_1}\, dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2(1-\rho^2)}} e^{\frac{\rho^2 x_2^2}{2(1-\rho^2)}} \int_{-\infty}^{\infty} x_1^4 e^{-\frac{x_1^2-\rho x_2 x_1+\rho^2 x_2^2}{2(1-\rho^2)}}\, dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2}} \int_{-\infty}^{\infty} x_1^4 e^{-\frac{(x_1-\rho x_2)^2}{2(1-\rho^2)}}\, dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2}} \int_{-\infty}^{\infty} (1-\rho^2)^{1/2}\left((1-\rho^2)^{1/2}u+\rho x_2\right)^4 e^{-\frac{u^2}{2}}\, du\, dx_2 \\
&= \int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2}} \int_{-\infty}^{\infty} \Big((1-\rho^2)^{5/2}u^4 + 4(1-\rho^2)^2\rho x_2 u^3 \\
&\qquad\qquad + 6(1-\rho^2)^{3/2}\rho^2 x_2^2 u^2 + 4(1-\rho^2)\rho^3 x_2^3 u \\
&\qquad\qquad + (1-\rho^2)^{1/2}\rho^4 x_2^4\Big) e^{-\frac{u^2}{2}}\, du\, dx_2 \\
&= \int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2}} \Big(3\sqrt{2\pi}(1-\rho^2)^{5/2} + 6\sqrt{2\pi}\rho^2(1-\rho^2)^{3/2}x_2^2 \\
&\qquad\qquad + \sqrt{2\pi}\rho^4(1-\rho^2)^{1/2}x_2^4\Big)\, dx_2 \\
&= \left(6\pi(1-\rho^2)^{5/2} + 12\pi\rho^2(1-\rho^2)^{3/2} + 6\pi\rho^4(1-\rho^2)^{1/2}\right)
\end{aligned} \tag{4.18}$$

where we used the substitution $u = (x_1-\rho x_2)/\sqrt{1-\rho^2}$, as well as the results that $\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}$ (see Rottmann, 2014/2003, p. 155, eq. 49), and

$$\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\, dx = \left[-xe^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx = \sqrt{2\pi},$$

and

$$\int_{-\infty}^{\infty} x^4 e^{-\frac{x^2}{2}}\, dx = \left[-x^3 e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 3x^2 e^{-\frac{x^2}{2}}\, dx$$
$$= \left[-3x e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 3e^{-\frac{x^2}{2}}\, dx = 3\sqrt{2\pi}.$$

The integrals $\int_{-\infty}^{\infty} x e^{-\frac{x^2}{2}}\, dx$ and $\int_{-\infty}^{\infty} x^3 e^{-\frac{x^2}{2}}\, dx$ are both equal to zero, since the integrands are odd functions.

The remaining integrals are solved in a similar fashion. A summary of the solutions is provided in table 4.2. Note that, due to the symmetry of the bivariate normal distribution (assuming, as we have, equal means and variances), we need only consider $(i, j) = (0, 2)$ or $(i, j) = (2, 0)$, not both.

Table 4.2: Solutions to the integral $\int_{-\infty}^{\infty} x_1^i x_2^j \exp\left(-\frac{x_1^2+x_2^2-2\rho x_1 x_2}{2(1-\rho^2)}\right) dx_1 dx_2$ for different values of i and j.

i	j	Solutions to equation (4.17)
0	2	$2\pi(1-\rho^2)^{3/2} + 2\pi\rho^2(1-\rho^2)^{1/2}$
1	1	$2\pi\rho(1-\rho^2)^{1/2}$
0	4	$6\pi(1-\rho^2)^{5/2} + 12\pi\rho^2(1-\rho^2)^{3/2} + 6\pi\rho^4(1-\rho^2)^{1/2}$
1	3	$6\pi\rho(1-\rho^2)^{3/2} + 6\pi\rho^3(1-\rho^2)^{1/2}$
2	2	$2\pi(1-\rho^2)^{3/2} + 6\pi\rho^2(1-\rho^2)^{1/2}$

Using the results from table 4.2 together with equation (4.16), cancelling equal terms,

we get

$$
\begin{aligned}
I(\rho) &= \mathbb{E}\left[\left(\frac{d}{d\rho}\log\left(p(x\,|\,\rho)\right)\right)^2 \,\middle|\, \rho\right] \\
&= \frac{\rho^2 - 2\rho^4 + \rho^6}{(1-\rho^2)^4} + \frac{4\rho^4 - 4\rho^2}{(1-\rho^2)^3} + \frac{4\rho^6 - 4\rho^4}{(1-\rho^2)^4} + \frac{2\rho^2 - 2\rho^6}{(1-\rho^2)^4} \\
&\quad + \frac{6\rho^2}{(1-\rho^2)^2} + \frac{12\rho^4}{(1-\rho^2)^3} + \frac{6\rho^6}{(1-\rho^2)^4} - \frac{12\rho^2 + 12\rho^4}{(1-\rho^2)^3} \\
&\quad - \frac{12\rho^4 + 12\rho^6}{(1-\rho^2)^4} + \frac{1 + 4\rho^2 + \rho^4}{(1-\rho^2)^3} + \frac{3\rho^2 + 12\rho^4 + 3\rho^6}{(1-\rho^2)^4} \\
&= \frac{6\rho^2}{(1-\rho^2)^2} + \frac{1 - 12\rho^2 + 5\rho^4}{(1-\rho^2)^3} + \frac{6\rho^2 - 6\rho^4}{(1-\rho^2)^4} \\
&= \frac{6\rho^2 - 6\rho^4 + 1 - 12\rho^2 + 5\rho^4 + 6\rho^2}{(1-\rho^2)^3} \\
&= \frac{1-\rho^4}{(1-\rho^2)^3} \\
&= \frac{1+\rho^2}{(1-\rho^2)^2}.
\end{aligned}
\tag{4.19}
$$

4.4.2 Deriving the prior

The Fisher information given by equation (4.19), in turn, implies that the Jeffreys prior becomes

$$
\pi_J(\rho) \propto \frac{\sqrt{1+\rho^2}}{1-\rho^2}, \tag{4.20}
$$

which is the same result as can be found in Fosdick and Raftery, 2012. The integral of equation (4.20) with respect to ρ does not converge, from which we conclude that the Jeffreys prior is improper. To see this, we can write

$$
\begin{aligned}
\int_{-1}^{1} \frac{\sqrt{1+\rho^2}}{1-\rho^2}\, d\rho &= 2\int_{0}^{1} \frac{\sqrt{1+\rho^2}}{1-\rho^2}\, d\rho \\
&\geq 2\int_{0}^{1} \frac{1}{1-\rho^2}\, d\rho = 2\int_{0}^{1} \frac{1}{(1+\rho)(1-\rho)}\, d\rho \\
&\geq \int_{0}^{1} \frac{1}{1-\rho}\, d\rho = \int_{0}^{1} \frac{1}{x}\, dx = \Big[\ln(x)\Big]_0^1 = \infty,
\end{aligned}
\tag{4.21}
$$

which proves the above statement.

Figure 4.4 shows a plot of the Jeffreys prior with an arbitrary scaling. It can be

readily seen that the prior gives a lot of weight to extreme cases – that is, a correlation close to -1 or 1 – and much less weight to correlations closer to 0.

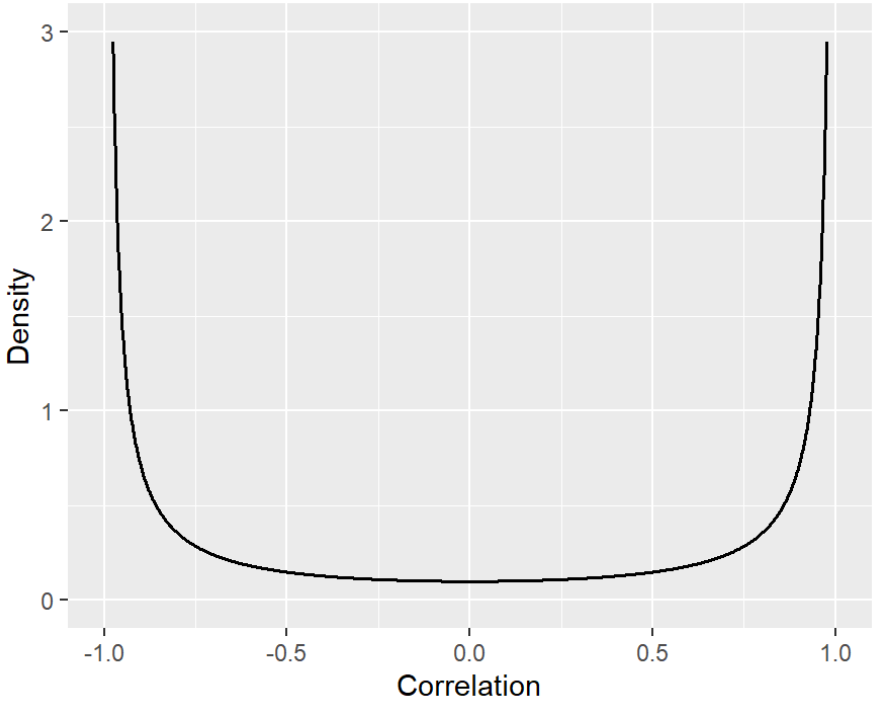


Figure 4.4: The density of the **Jeffreys prior** is shown. Note that it has been scaled more or less arbitrarily, but in a way that makes it easier to see the important features of the prior.

4.4.3 Corresponding posterior

Letting $\boldsymbol{x} = (x_1, ..., x_n)$ be the data, and using the statistic of equation (4.5) to simplify the expression, the posterior distribution using the Jeffreys prior in equation (4.20) becomes

$$\begin{aligned} &\pi_J(\rho \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \rho)\pi_J(\rho) \\ &= \frac{\sqrt{1+\rho^2}}{(2\pi)^n(1-\rho^2)^{n/2+1}} \exp\left(-\frac{T_1}{2(1-\rho^2)} + \frac{\rho T_2}{(1-\rho^2)}\right). \end{aligned} \tag{4.22}$$

Similarly to what we did for the flat posterior, we plot the posterior distribution in order to see how they behave for the sample sizes 3, 10 and 100. The posteriors are again calculated for 1000 different samples with $\rho = 0.5$, and the averages of these are displayed in figure 4.5.

We see that the behaviour of the posterior for the three chosen sample sizes is quite

similar to that of the flat posterior. However, for sample sizes 3 and 10, the density is shifted slightly towards the extremes (that is, $\rho \to \pm 1$). This seems reasonable, since the Jeffreys prior diverges in this situation (and hence a lot of weight is placed near ± 1), whereas the flat prior does not.

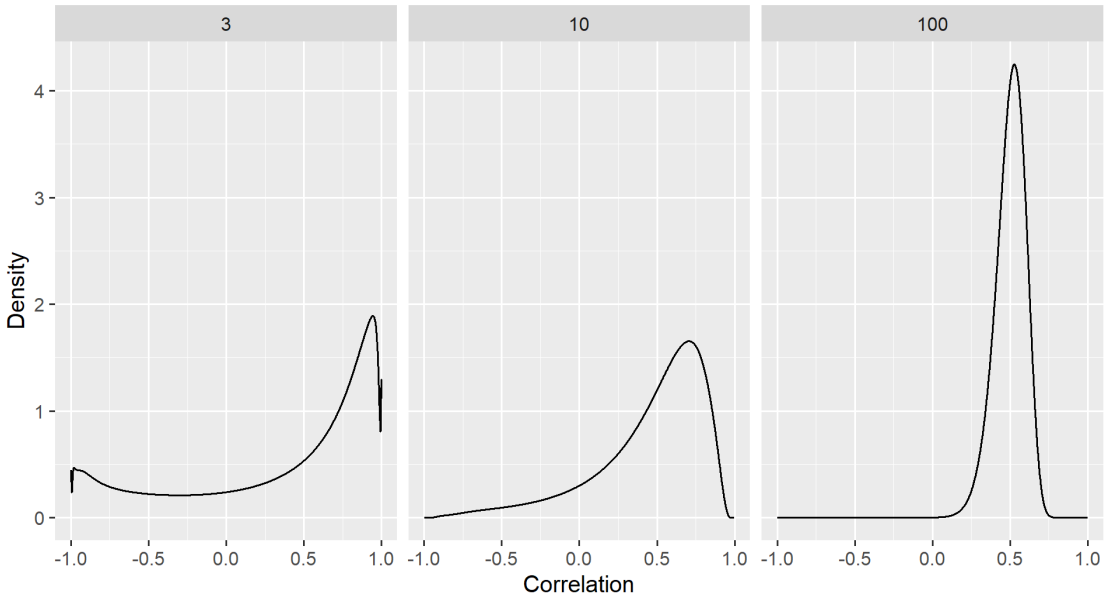


Figure 4.5: Using the **Jeffreys prior**, an average over 1000 posterior densities using equation (4.22) is shown. The densities have been calculated for sample sizes 3, 10 and 100, and with $\rho = 0.5$. Note that the number above each plot indicates the sample size for that averaged posterior.

4.4.4 The posterior is proper

In order to be sure that the posterior density in equation (4.22) can be applied safely, we should prove that it is a proper probability distribution.

Theorem 4.4.1. *The integral of the posterior density in equation* (4.22) *over* ρ *is finite, and hence the posterior distribution corresponding to the Jeffreys prior in equation* (4.20) *is proper.*

Proof. The theorem will be proved by finding a finite upper bound on the integral of the posterior. We ignore the constant $1/(2\pi)^n$ throughout this section. We also note that, since the integrand is always positive, replacing $\sqrt{1+\rho^2}$ by $\sqrt{2}$ in the numerator will only make the value of the integral larger, and since this is a finite constant we simply

ignore it. We therefore want to show that

$$\int_{-1}^{1} \frac{1}{(1-\rho^2)^{n/2+1}} \exp\left(-\frac{T_1 - 2\rho T_2}{2(1-\rho^2)}\right) \, d\rho < \infty. \tag{4.23}$$

We will make the assumption that $T_2 > 0$ and prove that the integral in equation (4.23) is finite. The case where $T_2 < 0$ can be proved in a similar fashion. The case $T_2 = 0$ then follows easily by reusing arguments from the other cases. We will also utilise the fact that $T_1 \geq 2T_2$, shown in section 4.2.1.

By integrating from -1 to 0 and from 0 to 1 separately, we can make use of different arguments to show that they are both finite. For $\rho \geq 0$ we have that $T_1 - 2\rho T_2 \geq T_1 - 2T_2 \geq 0$, giving $\exp(-(T_1 - 2\rho T_2)) \leq \exp(-(T_1 - 2T_2))$. For $\rho \leq 0$ we can simply use that $\exp(-(T_1 - 2\rho T_2)) \leq \exp(-T_1)$. Hence, we get that the expression in (4.23) is smaller than

$$\begin{aligned} &\leq \int_{0}^{1} \frac{1}{(1-\rho^2)^{n/2+1}} \exp\left(-\frac{T_1 - 2T_2}{2(1-\rho^2)}\right) \, d\rho \\ &\quad + \int_{-1}^{0} \frac{1}{(1-\rho^2)^{n/2+1}} \exp\left(-\frac{T_1}{2(1-\rho^2)}\right) \, d\rho, \end{aligned} \tag{4.24}$$

Using the substitution $u = 1/(1-\rho^2)$, we get

$$\begin{aligned} &\leq \frac{1}{2} \int_{1}^{\infty} u^{n/2-1} \exp\left(-\frac{T_1 - 2T_2}{2} u\right) \, du \\ &\quad + \frac{1}{2} \int_{1}^{\infty} u^{n/2-1} \exp\left(-\frac{T_1}{2} u\right) \, du, \end{aligned} \tag{4.25}$$

where we have replaced $1/\rho$ by 1 in the first integral and -1 in the second, since this will give us something larger or equal. By substituting $v = (T_1 - 2T_2)u/2$ in the first integral and $v = T_1 u/2$ in the second, we simply get

$$\begin{aligned} &= \frac{1}{2} \left(\frac{2}{T_1 - 2T_2}\right)^{-n/2} \Gamma\left(\frac{n}{2}, \frac{T_1 - 2T_2}{2}\right) \\ &\quad + \frac{1}{2} \left(\frac{2}{T_1}\right)^{-n/2} \Gamma\left(\frac{n}{2}, \frac{T_1}{2}\right), \end{aligned} \tag{4.26}$$

with $\Gamma(\cdot, \cdot)$ the upper incomplete Gamma function (see MathWorld, n.d.(b)). This result is finite, and hence our posterior density is proper. □

4.5 Reference prior

In order to derive the exact reference prior, we should apply theorem 3.4.4. In lack of any obvious statistic to use in the calculations, we try using the (obviously sufficient) statistic given by the whole sample $\boldsymbol{x} = (x_1, ..., x_n)$. Since the data is assumed to be independent and identically distributed, the density $p(t_n \mid \theta)$ in equation (3.11) is simply given by the likelihood function of equation (4.4). By letting $\pi^*(\theta) = 1$ in equation (3.12), we can derive $\pi^*(\theta \mid t_n)$ by calculating

$$\int_{-1}^{1} L(\rho)\, d\rho = \frac{1}{(2\pi)^n} \int_{-1}^{1} \frac{1}{(1-\rho^2)^{n/2}} \cdot \exp\left(-\frac{T_1}{2(1-\rho^2)} + \frac{\rho T_2}{1-\rho^2}\right) d\rho. \tag{4.27}$$

We have not, however, been able to find an analytic solution to this integral, and hence we are forced to rely on other methods of derivation.

4.5.1 Jeffreys prior equals reference prior

In some situations the one-parameter reference prior is simply given by the Jeffreys prior. Indeed, Wikipedia, n.d. provides a proof indicating that this is always the case. However, there seems to be an assumption that there exists an asymptotically normal estimator for the parameter, which allows for the use of Bernstein-von Mises theorem (see e.g. Mathematics, n.d.(b)). Hence, there are conditions that need to be satisfied before we can conclude that the priors are identical. Despite some efforts to show equality, we have not been able to produce anything conclusive. Hence, we move on to a numerical approximation of the reference prior, which gives an indication that the priors are in fact equal. We will therefore assume equality and work only with the Jeffreys prior in what follows.

4.5.2 Numerical approximation

A numerical scheme as presented in algorithm 1 lets us approximate the true reference prior. Figure 4.6 shows the numerical approximation to the density. The number of values of ρ was set to 1000, and they have been placed uniformly on the interval $(0, 1)$. For each value, 10 000 iterations were performed. The sample size for each iteration was set to 100. The Jeffreys prior given by equation (4.20) is displayed together with the

numerically approximated reference prior, showing that the similarity is indeed striking. Note that the Jeffreys prior has been scaled by a factor of 2.45 – chosen to make it align well with the approximated reference prior.

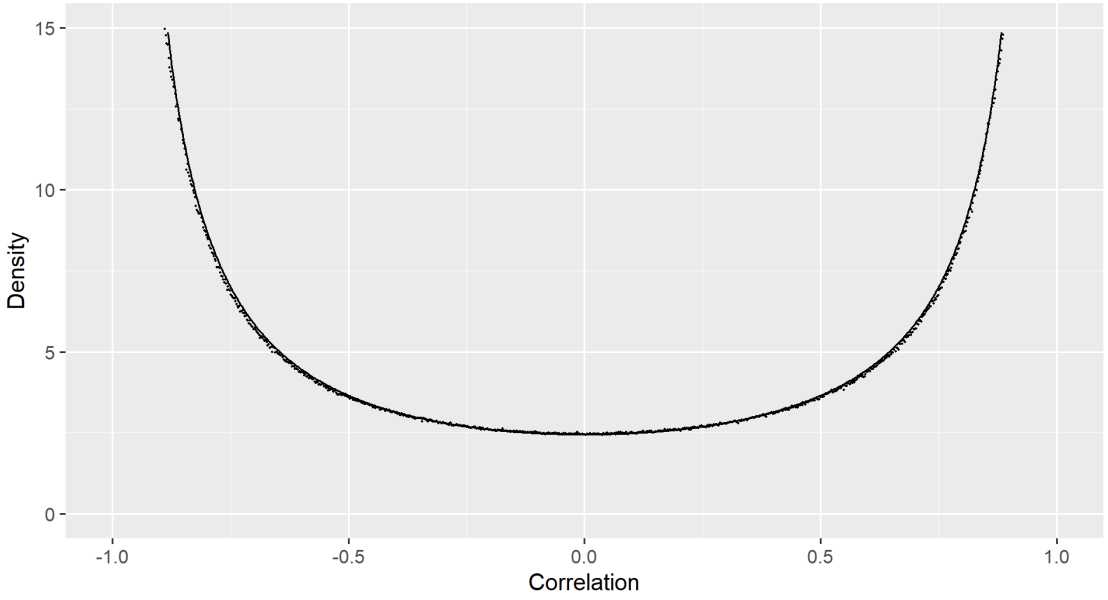


Figure 4.6: The numerical approximation to the density of the reference prior is shown. The prior was calculated for 1000 values of ρ placed uniformly on the interval $(0, 1)$, with 10 000 iterations for each value. The sample size for each iteration was set to 100. Together with the approximated reference prior, the Jeffreys priors given by equation (4.20) – scaled by a factor of 2.45 – is shown as a solid line, showing that they do indeed look very similar.

4.6 Penalised Complexity prior

A natural choice for the base model is to set $\rho = 0$, that is, the two components of the bivariate normal are independent. We will denote the full model – that is, the bivariate normal with $\rho \in (-1, 1)$ unknown – by N_f, and the base model by N_b.

4.6.1 Kullback–Leibler divergence

We will first find the Kullback–Leibler divergence between two bivariate normal distributions with parameters ρ_1 and ρ_2.

$$
\begin{aligned}
&\kappa\left(p(x \mid \rho_2) \mid p(x \mid \rho_1)\right) \\
&= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho_1^2}} e^{-\frac{x_1^2+x_2^2-2\rho_1 x_1 x_2)}{2(1-\rho_1^2)}} \\
&\qquad\qquad \cdot \ln\left(\frac{\frac{1}{2\pi\sqrt{1-\rho_2^2}} e^{-\frac{x_1^2+x_2^2-2\rho_2 x_1 x_2}{2(1-\rho_2^2)}}}{\frac{1}{2\pi\sqrt{1-\rho_1^2}} e^{-\frac{x_1^2+x_2^2-2\rho_1 x_1 x_2}{2(1-\rho_1^2)}}}\right) dx_1 dx_2 \\
&= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho_1^2}} e^{-\frac{x_1^2+x_2^2-2\rho_1 x_1 x_2}{2(1-\rho_1^2)}} \left(\frac{1}{2}\ln\left(\frac{1-\rho_1^2}{1-\rho_2^2}\right)\right. \\
&\qquad\qquad \left. - \frac{x_1^2+x_2^2-2\rho_2 x_1 x_2}{2(1-\rho_2^2)} + \frac{x_1^2+x_2^2-2\rho_1 x_1 x_2}{2(1-\rho_1^2)}\right) dx_1 dx_2 \\
&= -\frac{1}{2}\ln\left(\frac{1-\rho_1^2}{1-\rho_2^2}\right) + \frac{1}{2\pi\sqrt{1-\rho_1^2}} \\
&\quad \cdot \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\left(\frac{1}{2(1-\rho_2^2)} - \frac{1}{2(1-\rho_1^2)}\right)(x_1^2+x_2^2)\right. \\
&\quad \left. - \left(\frac{\rho_2}{(1-\rho_2^2)} - \frac{\rho_1}{(1-\rho_1^2)}\right) x_1 x_2\right) e^{-\frac{x_1^2+x_2^2-2\rho_1 x_1 x_2}{2(1-\rho_1^2)}} \, dx_1 dx_2.
\end{aligned}
\tag{4.28}
$$

Using the results summarised in table 4.2, we can find the solution to the integrals, and get

$$
\begin{aligned}
&\kappa\left(p(x \mid \rho_2) \mid p(x \mid \rho_1)\right) \\
&= -\frac{1}{2}\ln\left(\frac{1-\rho_1^2}{1-\rho_2^2}\right) + \frac{1-\rho_1^2}{(1-\rho_2^2)} - 1 + \frac{\rho_1^2}{(1-\rho_2^2)} - \frac{\rho_1^2}{(1-\rho_1^2)} \\
&\qquad\qquad - \frac{\rho_1\rho_2}{(1-\rho_2^2)} + \frac{\rho_1^2}{(1-\rho_1^2)} \\
&= -\frac{1}{2}\ln\left(\frac{1-\rho_1^2}{1-\rho_2^2}\right) + \frac{1-\rho_1\rho_2}{(1-\rho_2^2)} - 1.
\end{aligned}
\tag{4.29}
$$

Now, the Kullback–Leibler divergence between the model and the base model becomes

$$
\kappa\left(N_b \mid N_f\right) = -\frac{1}{2}\ln(1-\rho^2), \tag{4.30}
$$

by setting $\rho_1 = \rho$ and $\rho_2 = 0$.

4.6.2 Deriving the prior

From equations (3.14) and (4.30), we get

$$d(\rho) = \sqrt{-\ln(1-\rho^2)}. \tag{4.31}$$

Therefore, we have that

$$\left|\frac{d}{d\rho}d(\rho)\right| = \left|\frac{\rho}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}}\right| = \frac{\rho \cdot \text{sgn}(\rho)}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}}, \tag{4.32}$$

which yields the prior

$$\pi_{PC}(\rho) \propto \frac{\lambda\rho \cdot \text{sgn}(\rho)}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} e^{-\lambda\sqrt{-\ln(1-\rho^2)}}. \tag{4.33}$$

with $\lambda > 0$.

Now, the PC prior in equation (4.33) does indeed look rather complicated as compared to a flat prior or the Jeffreys prior of equation (4.20), and it is difficult to imagine someone formulating such a prior without going through the above calculation using the definition of the PC prior. Hence, it would be interesting to see if this added complexity does actually improve the performance of the subsequent analysis.

Before we can start applying the prior in our analysis, however, we should look more closely at its behaviour at the point $\rho = 0$, since at this point we seem to get $0/0$. Our first observation is that the prior is symmetric around $\rho = 0$, due to the sign function counteracting the negative sign of ρ in the numerator. For $\rho = 0$, both the exponential function and the factor $(1-\rho^2)$ in the numerator becomes 1. Using the result that $\ln(1-x) \approx x$ for $|x| \ll 1$, we get $\sqrt{-\ln(1-\rho^2)} \approx \sqrt{-(-\rho^2)} = \rho \cdot \text{sgn}(\rho)$, and hence the square root in the denominator cancels the terms ρ and $\text{sgn}(\rho)$ in the numerator. What we are left with, then, is the conclusion that $\pi_{PC}(0) \propto \lambda$.

Now, if the prior is proper we would like to integrate over the whole parameter space to find the normalising constant. Using the substitution $u = \sqrt{-\ln(1-\rho^2)}$, we get

$$\begin{aligned} &\int_{-1}^{1} \frac{\lambda\rho \cdot \text{sgn}(\rho)}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} e^{-\lambda\sqrt{-\ln(1-\rho^2)}}\, d\rho \\ &= 2\int_0^{\infty} \lambda e^{-\lambda u}\, du = \left[2\lambda e^{-\lambda u}\right]_0^{\infty} = 2, \end{aligned} \tag{4.34}$$

implying that the prior is indeed proper, and that the normalising constant is independent of λ and equal to 2. In conclusion, the PC prior becomes

$$\pi_{PC}(\rho) = \frac{\lambda\rho \cdot \text{sgn}(\rho)}{2(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} e^{-\lambda\sqrt{-\ln(1-\rho^2)}}, \tag{4.35}$$

and we have that $\pi_{PC}(0) = \lambda/2$.

4.6.3 User-defined scaling

Before the prior is fully defined, the choice of user-defined scaling, that is the value of λ, must be made. Looking at equation (3.22), we choose Q equal to the absolute value function, $U = \rho_0 \in (0,1)$, and $\alpha \in (0,1)$. From this, we need to solve the equation $P(|\rho| > \rho_0) = \alpha$ with respect to λ, which becomes

$$\begin{aligned}
\left(\int_{-1}^{-\rho_0} + \int_{\rho_0}^{1}\right) \pi_{PC}(\rho)\, d\rho &= \alpha \\
2\int_{\rho_0}^{1} \pi_{PC}(\rho)\, du &= \alpha \\
\int_{\rho_0}^{1} \lambda e^{-\lambda u}\, du &= \alpha \\
\left[-e^{-\lambda u}\right]_{\rho_0}^{1} &= \alpha \\
e^{-\lambda\sqrt{-\ln(1-\rho_0^2)}} &= \alpha,
\end{aligned} \tag{4.36}$$

using the same substitution as in equation (4.34). Equation (4.36) gives

$$\lambda = \frac{-\ln(\alpha)}{\sqrt{-\ln(1-\rho_0^2)}}. \tag{4.37}$$

Now, how do we actually choose α and ρ_0? Keeping ρ_0 constant, a larger value of α would mean more weight given to extreme values of ρ. At the same time, keeping α constant, a larger value of ρ_0 would have a similar effect, shifting the weight defined by α to a smaller interval of ρ in the vicinity of ± 1. Increasing α or increasing ρ_0 both result in a smaller value of λ, whereas doing the opposite results in larger values of λ, and hence we conclude that setting λ relatively close to 0 would give large weight to extreme values of ρ, whereas setting λ equal to a large positive number would focus the density around $\rho = 0$, which is our base model. In other words, large values of λ would force the resulting estimates of ρ closer to the value of the base model.

It is not immediately evident which values of α and ρ_0, and hence which value of λ, is most suitable. We will therefore choose three combinations aiming to capture the differences in the PC prior in order to gauge the effects and try to deduce which values of λ seem to perform better. The combinations chosen can be seen in 4.3.

Table 4.3: The chosen values of α and ρ_0 for the user-defined scaling of the PC prior. The values of λ have been calculated using equation (4.37).

α	ρ_0	λ (approx.)
0.9	0.9	0.0818
0.2	0.9	1.249
0.001	0.9	5.360

Figure 4.7 shows the PC prior of equation (4.33) for the three λ values displayed in table 4.3. We immediately see that our analysis above is confirmed, as smaller values of λ leads to a flatter prior around $\rho = 0$ (see the solid line), whereas a higher value of λ focuses the prior around $\rho = 0$ and gives little weight to $\rho \to \pm 1$ (see the dotted line). A value of λ close to 1 (see the dash-dotted line) seems to balance these two tendencies, and does in fact cause the prior to closely follow a flat prior for $\rho \in (-0.8, 0.8)$.

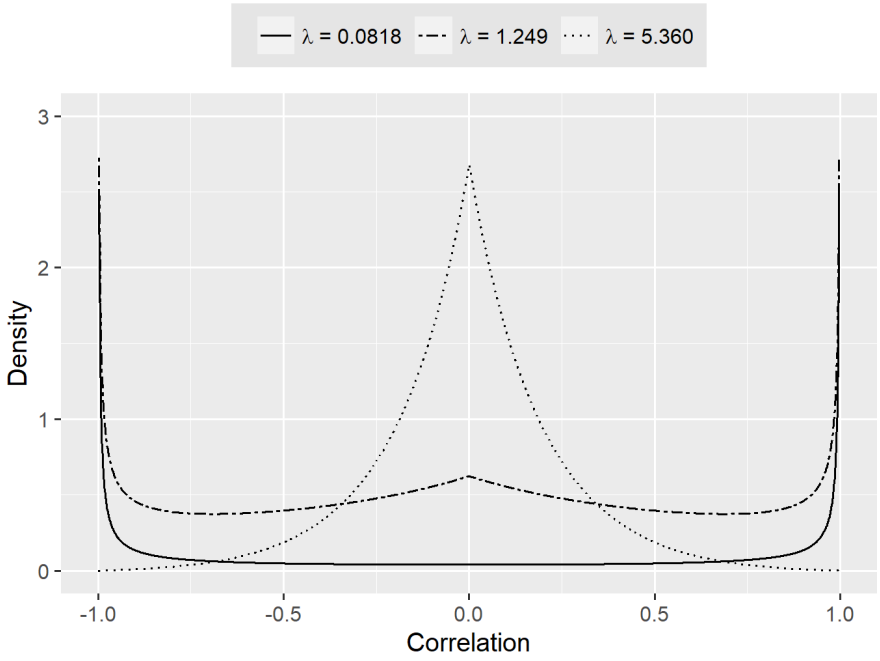


Figure 4.7: The densities of the **PC prior** from equation (4.35) with $\lambda = 0.0818$, 1.249 and 5.360 (as summarised in table 4.3) are shown.

4.6.4 Limiting behaviour of user-defined scaling

Letting λ approach zero, we get that

$$\lim_{\lambda \to 0} \pi_{PC}(\rho) = 0 \tag{4.38}$$

for $\rho \in (0, 1)$. Hence, in the limit, all weight seems to be located at the extreme values $\rho = \pm 1$. Figure 4.8 indicates that this is indeed the case, showing the prior for $\lambda = 0.01$, 0.001 and 0.0001 zoomed in on the interval $[0.999, 1]$. In other words, the density seems to be approaching

$$\frac{1}{2}\delta(\rho + 1) + \frac{1}{2}\delta(\rho - 1), \tag{4.39}$$

with $\delta(\rho - a)$ the Dirac delta function with value infinity at point $\rho = a$ and zero elsewhere.

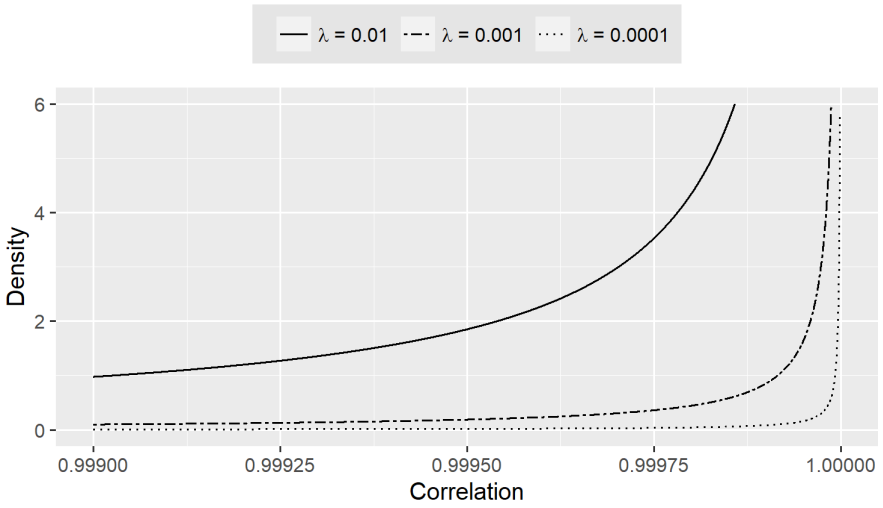


Figure 4.8: The densities of the **PC prior** from equation (4.35) with $\lambda = 0.01$, 0.001 and 0.0001, zoomed in on the interval $[0.999, 1]$, are shown.

If $\lambda \to \infty$, all mass will be located at $\rho = 0$, that is

$$\lim_{\lambda \to \infty} \pi_{PC}(\rho) = \delta(\rho), \tag{4.40}$$

which means that the model will be forced towards the base model regardless of the data.

Any prior distribution that gives zero weight to some region of the parameter space will yield bad results for some parameter values, even when using vast amounts of data, conflicting with our goal of finding a prior that performs well for most or all situations. The limits of the PC prior will force the posterior towards ± 1 or 0 regardless of the true

parameter value, and hence, in the limits $\lambda \to 0$ and $\lambda \to \infty$, the PC prior should not be used.

It should, however, be noted that this by no means disqualifies the PC prior for values of λ between 0 and infinity, for which the prior will be strictly positive for all possible values of ρ.

4.6.5 Corresponding posterior

The posterior distribution, using the statistic given by equation (4.5), becomes

$$\begin{aligned} &\pi_{PC}(\rho \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \rho)\pi_{PC}(\rho) \\ &= \frac{\lambda\rho \cdot \operatorname{sgn}(\rho)}{2^{n+1}\pi^n(1-\rho^2)^{n/2+1}\sqrt{-\ln(1-\rho^2)}} \\ &\qquad \cdot \exp\left(-\frac{T_1}{2(1-\rho^2)} + \frac{\rho T_2}{(1-\rho^2)} - \lambda\sqrt{-\ln(1-\rho^2)}\right), \end{aligned} \tag{4.41}$$

which, since $\pi_{PC}(\rho)$ is proper, is guaranteed to be a proper posterior.

Similarly to what we did for the flat and Jeffreys posteriors, we plot the posterior distributions in order to see how they behave for the sample sizes 3, 10 and 100. The posteriors are again calculated for 1000 different samples with $\rho = 0.5$, and the averages of these are displayed in figure 4.9.

For a sample size of 3, all three posterior densities clearly retain some of the features from their corresponding priors. Both for $\lambda = 0.0818$ and $\lambda = 1.249$, the posterior places much weight on values close to -1 and 1. This effect is much less prominent for a sample size of 10. For $\lambda = 1.249$ and especially $\lambda = 5.360$, the local maximum seen in the priors at $\rho = 0$ continues to be visible in the form of a notch in the density both for samples of size 3 and 10. For a sample size of 100, the posterior densities are very close to those seen for the Jeffreys posterior in figure 4.5 – that is, seemingly close to univariate normal distributions with mean 0.5.

4.7 Bayes estimators

In this section, we will find formulas for the Bayes estimator when using Kullback–Leibler divergence and the Fisher information metric as loss functions, as discussed in section 2.2.2.

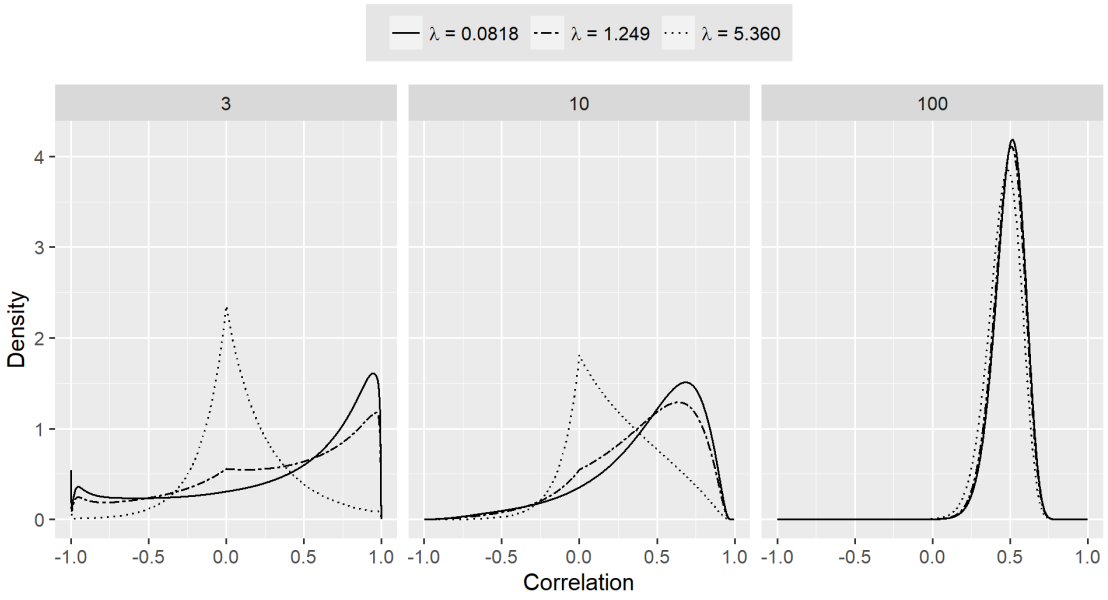


Figure 4.9: Using the **PC prior**, an average over 1000 posterior densities using equation (4.41) is shown. The densities have been calculated for sample sizes 3, 10 and 100, and for $\lambda = 0.0818$, 1.249 and 5.360, with $\rho = 0.5$. Note that the number above each plot indicates the sample size for those averaged posteriors.

4.7.1 Kullback–Leibler divergence

Using the result from equation (4.29) with $\rho_1 = \rho$ and $\rho_2 = \hat{\rho}$, the Bayes estimator becomes

$$\hat{\rho}_{KL} = \underset{\hat{\rho}}{\operatorname{argmin}} \int_{-1}^{1} \left(-\frac{1}{2} \ln \left(\frac{1-\rho^2}{1-\hat{\rho}^2} \right) + \frac{1-\hat{\rho}\rho}{(1-\hat{\rho}^2)} - 1 \right) \pi(\rho \,|\, x) \; d\rho. \tag{4.42}$$

Since we can ignore any part of equation (4.42) that is constant with respect to $\hat{\rho}$, we can simplify the expression to get

$$\hat{\rho}_{KL} = \underset{\hat{\rho}}{\operatorname{argmin}} \left(\frac{1}{2} \ln \left(1-\hat{\rho}^2\right) + \frac{1-\hat{\rho}}{(1-\hat{\rho}^2)} \mathbb{E}_{\pi} \left[\rho \,|\, x\right] \right), \tag{4.43}$$

with $\mathbb{E}_{\pi} \left[\rho \,|\, x\right] = \int_{-1}^{1} \rho \, \pi(\rho \,|\, x) \, d\rho$. Interestingly, it turns out that the Bayes estimator with Kullback–Leibler divergence as loss function is a function of ρ only through the Bayes estimator with MSE as loss function – which is just the expected value of the posterior density (as seen in section 2.2.2).

4.7.2 Fisher information metric

In order to find the Bayes estimator with the Fisher information metric as loss function, we first need to find the integral of the square root of the fisher information. Using the result from equation (4.19), we first note that

$$\begin{aligned} &\frac{\sqrt{1+x^2}}{1-x^2} = \frac{1+x^2}{(1-x^2)\sqrt{1+x^2}} \\ &= \frac{2}{(1-x^2)\sqrt{1+x^2}} - \frac{1-x^2}{(1-x^2)\sqrt{1+x^2}} = \frac{2}{(1-x^2)\sqrt{1+x^2}} - \frac{1}{\sqrt{1+x^2}}. \end{aligned} \tag{4.44}$$

The integral of the second term in the last expression of equation (4.44) is equal to $\operatorname{arcsinh}(x)$ plus a constant. Using the, perhaps not immediately obvious, substitution $u = \sqrt{2}x/\sqrt{1+x^2}$ for the first term, we get

$$\begin{aligned} &\int \frac{2}{(1-x^2)\sqrt{1+x^2}}\, dx = \sqrt{2}\int \frac{1+\frac{u^2}{2-u^2}}{1-\frac{u^2}{2-u^2}}\, du \\ &= \sqrt{2}\int \frac{1}{1-u^2}\, du = \sqrt{2}\operatorname{arctanh}(u) + c = \sqrt{2}\operatorname{arctanh}\left(\frac{\sqrt{2}x}{\sqrt{1+x^2}}\right) + c, \end{aligned} \tag{4.45}$$

with $c \in \mathbb{R}$ some constant. Consequently, the loss function is given by

$$\begin{aligned} L(\rho,\hat{\rho}) &= \left|\int_{\rho}^{\hat{\rho}} \frac{\sqrt{1+x^2}}{1-x^2}\, dx\right| \\ &= \left|\left[\sqrt{2}\operatorname{arctanh}\left(\frac{\sqrt{2}x}{\sqrt{1+x^2}}\right) - \operatorname{arcsinh}(x)\right]_{\rho}^{\hat{\rho}}\right|. \end{aligned} \tag{4.46}$$

The Bayes estimator then becomes

$$\begin{aligned} \hat{\rho}_I = \operatorname*{argmin}_{\hat{\rho}} \int_{-1}^{1} \Bigg| &\sqrt{2}\operatorname{arctanh}\left(\frac{\sqrt{2}\hat{\rho}}{\sqrt{1+\hat{\rho}^2}}\right) - \operatorname{arcsinh}(\hat{\rho}) \\ &- \sqrt{2}\operatorname{arctanh}\left(\frac{\sqrt{2}\rho}{\sqrt{1+\rho^2}}\right) + \operatorname{arcsinh}(\rho)\Bigg| \pi(\rho \,|\, x)\, d\rho. \end{aligned} \tag{4.47}$$

Fixing $\hat{\rho}$ and looking at the loss function as a function of ρ only, we have that $L(\rho,\hat{\rho}) \geq 0$, and is equal to zero if and only if $\rho = \hat{\rho}$. This is as expected, since equation (4.46) should

define a metric. Further, since the value of $L(\rho, \hat{\rho}) \to \infty$ as $\rho \to \pm 1$, and $\pi(\rho \mid x) \geq 0$ and bounded, the minimum does exist. Note, however, that the value of the estimator, as was the case with the Bayes estimator using Kullback–Leibler divergence as loss function, has to be found numerically.

4.8 Fiducial approach

In Taraldsen and Lindqvist, 2018, a method for sampling from the fiducial of ρ is discussed. We will not show the correctness of the method here, but simply apply the resulting sampling procedure in good faith. As discussed in section 2.3, such a sampling procedure might be used to approximate the fiducial, which in turn can be viewed as a posterior distribution that does not correspond to any prior distribution.

Let r denote the empirical correlation introduced in section 4.2.2, and calculate

$$x = \frac{r}{\sqrt{1 - r^2}}. \tag{4.48}$$

Next, sample $u_1 \sim \chi^2_{n-1}$, $u_2 \sim \chi^2_{n-2}$ and $u_3 \sim \mathrm{N}(0, 1)$, and calculate

$$\theta = \frac{x u_2 - u_3}{u_1}. \tag{4.49}$$

The sample from the fiducial is then given by

$$\rho = \frac{\theta}{\sqrt{1 + \theta^2}}. \tag{4.50}$$

The obvious choice of r would be the empirical correlation given in equation (4.8). However, knowing that the means of our bivariate normal model are both 0, it might make sense to use equation (4.9) instead. We will here make use of both, in order to see which choice is better for our model. Note that, since the empirical correlation given by equation (4.10) would give values of r outside the interval $(-1, 1)$, which would subsequently give the square root of a negative number in equation (4.48), it does not make sense to use this quantity in the fiducial setting.

In order to see how the fiducial behaves, we can sample several times from the bivariate normal for a given ρ, and for each sample calculate the sample value from the fiducial by using the procedure above. This, in effect, gives us an empirical distribution approximating the true fiducial. By applying a kernel density estimator, we can get a continuous approximation to the fiducial. In figure 4.10 we have done this for $\rho = 0.5$,

bivariate sample sizes 3, 10 and 100, and fiducial sample size 10 000. The fiducial has been calculated using both the standard empirical correlation and the one where we assume known means. For each bivariate sample size, the fiducial was approximated 100 times, and the average of these is displayed in the figure.

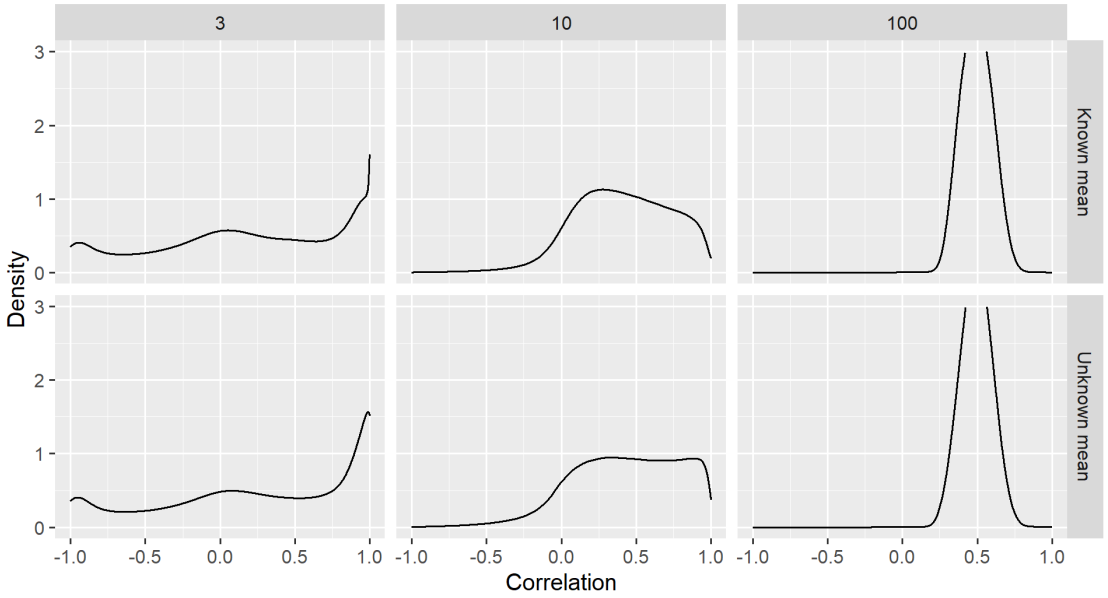


Figure 4.10: An average over 100 approximated fiducials calculated for sample sizes 3, 10 and 100, with $\rho = 0.5$. The fiducials were approximated with 10 000 sample points, before using a kernel density estimator. Note that the number above each plot indicates the sample size for those averaged fiducials, whereas the labels on the right indicate whether or not known means were assumed.

First of all, we see that the two versions of the fiducial behave quite similarly, but that there are some differences. For a sample size of 3 we seem to get some features similar to that of the PC priors, and some features closer to the flat and Jeffreys priors. More concretely, the distribution places a lot of weight on $\rho \to 1$, while at the same time having a local maximum close to $\rho = 0$.

For a sample size of 10, the weight of the distribution is shifted towards the true value of the parameter. Using known means we have a larger weight on values between $\rho = 0$ and 0.5 than is seen for the priors (except for the PC priors with $\lambda = 5.360$ which behaves rather strangely), whereas the fiducial is more balanced around $\rho = 0.5$ for the regular empirical correlation.

When the sample size is increased to 100, the distribution seems to approach a univariate normal distribution with mean 0.5, as was the case for all the priors as well, and there are no clear distinctions between the two versions of the fiducial.

Chapter 5

Simulations

If we are to make any statements about the suitability of the various methods of estimation discussed in the previous chapters, we need to perform experiments in which we compare the performance of the estimators on simulated data. This chapter includes such experiments, and subsequent discussion of their results. We will start out by looking at the suitability of the posterior distributions as tools to construct confidence intervals. Then, we will compare the point estimators using the empirical distribution of the estimators themselves, as well as the distance between the point estimates and the true value of the correlation. The distance measures used will be Kullback–Leibler divergence and the Fisher information metric. Lastly, we will compare the different methods for constructing confidence intervals.

The code used in this section can be found in an open repository in Github, by following the link https://github.com/erikhide/TMA4900. Note that the code is not written to be easy to read, and there are no comments. However, it serves as documentation for the simulations performed in this thesis.

Throughout this section, the number of iterations are relatively low. The reason for this was a combination of too little time and too high computational complexity. Some results might therefore be unreliable, and the conclusions should be viewed in light of this. In order to improve upon the work done in this thesis, a similar simulation study should be carried out with more iterations and hence a higher accuracy.

5.1 Frequentist coverage

Credible intervals derived from posteriors (as explained in section 2.2.3) might not actually deliver the promised degree of certainty concerning the parameter value, when viewed as confidence intervals. That is, for a $1-\alpha$ credible interval, the probability of covering the actual parameter value might not be $1-\alpha$. By simulating a large number of random samples from the bivariate normal distribution, and for each sample calculate

the credible interval in order to check whether or not the (now known) value of ρ is actually inside the interval, we can approximate the frequentist coverage. Specifically, the approximation to the frequentist coverage becomes equal to the number of times the interval actually covers ρ divided by the number of iterations performed. Mathematically, using the definition of $C(\boldsymbol{x})$ from section 2.2.3, we use the following approximation

$$P\big[\rho \in C(\boldsymbol{x}) \,|\, \rho\big] \approx \frac{1}{m} \sum_{i=1}^{m} \big[\rho \in C(\boldsymbol{x}_i)\big], \tag{5.1}$$

with m equal to the number of iterations, and $\boldsymbol{x}_i = (x_{i1}, ..., x_{in})$ with n equal to the sample size and x_{ij} being one realisation from the bivariate normal distribution. The square bracket notation $[\cdot]$ surrounding $\rho \in C(\boldsymbol{x}_i)$ denotes the Iverson bracket (see Knuth, 1992), which returns 1 if the logical statement inside is true, and 0 otherwise.

In the following simulations, we will consistently use 10 000 iterations per value of ρ in an attempt to strike a balance between accuracy in the estimates on the one hand and computational complexity on the other. In order to test the prior distributions on the most extreme case, and also see how they perform as the sample size increases, simulations will be done for samples of size 3, 10 and 100. For each combination of sample size and prior distribution, the frequentist coverage will be estimated for $\rho \in \{-0.9, -0.8, ..., 0.8, 0.9\}$. Due to the approximate nature of the estimation method, we would expect some variation in the computed frequentist coverages. Hence, some deviation from the desired 0.95 coverage probability is natural. However, due to the relatively high number of iterations, larger deviations will likely be caused by a systematic difference between the promised and actual coverage probabilities. Even though the coverage probabilities should theoretically be symmetric around $\rho = 0$, the relatively small number of iterations results in some deviations from this.

5.1.1 Flat prior

Figure 5.1 shows the estimated coverage probabilities when using a flat prior. An immediate observation is that the coverage probabilities are consistently quite close to 0.95. For a sample size of 3 the coverage probabilities are a bit too low, fluctuating around 0.945, whereas for sample sizes of 10 and 100 the coverage probabilities fluctuate around 0.95, as promised.

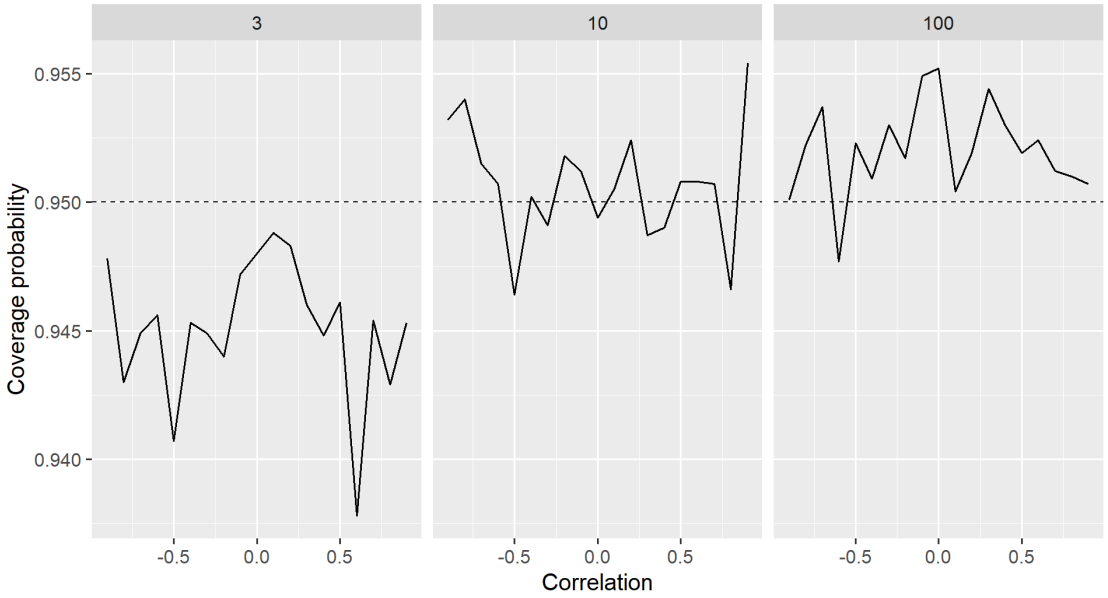


Figure 5.1: The estimated coverage probabilities for the **flat prior** is shown for sample sizes 3, 10 and 100, and for ρ values between -0.9 and 0.9 with step size 0.1. A total of 10 000 iterations was performed per value of ρ. Note that the number above each plot indicates the sample size.

5.1.2 Jeffreys prior

Figure 5.2 shows the estimated coverage probabilities when using the Jeffreys prior. For sample size 3 the coverage probabilities are all close to 0.85, which is quite low as compared to the promised 0.95. Increasing the sample size to 10 shifts the coverage probabilities to 0.92, which is closer to 0.95 but still quite a lot off. For a sample size of 100, the coverage probabilities are all close to 0.95.

5.1.3 Penalised Complexity prior

Figure 5.3 shows the estimated coverage probabilities when using the PC prior with λ equal to 0.0818, 1.249 and 5.360. One of the most striking features common for all λs is how the behaviour of the coverage probabilities changes significantly close to -1 and 1. Further, both for $\lambda = 1.249$ and $\lambda = 5.360$, the coverage probability is growing as we move closer to $\rho = 0$.

Both for $\lambda = 0.0818$ and $\lambda = 1.249$, the coverage probabilities fairly close to the goal of 0.95, especially as the sample size grows. However, the coverage probability is not stable for different values of ρ, but rather it varies quite a lot from values well above 0.95 to values a great deal lower than what we were hoping for.

For $\lambda = 5.360$, the coverage probability estimates are behaving very strangely, and

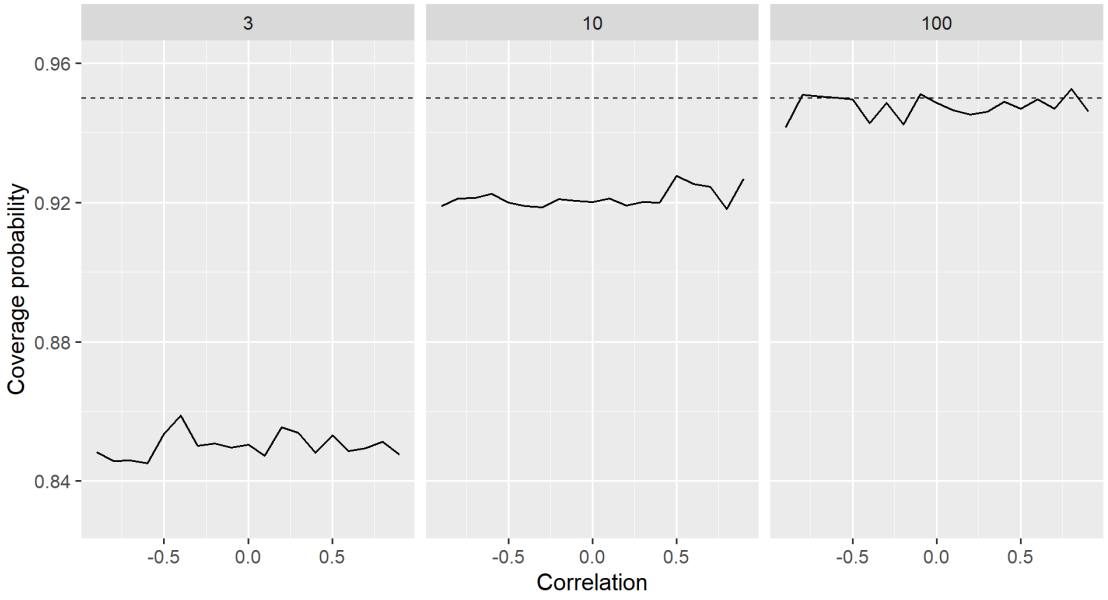


Figure 5.2: The estimated coverage probabilities for the **Jeffreys prior** is shown for sample sizes 3, 10 and 100, and for ρ values between -0.9 and 0.9 with step size 0.1. A total of 10 000 iterations was performed per value of ρ. Note that the number above each plot indicates the sample size.

the sample size has to reach 100 before they can be said to be anywhere close to 0.95.

5.1.4 Fiducial approach

Since the fiducial might be viewed as a posterior distribution, it would be interesting to see how well it performs with regards to frequentist coverage. For computational reasons, the number of iterations was restricted to 1000 for each value of ρ, and each fiducial distribution was approximated using just 1000 samples.

First of all, figure 5.4, which shows the estimated coverage probabilities using both versions of the empirical correlation, shows that there are no discernible differences between the two. We then note that for samples of size 3, the coverage probability is close to or equal to 1 for all values of ρ. For samples of size 10, the coverage probability falls towards 0.75 as ρ approaches 0, but stays close to 1 at the extreme ends. Then, for samples of size 100, the tendencies we say for samples of size 10 are reinforced, with the coverage probability being as low as 0.20 for $\rho = 0$.

5.1.5 Summary of the coverage probabilities

In conclusion, it seems as though the only posterior (including the fiducials) able to achieve something resembling exact frequentist matching for samples of sizes 3, 10 and

100, is the posterior given by a flat prior. Jeffreys prior does give almost the same coverage probability for all values of ρ, but the values are close to 0.85 for samples of size 3, close to 0.92 for samples of size 10, and only for samples of size 100 do the values approach 0.95. This result is a bit surprising, since one would assume that the Jeffreys prior would outperform the flat prior when comparing the frequentist coverage results.

For the PC priors, the version with $\lambda = 5.360$ behaves very strangely for all sample sizes, whereas the other two have problems for extreme values of ρ. The best choice out of the three PC priors seem to be $\lambda = 1.249$, which lies close to 0.95 for all values of ρ and all sample sizes. The fiducials do not seem to be appropriate for interval estimation at all, displaying some very strange behaviour.

When looking closely at the coverage probability plots, there seems to be a lack of symmetry especially for the flat and Jeffreys priors. This is most likely caused by having relatively few iterations when calculating the coverage probabilities.

5.2 Evaluating point estimators

There are several approaches to evaluating point estimators. We might for example run a simulation in which we calculate the estimator a large number of times, and view the resulting list of values as a sample from the distribution of the estimator. Using this empirical distribution, we can either calculate the mean and variance directly, or we can apply kernel density estimation to get an approximate density for the estimator and use this to find the same quantities. In the end, we would like the mean of the distribution to be close to the actual value of the parameter, while keeping the variance as low as possible. A biased estimator – that is, one that in expectation gives a too low or too high estimate – is undesirable.

Another, possibly simpler, approach is to subtract the true value from the estimated value and square the result. We then do this a large number of times before finding the average. This is the mean square error (MSE), and it is commonly used in the evaluation of estimates.

If we would like to be a bit more clever, we might want to measure distances between distributions, instead of just the difference between the estimated and true value the parameter. Indeed, one could argue that the goal of estimating the parameter is to get an estimate of the true underlying model, and in this sense, comparing the models seems more correct than simply comparing the values of the parameters. In section 2.2.2

we discussed distances between distributions in relation to the Bayes estimator, and two distance measures were introduced. The expression for the Kullback–Leibler divergence was derived in equation (4.29), whereas the Fisher information metric is given in equation (4.46). For both of these measures of deviation between the true and estimated distributions, we would like the value to be as close to 0 as possible.

In this section we will look closer at how our estimators perform both when looking at the mean and variances of their distributions, and when using Kullback–Leibler divergence and the Fisher information metric to calculate the distance between the estimated and true models. Due to symmetry, we will only look at values of ρ in the interval $[0, 1)$. Note that the approximative nature of the numerical methods used to estimate the parameters, as well as the finite sample size used to get the empirical distribution of the estimators, there would have been some deviation between the results for ρ and $-\rho$. Theoretically, however, the results should be the same, and indeed, when performing some tests, they were quite similar even for modestly large samples.

5.2.1 The distribution of the estimators

Figure 5.5 shows the approximated densities of the point estimators when using prior distributions and Bayes esimators. One apparent feature of these plots is the strange behaviour of the PC prior with $\lambda = 5.360$. No matter the true value of ρ, the distribution of the estimators all seem to be quite close to 0, which reflects the fact that the prior gives a lot of weight to values close to this point. Due to this, we will focus on the four other prior choices. Another important point is that using Kullback–Leibler divergence and mean square error as loss functions produce very similar estimator densities. It therefore makes sense to compare these two taken together against the estimator distributions when using the Fisher information metric as loss function.

Looking first at $\rho = 0$ (i.e. the first row), we wee that all distributions are close to symmetric around 0, and hence are likely to provide good estimates in expectation for true values of ρ close to 0. It seems that the Fisher information metric gives a flatter distribution than the other two, with especially Jeffreys prior and the PC prior with $\lambda = 0.0818$ displaying local maxima close to ± 0.9. Overall, the flat prior and the PC prior with $\lambda = 1.249$ using Kullback–Leibler divergence or mean square error as loss function are the ones that seem to give the estimator distributions with most weight located close to 0, and hence the ones that will likely perform best in this case.

Moving on to $\rho = 0.4$ (i.e. the second row), we see that also in this case do the Jeffreys

prior and PC prior with $\lambda = 0.0818$ give more weight to values close to ± 1. The both have a small bump around -0.9 with the Fisher information metric as loss function, and a lot of weight focused between 0.3 and 0.9 for all loss functions. The flat prior seems to give a density with a maximum close to the true value of ρ, with the Fisher information metric case giving a distribution that is farther away from 0 than the other two. The PC prior with $\lambda = 1.249$ is again quite close to the flat prior, but now slightly closer to 0, especially in the Fisher information metric case, where there is a clear difference between the two. Based on just looking at the plots, it seems a flat prior with loss function given by the Fisher information metric is a good choice for true values of ρ close to 0.4.

For $\rho = 0.8$ (i.e. the third row), we again see that using the Fisher information metric as loss function gives more weight closer to the extremes. Further, it seems to give more weight close to the true value of ρ, whereas the other two loss functions tend to be drawn towards 0. For this case, the Jeffreys prior and PC prior with $\lambda = 0.0818$ seem to be the best choices. Which loss function is better is not readily apparent.

Figure 5.6 shows the approximated densities of the point estimators when using fiducials and Bayes esimators. These results are indeed quite strange, and after approximating the distributions a few times it becomes clear that the behaviour changes from one simulation to the next. One likely explanation for this strange behaviour is the fact that we have to sample from and then approximate the fiducial, whereas for the priors we had a closed form expression. The approximations of the fiducials would improve with a larger sample, but due to time constraints, a relatively small sample size (i.e. 1000) was used for this part of the simulation.

Figure 5.7 shows the approximated densities of the point estimators when using the empirical correlation with known mean and the maximum likelihood estimator. For $\rho = 0$, we again have the symmetry around the true value. However, both densities give a lot of weight to more extreme values. Indeed, the MLE resembles the behaviour seen when using Jeffreys prior with the Fisher information metric as loss function, with local maxima close to ± 0.8.

For $\rho = 0.4$, we se that more weight is placed on the interval $(0, 1)$, but that both densities still exhibit much of the same behaviour as for the case $\rho = 0$. For $\rho = 0.8$, however, both densities place much weight on values close to the true parameter value. Note that the MLE still places some weight on values of ρ between -1 and -0.8, but almost no weight on values between -0.8 and 0.5. Especially for high values of 0.9, the two estimators considered in this last figure showed a lesser tendency than the prior

approaches for the estimates to lie closer to 0 than the true value.

5.2.2 Mean and variance

In order to find the mean and variance of the estimator distributions, we sample 10 000 values from each estimator, and use these empirical distributions to approximate a density using kernel density estimation. The mean and variance are then found through integration. Table 5.1 shows the results of such an experiment for all the estimators used in this chapter, with values calculated for $\rho = 0.1$, 0.3, 0.5, 0.7 and 0.9.

A comment about the fiducials: The results here seem to be all over the place, and it is not immediately evident that they are of any value. There might be several reasons for this, but one likely explanation is the approximative nature of the fiducials used in the calculations, as mentioned in section 5.2.1. In other words, using a larger number of samples to approximate each fiducial might result in a significant improvement. However, due to time constraints, tests to verify this claim were not performed.

A general observation is that all estimated means (ignoring the fiducials) are below the true value of ρ. Hence, all estimators seem to be biased towards the centre of the parameter domain. Ideally, we would like to have unbiased estimators, but in the absence of such an estimator, the best choice seems to be the one with lowest bias. However, we should still consider the variance before concluding.

If we shift our attention towards the last two rows, containing the results for the maximum likelihood estimator and the empirical correlation with known mean, the MLE is the least unbiased estimator, with an estimated mean between 0.58 and 0.12 below the true value of ρ. In fact, the only value for which the empirical correlation performs better is $\rho = 0.1$, and here the difference is tiny. However, if we look at the estimated variance, there is a huge difference between the estimators, with the empirical correlation performing much better. The difference is not that large for $\rho = 0.9$, but for the other values the difference ranges between 0.75 and 1.12. In conclusion, it is not immediately apparent which estimator is better.

Moving on to the estimators we get when using priors together with Bayes estimators, we have emphasised the three least biased estimators for each value of ρ by showing their estimated means in bold. First, we see that we can probably ignore the PC prior with $\lambda = 5.360$, since the estimated means are so far off. Two estimators that clearly stand out are the Jeffreys prior and PC prior with $\lambda = 0.0818$ using the Fisher information metric as loss function. In fact, it seems that the Fisher information metric should be

the preferred loss function if only the estimated means are taken into account.

If we look closely at the variances, however, using either mean square error or Kullback–Leibler divergence as loss function seems to give much lower estimated variances for all values of ρ. Hence, there appears to be a bias–variance tradeoff between the Fisher information metric on one side and the MSE and Kullback–Leibler divergence on the other.

Anther bias–variance tradeoff is between the Jeffreys prior and PC prior with $\lambda = 0.0818$ on one side and the flat prior and PC prior with $\lambda = 1.249$ on the other. This was also seen and discussed in relation to figure 5.5 in section 5.2.1. In the end, one might prefer a heavily biased estimator with low variance, in which case a flat prior or a PC prior with $\lambda = 1.249$ with loss function given by either the MSE or Kullback–Leibler divergence seems appropriate, or one might prefer a less biased estimator with a higher variance, in which case choosing Jeffreys prior or the PC prior with $\lambda = 0.0818$ with loss function given by the Fisher information metric seems reasonable. One might also prefer something in between, in which case using Jeffreys prior together with Kullback–Leibler divergence as loss function seems to be a good choice.

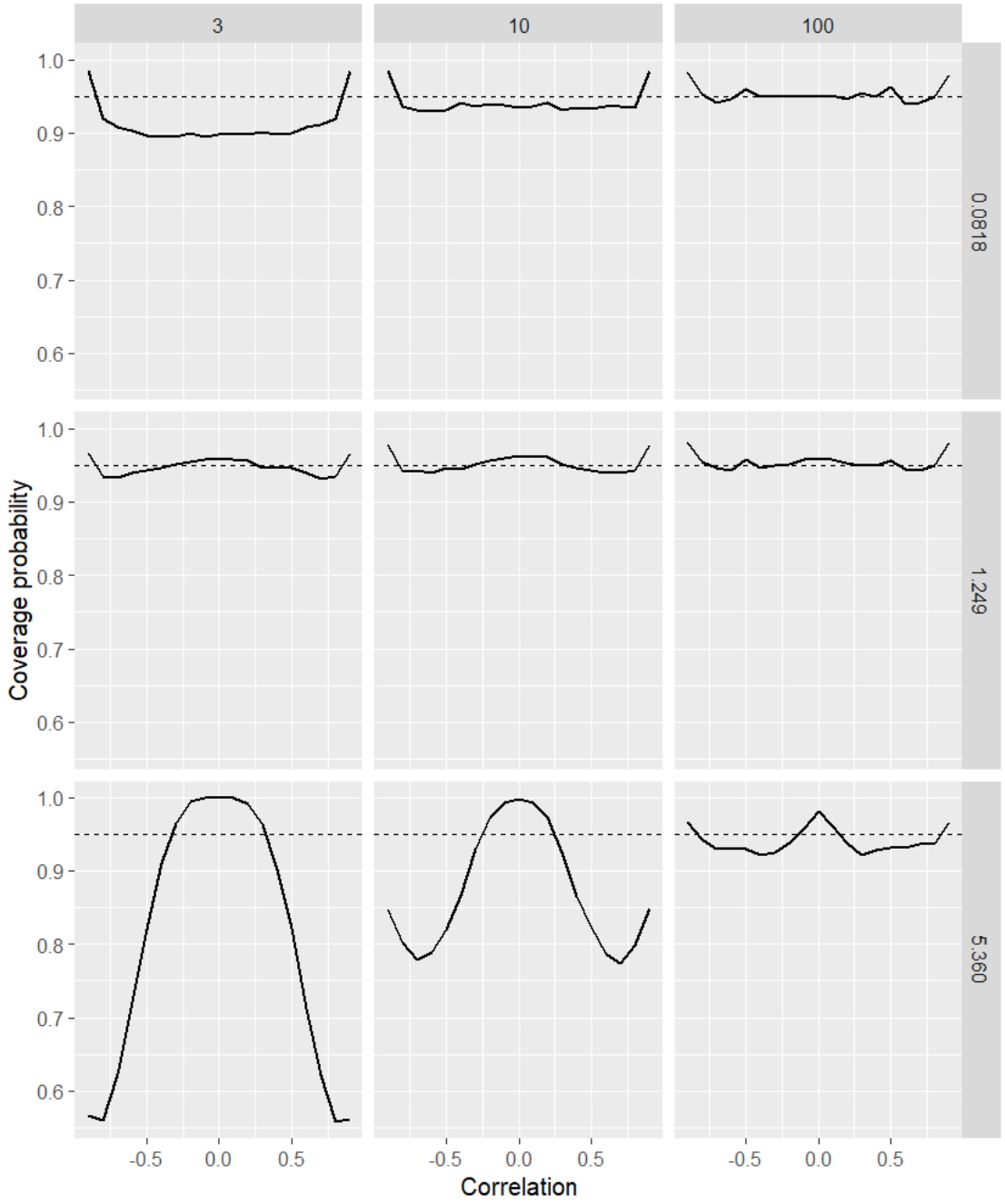


Figure 5.3: The estimated coverage probabilities for the **PC prior** with $\lambda = 0.0818$, 1.249 and 5.360 is shown for sample sizes 3, 10 and 100, and for values of ρ between -0.9 and 0.9 with step size 0.1. A total of 10 000 iterations was performed per value of ρ. Note that the number on top of each column indicates the sample size, whereas the number to the far right of each row indicates the value of λ.

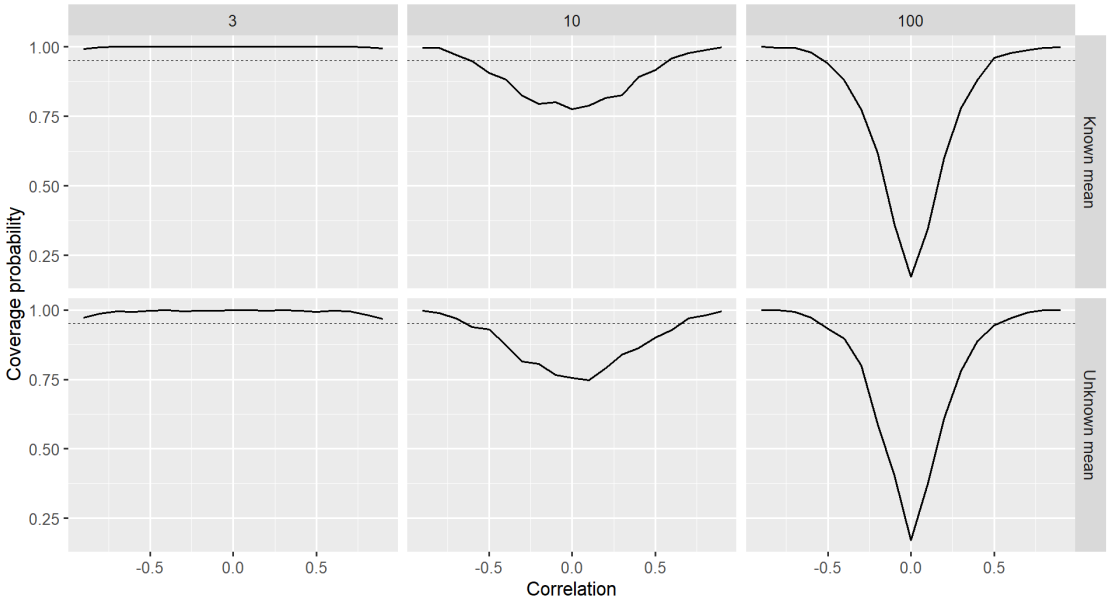


Figure 5.4: The estimated coverage probabilities for the **fiducial** is shown for sample sizes 3, 10 and 100, and for values of ρ between -0.9 and 0.9 with step size 0.1. A total of 1000 iterations was performed per value of ρ, and the fiducials were approximated using 1000 samples. Note that the number on top of each column indicates the sample size, whereas the number to the far right of each row indicates whether the empirical correlation with known or unknown means was used.

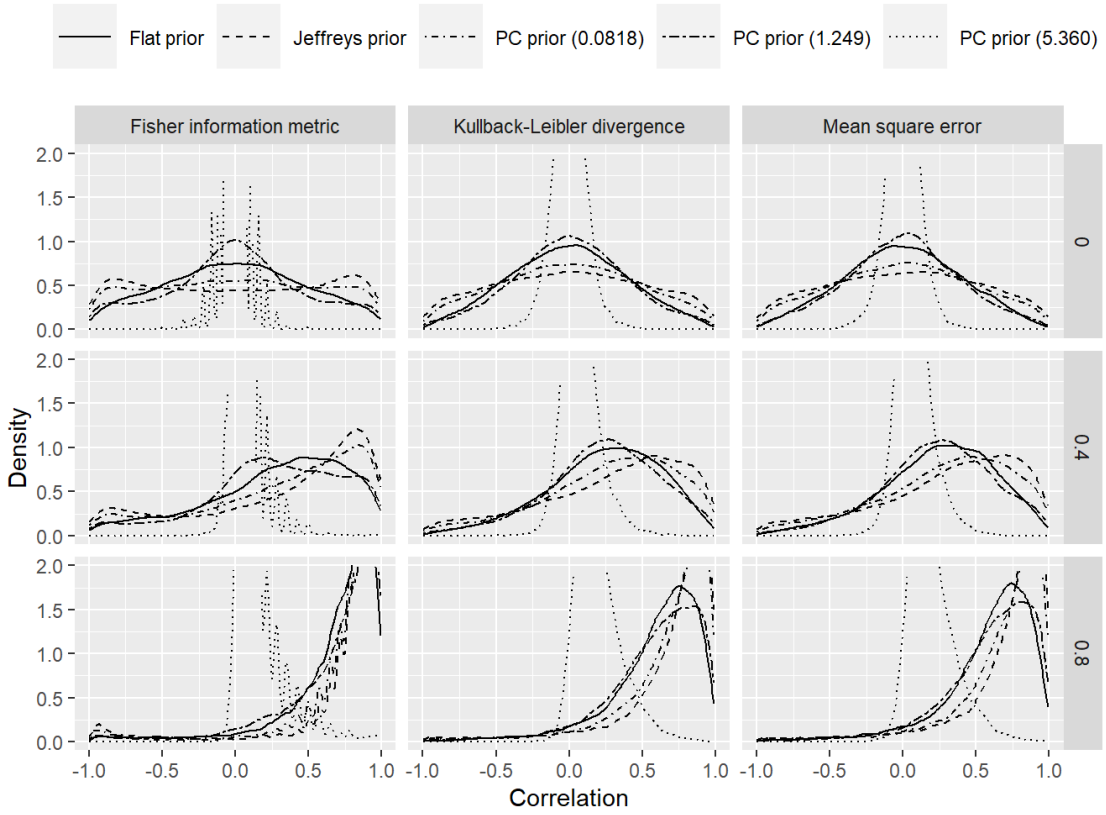


Figure 5.5: Approximations of the distributions of point estimators using **prior distributions and Bayes estimators** are shown, using samples of size 3. 10 000 values of each estimator were found, before approximating the resulting empirical distribution with kernel density estimation. Note that the values to the right of the plots indicate the value of ρ used, whereas the information on top indicate which loss function was used in the Bayes estimator.

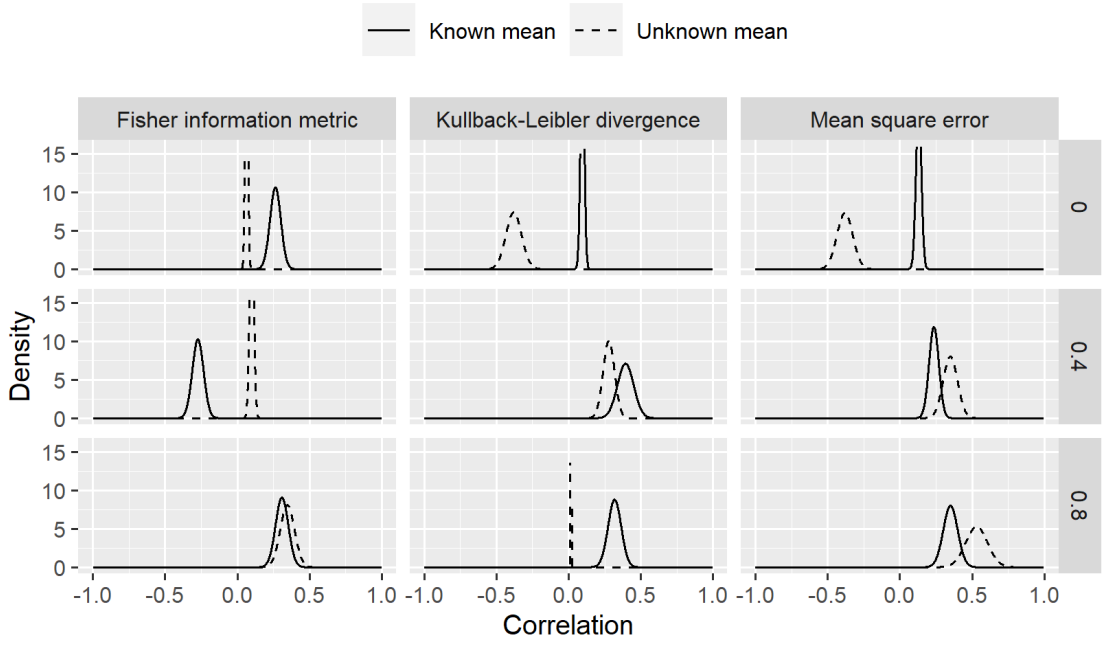


Figure 5.6: Approximations of the distributions of point estimators using **the fiducials and Bayes estimators** are shown, using samples of size 3. The fiducials were approximated using 1000 samples and a kernel density estimation, before calculating the estimator. 10 000 values of each estimator were found, before approximating the resulting empirical distribution with yet another kernel density estimation. Note that the values to the right of the plots indicate the value of ρ used, whereas the information on top indicate which loss function was used in the Bayes estimator.

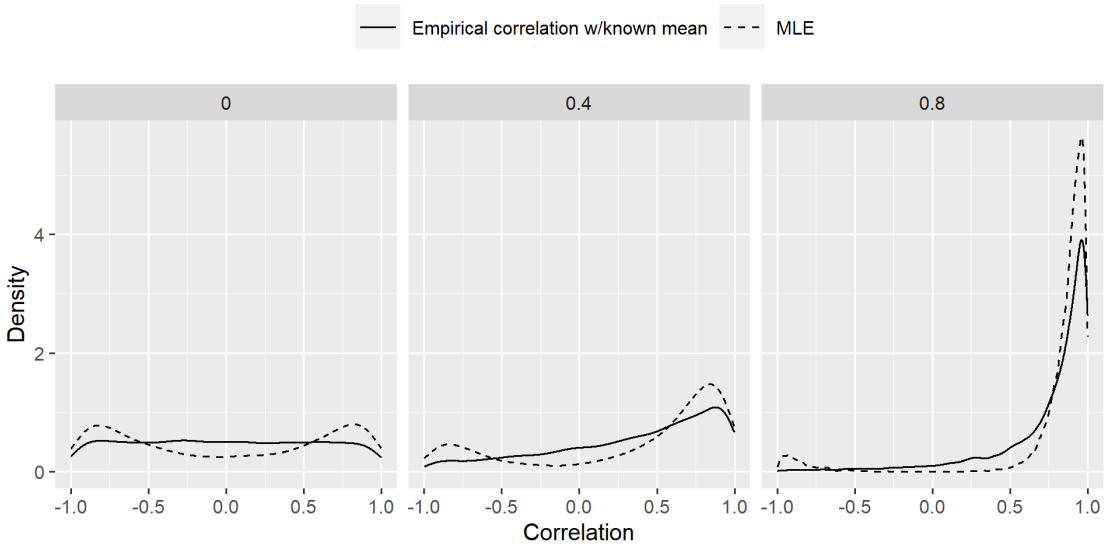


Figure 5.7: Approximations of the distributions of point estimators using **empirical correlation** with known mean and the **maximum likelihood estimator** are shown, using samples of size 3. 10 000 values of each estimator were found, before approximating the resulting empirical distribution with kernel density estimation. Note that the values on top of the plots indicate the value of ρ used.

Table 5.1: Estimated means and variances of the distribution of point estimators are shown in the form $(\hat{\mu}, \hat{\sigma}^2)$. The point estimators were calculated using samples of size 3. The means and variances were then found by first sampling 10 000 values of each estimator, before approximating this empirical distribution with kernel density estimation. The resulting function was then used to derive the values in the table. For each ρ value, the three estimates among the Bayes estimators closest to the true value of the mean are shown in bold.

Estimator		**ρ value**				
Bayesian						
Prior	*Loss function*	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*
Flat	MSE	(0.058, 0.160)	(0.178, 0.154)	(0.307, 0.142)	(0.473, 0.115)	(0.727, 0.056)
Jeffreys	MSE	(0.064, 0.254)	(**0.209**, 0.236)	(**0.370**, 0.205)	(0.558, 0.153)	(0.822, 0.053)
PC (0.0818)	MSE	(**0.065**, 0.215)	(0.196, 0.205)	(0.344, 0.183)	(0.527, 0.138)	(0.795, 0.054)
PC (1.249)	MSE	(0.051, 0.154)	(0.174, 0.148)	(0.298, 0.141)	(0.461, 0.116)	(0.728, 0.059)
PC (5.360)	MSE	(0.020, 0.014)	(0.057, 0.016)	(0.100, 0.017)	(0.159, 0.021)	(0.269, 0.034)
Flat	KL div.	(0.058, 0.163)	(0.167, 0.156)	(0.308, 0.141)	(0.471, 0.112)	(0.727, 0.054)
Jeffreys	KL div.	(0.063, 0.248)	(0.203, 0.235)	(0.359, 0.210)	(**0.561**, 0.146)	(**0.823**, 0.051)
PC (0.0818)	KL div.	(**0.068**, 0.214)	(0.200, 0.206)	(0.339, 0.187)	(0.527, 0.141)	(0.795, 0.053)
PC (1.249)	KL div.	(0.048, 0.154)	(0.161, 0.150)	(0.288, 0.142)	(0.455, 0.121)	(0.727, 0.061)
PC (5.360)	KL div.	(0.019, 0.014)	(0.058, 0.015)	(0.100, 0.017)	(0.159, 0.022)	(0.274, 0.033)
Flat	FI metric	(0.059, 0.226)	(0.204, 0.214)	(0.346, 0.197)	(0.528, 0.156)	(0.802, 0.062)
Jeffreys	FI metric	(**0.069**, 0.334)	(**0.228**, 0.316)	(**0.395**, 0.275)	(**0.614**, 0.198)	(**0.871**, 0.059)
PC (0.0818)	FI metric	(0.063, 0.293)	(**0.224**, 0.276)	(**0.380**, 0.246)	(**0.575**, 0.190)	(**0.857**, 0.063)
PC (1.249)	FI metric	(0.057, 0.207)	(0.178, 0.203)	(0.332, 0.183)	(0.515, 0.149)	(0.807, 0.063)
PC (5.360)	FI metric	(0.015, 0.009)	(0.045, 0.010)	(0.076, 0.014)	(0.126, 0.021)	(0.229, 0.047)
Fiducial						
Known mean?	*Loss function*					
No	MSE	(0.223, 0.001)	(0.216, 0.001)	(-0.119, 0.000)	(0.252, 0.001)	(0.302, 0.002)
No	KL div.	(0.031, 0.071)	(0.086, 0.069)	(0.149, 0.061)	(0.225, 0.048)	(0.327, 0.028)
No	FI metric	(0.041, 0.186)	(0.125, 0.182)	(0.220, 0.167)	(0.340, 0.152)	(0.529, 0.090)
Yes	MSE	(0.257, 0.001)	(0.211, 0.001)	(0.299, 0.002)	(0.236, 0.001)	(-0.185, 0.001)
Yes	KL div.	(0.032, 0.040)	(0.086, 0.037)	(0.151, 0.031)	(0.219, 0.023)	(0.320, 0.011)
Yes	FI metric	(0.029, 0.087)	(0.132, 0.094)	(0.224, 0.089)	(0.356, 0.065)	(0.529, 0.046)
Other approaches						
Maximum likelihood		(0.077, 0.410)	(0.243, 0.379)	(0.442, 0.333)	(0.649, 0.235)	(0.888, 0.064)
Empirical correlation w/known mean		(0.078, 0.298)	(0.234, 0.275)	(0.387, 0.231)	(0.567, 0.160)	(0.786, 0.059)

5.2.3 Error using Kullback–Leibler divergence

In order to get more information concerning which estimators should be preferred, we have estimated the Kullback–Leibler divergence between the model using the true parameter value and the estimated value. For each estimator, and for the same values of ρ as in section 5.2.2, 10 000 estimated values were found and used to calculate the Kullback–Leibler divergence. The averaged results multiplied by a factor of 1000 are shown in table 5.2.

Since we do not know the value of ρ, we would, ideally, like the error to be as small as possible for all possible values of ρ. For our simple experiment, this entails choosing an estimator that gives relatively low errors for all five values of ρ used.

An interesting point is that the Kullback–Leibler divergence penalises deviations that are too extreme more than deviations that are too close to 0. Hence, an estimator whos distribution places most weight near the centre of the domain will give smaller values than one which e.g. places half the weight above the true parameter value, even though the latter should usually be preferred. Hence, we need to be careful when looking at the errors in table 5.2, and make sure to analyse them in relation to other results such as the mean and variance approximations discussed in section 5.2.2.

Due to the preference for small absolute values of ρ on the part of the error using Kullback–Leibler divergence, the PC prior with $\lambda = 5.360$ performs very well, especially for values of ρ up to 0.7. However, there is a sharp increase in the error from 11 to 48 for $\rho = 0.1$, up to 607 to 735 for $\rho = 0.9$. This indicates that, even though the estimator seemingly performs quite well for low values of ρ, it does not work well for all values of ρ, and it gets consistently worse as the absolute value of ρ increases. Indeed, this aligns well with the results from sections 5.2.1 and 5.2.2.

Shifting attention to the fiducials, we see a similar pattern as for the PC prior with $\lambda = 5.360$, with the error being quite small for small absolute values of ρ, but increasing by a lot as ρ approaches 1. Due to the quite strange performance of these estimators, as seen in section 5.2.1, it is difficult to draw conclusions, but it does seem like the estimators return values close to 0, in the same way that the PC prior with $\lambda = 5.360$ does. One exception is the fiducial estimator found when using the full empirical correlation with the Fisher information metric as loss function, where the values are much higher, and do not follow the same pattern with higher values for ρ closer to 1. We have not been able to find any reasonable explanations for this behaviour, but the same results are seen

when rerunning the simulation.

Looking at the two last rows of table 5.2, we see that both the maximum likelihood estimator and the empirical correlation with known mean produces very large errors, with the latter giving errors about twice as large as the former. One might have thought that with the MLE being less biased (i.e. further away from 0 in expectation) and having a higher variance, would have caused it to have higher error when using Kullback–Leibler divergence as well. This, however, does not seem to be the case.

Moving on to the Bayes estimators (excluding the PC prior with $\lambda = 5.360$), we see that the less biased estimators, namely Jeffreys prior and the PC prior with $\lambda = 0.0818$, give higher errors than the more biased estimators, namely the flat prior and PC prior with $\lambda = 1.249$. Since the Kullback–Leibler divergence is larger for deviations closer to the extremes, this result should be expected; the more biased estimators have a lot of weight between 0 and the true value of ρ, whereas the less biased estimators place more weight on values above the true value of ρ.

The differences between using mean square error or the Kullback–Leibler divergence as loss functions are mostly quite small, and it is difficult to say anything conclusive about differences in performance. However, as was the case in section 5.2.2, there are clear differences between these two loss functions on one side, and the Fisher information metric on the other. The latter did produce less biased estimators with higher variance, and it also gives a sharp increase in the Kullback–Leibler divergence. In short, it seems that the Kullback–Leibler divergence penalises estimators that give frequent and large deviations from 0, but gives low errors to estimators that frequently returns estimates close to 0, no matter the true value of ρ.

How do we pick the best estimator based solely on the error using Kullback–Leibler divergence? One way to do this, is by looking at the largest error produced by each estimator, and pick the estimtor whose largest error is smallest. In such a case, the flat prior with mean square error as loss function would be the preferred choice. However, it is not at all clear that this is the best approach, and in fact it does not fit the conclusions when considering means and variances of the estimator distributions, and preferring a less biased estimator, as we did in section 5.2.2. The result does fit with the desire to minimise the variance, at the cost of getting a larger bias. Note also that it does not pick the PC prior with $\lambda = 5.360$ (or any of the fiducials) even though their variances are very small, which is good, since all of these seem to be useless.

Table 5.2: The error using **Kullback–Leibler divergence as measure of error** for the point estimators are shown. The point estimators were calculated using samples of size 3. The error was found by sampling 10 000 values of each estimator, and calculating the Kullback–Leibler divergence between the model using the estimator and the true value. The values shown are the average of these *errors multiplied by a factor of 1000*.

Estimator		**ρ value**				
Prior	*Loss function*	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*
Bayesian						
Flat	MSE	312	307	323	292	367
Jeffreys	MSE	1154	1197	1099	1226	888
PC (0.0818)	MSE	946	937	935	721	795
PC (1.249)	MSE	472	447	524	538	715
PC (5.360)	MSE	12	40	105	243	607
Flat	KL div.	318	308	370	341	509
Jeffreys	KL div.	1180	1030	1100	1022	1049
PC (0.0818)	KL div.	964	881	806	818	869
PC (1.249)	KL div.	484	425	601	530	788
PC (5.360)	KL div.	11	42	107	611	729
Flat	FI metric	1052	1127	955	900	682
Jeffreys	FI metric	1886	2029	2202	2093	1516
PC (0.0818)	FI metric	1796	1706	1682	2105	1491
PC (1.249)	FI metric	1231	1190	1211	1296	1102
PC (5.360)	FI metric	48	153	137	297	735
Fiducial						
Known mean?	*Loss function*					
No	MSE	68	364	113	159	503
No	KL div.	45	65	110	215	565
No	FI metric	2278	1048	1913	2201	2329
Yes	MSE	31	68	52	205	536
Yes	KL div.	25	43	92	207	572
Yes	FI metric	69	79	100	174	435
Other approaches						
Maximum likelihood		2512	2691	2408	2384	2145
Empirical correlation w/known mean		4870	6498	4083	4014	5225

5.2.4 Error using the Fisher information metric

In addition to estimating errors using Kullback–Leibler divergence, we have used the Fisher information metric to estimate the distance between the model using the true parameter value and the estimated value. For each estimator, and for the same values of ρ as in sections 5.2.2 and 5.2.3, 10 000 estimated values were found and used to calculate the Fisher information metric. The averaged results multiplied by a factor of 1000 are shown in table 5.2.

Again, the PC prior with $\lambda = 5.360$ and the fiducials give very low errors for small values of ρ, but shows a large increase in error as ρ increases towards 1. The only difference from section 5.2.3 is that the fiducial estimator using the full empirical correlation and

the Fisher information as loss function does not behave differently from the other fiducial estimators.

The maximum likelihood estimator and empirical correlation with known mean do still give quite large errors. However, the latter is not double the size of the former, as was the case when using Kullback–Leibler divergence. Rather, the MLE has a larger error for small values of ρ, whereas the empirical correlation gives a larger error for values of ρ closer to 1. Also, the error for the empirical correlation seems to be more stable across different values of ρ.

The priors (excluding the PC prior with $\lambda = 5.360$) show a similar behaviour as in section 5.2.3, with the mean square error and Kullback–Leibler divergence as loss functions giving more or less the same results, and the Fisher information metric giving larger errors. However, if we again choose the estimator with the smallest maximum value over all values of ρ tested, we would now choose the flat prior with the Fisher information metric as loss function.

Table 5.3: The error using the **Fisher information metric as measure of error** for the point estimators are shown. The point estimators were calculated using samples of size 3. The error was found by sampling 10 000 values of each estimator, and calculating the Kullback–Leibler divergence between the model using the estimator and the true value. The values shown are the average of these *errors multiplied by a factor of 1000*.

Estimator		**ρ value**				
Prior	*Loss function*	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*
Flat	MSE	406	416	450	517	655
Jeffreys	MSE	605	619	619	632	618
PC (0.0818)	MSE	544	546	564	605	633
PC (1.249)	MSE	396	427	482	570	687
PC (5.360)	MSE	116	264	471	779	1427
Flat	KL div.	410	419	455	528	647
Jeffreys	KL div.	611	605	614	633	619
PC (0.0818)	KL div.	532	544	579	607	627
PC (1.249)	KL div.	393	426	484	565	699
PC (5.360)	KL div.	116	265	472	779	1423
Flat	FI metric	561	559	598	608	597
Jeffreys	FI metric	831	837	818	794	700
PC (0.0818)	FI metric	733	747	752	746	670
PC (1.249)	FI metric	535	564	623	674	678
PC (5.360)	FI metric	107	280	501	817	1469
Fiducial						
Known mean?	*Loss function*					
No	MSE	354	822	508	633	1289
No	KL div.	242	274	417	703	1367
No	FI metric	473	459	541	695	1089
Yes	MSE	248	377	340	728	1339
Yes	KL div.	180	237	416	712	1387
Yes	FI metric	257	292	384	588	1097
Other approaches						
Maximum likelihood		1015	999	963	892	776
Empirical correlation w/known mean		785	793	805	849	891

Chapter 6

Discussion

In this thesis we have seen that different approaches to statistical analysis might be taken, and that this to some extent affects the point and interval estimators that we use to make statements about unknown parameters of our model. We have introduced some ways to perform parameter estimation for the frequentist and Bayesian schools, and also looked briefly at fiducial analysis.

With a particular focus on Bayesian methods, the choice of prior distribution for our parameter was discussed, and the need for default objective priors was explained. After a discussion around objectivity in Bayesian analysis, three frameworks for finding an objective prior were presented. Jeffreys prior focuses on invariance, and works well in one-parameter cases (but is not recommended for any multi-parameter distributions). Reference priors might be seen as an extension to Jeffreys prior, with a more complete mathematical foundation. The concept of Penalised Complexity priors is a more pragmatic approach, in which parsimonious models and computational tractability are central.

In order to apply the theory, the bivariate normal distribution with known zero means and unit variances was chosen. The priors and estimators introduced in previous chapters were then applied to this one-parameter model. Simulations were then performed to investigate the performance of the estimators for the chosen model.

The simulations show that there are huge differences in the behaviour and performance of the estimators, but also that some estimators perform similarly and are essentially indistinguishable based on the test performed. Some estimators were very likely to give estimates close to 0, regardless of the true value of ρ, whereas others were much closer to the true parameter value when ρ approached ± 1.

It seems clear that a tradeoff exists between low bias and low variance, both of which are desirable features of an estimator. Also, in order to choose an estimator that works well for most or all possible values of ρ, it is important to consider the behaviour of the estimator for ρ values that are both close to 0 and close to the extremes, as well as the

possible values in between. Hence, an estimator that works well either for large or small values of ρ, might not be preferred, if it works badly for the other case. In the end, the best choice seems to be an estimator that manages to strike a balance between bias and variance, and that works well enough for all values of ρ.

Moving forward, there are a lot of things one might want to investigate further. First of all, running larger experiments with more iterations would provide more accurate results, and perhaps indicate which effects, if any, were merely caused by noise. In particular, increasing the sample size used to approximate the fiducials would, most likely, improve the performance of estimators based on this approach. Indeed, the fiducials proved almost useless in our simulations, and it would be interesting to look into why this was the case.

With regards to the Bayes estimators, one could use the square root of the Kullback–Leibler divergence and the Fisher information metric, or the square of these. It would be interesting to see how the results from such loss functions would differ from the ones obtained in this thesis.

Even though it is not part of the definition of the PC priors, one could try to place a prior distribution on the user-defined scaling λ (e.g. $1/\lambda$, as suggested in section 3.5). This would at least eliminate the need to define a tail event and its weight, in order to get a numerical value for λ.

Fosdick and Raftery, 2012 uses a prior which they call the *arc–sine prior*, which does give quite good results for the model we have used in this thesis. They do not, however, use Bayes estimator with Kullback–Leibler divergence and the Fisher information metric as loss functions. Comparing this prior using these Bayes estimators to the Jeffreys, flat and PC priors that we have looked at here seems reasonable.

One further possible estimation procedure comes from viewing the problem as one of regression. This approach was taken in Castro and Vidal, 2019, and adapting this to our model could provide us with yet another possible estimator.

Lastly, when considering interval estimation using posteriors and the fiducials, future studies should look into the lengths of the intervals produced (which we would like to be as small as possible). Also, estimating the power of hypothesis tests when using the different estimators, as is done in Fosdick and Raftery, 2012, would be appropriate.

Bibliography

[1] C. Atkinson and A. F. S. Mitchell. "Rao's Distance Measure". In: *Sankhyā: The Indian Journal of Statistics, Series A* 43.3 (1981). https://www.jstor.org/stable/25050283, pp. 345–365.

[2] J. O. Berger. "The Case for Objective Bayesian Analysis". In: *Bayesian Analysis* 1.3 (2006). https://doi.org/10.1214/06-BA115, pp. 385–402.

[3] J. O. Berger, J. M. Bernardo, and D. Sun. "Overall Objective Priors". In: *Bayesian Analysis* 10.1 (2015). https://doi.org/10.1214/14-BA915, pp. 189–221.

[4] J. O. Berger, J. M. Bernardo, and D. Sun. "The Formal Definition of Reference Priors". In: *The Annals of Statistics* 37.2 (2009). https://doi.org/10.1214/07-AOS587, pp. 905–938.

[5] J. O. Berger and D. Sun. "Objective priors for the bivariate normal model". In: *The Annals of Statistics* 36.2 (2008). https://doi.org/10.1214/07-AOS501, pp. 963–982.

[6] J. M. Bernardo. "Reference Posterior Distributions for Bayesian Inference". In: *Journal of the Royal Statistical Society. Series B* 41.2 (1979). https://doi.org/10.1111/j.2517-6161.1979.tb01066.x, pp. 113–147.

[7] C. Bioche and P. Druilhet. "Approximation of improper priors". In: *Bernoulli* 22.3 (2016). https://doi.org/10.3150/15-BEJ708, pp. 1709–1728.

[8] G. Casella and R. L. Berger. *Statistical Inference*. 2nd. Brooks/Cole, 2002.

[9] M. de Castro and I. Vidal. "Bayesian inference in measurement error models from objective priors for the bivariate normal distribution". In: *Stat Papers* 60 (2019). https://doi.org/10.1007/s00362-016-0863-7, pp. 1059–1078.

[10] G. Consonni, D. Fouskakis, B. Liseo, and I. Ntzoufras. "Prior Distributions for Objective Bayesian Analysis". In: *Bayesian Analysis* 13.2 (2018). https://doi.org/10.1214/18-BA1103, pp. 627–679.

[11] B. K. Fosdick and A. E. Raftery. "Estimating the Correlation in Bivariate Normal Data With Known Variances and Small Sample Sizes". In: *The American Statistician* 66.1 (2012). https://doi.org/10.1080/00031305.2012.676329, pp. 34–41.

[12] D. Fraser, G. Monette, and K.-W. Ng. "Marginalization, likelihood and structured models". In: *Multivariate Analysis* 25.1 (1985), pp. 209–217.

[13] M. Ghosh, B. Mukherjee, U. Santra, and D. Kim. "Bayesian and likelihood-based inference for the bivariate normal correlation coefficient". In: *Journal of Statistical Planning and Inference* 140.6 (2010). https://doi.org/, pp. 1410–1416.

[14] L. Huisman. "Infinitesimal Distributions, Improper Priors and Bayesian Inference". In: *Sankhyā: The Indian Journal of Statistics* 78 (2016). https://doi.org/10.1007/s13171-016-0092-0, pp. 324–346.

[15] W. H. Jefferys and J. O. Berger. "Ockham's Razor and Bayesian Analysis". In: *American Scientist* 80.1 (1992). https://www.jstor.org/stable/29774559, pp. 64–72.

[16] H. Jeffreys. "An invariant form for the prior probability in estimation problems". In: *Proc. R. Soc. Lond.* 186.1007 (1997). https://doi.org/10.1098/rspa.1946.0056, pp. 453–461.

[17] D. H. Kim, S. G. Kang, and W. D. Lee. "Noninformative priors for the common mean in the bivariate normal distribution". In: *Journal of the Korean Statistical Society* 38.2 (2009). https://doi.org/, pp. 167–174.

[18] D. E. Knuth. "Two Notes on Notation". In: *The American Mathematical Monthly* 99.5 (1992). https://doi.org/10.2307/2325085, pp. 403–422.

[19] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *Annals of Mathematical Statistics* 22.1 (1951). https://doi.org/10.1214/aoms/1177729694, pp. 79–86.

[20] E. of Mathematics. Absolutely continuous measures. https://encyclopediaofmath.org/wiki/Absolutely_continuous_measures [Accessed: July 8, 2020].

[21] E. of Mathematics. Bernstein-von Mises theorem. https://encyclopediaofmath.org/wiki/Bernstein-von_Mises_theorem [Accessed: August 23, 2020].

[22] E. of Mathematics. Degenerate distribution. https://encyclopediaofmath.org/wiki/Degenerate_distribution [Accessed: July 1, 2020].

[23] E. of Mathematics. Jensen inequality. https://encyclopediaofmath.org/wiki/Jensen_inequality [Accessed: August 10, 2020].

[24] W. MathWorld. Cubic Formula. https://mathworld.wolfram.com/CubicFormula.html [Accessed: July 3, 2020].

[25] W. MathWorld. Incomplete Gamma Function. https://mathworld.wolfram.com/IncompleteGammaFunction.html [Accessed: July 24, 2020].

[26] K. Miura. "An Introduction to Maximum Likelihood Estimation and Information Geometry". In: *Interdisciplinary Information Sciences* 17.3 (2011). https://doi.org/10.4036/iis.2011.155, pp. 155–174.

[27] F. Nielsen. *An elementary introduction to information geometry*. https://arxiv.org/abs/1808.08271. 2018.

[28] C. R. Rao. "*Fisher-Rao metric*". In: *Scholarpedia* 4.2 (2009). http://doi.org/10.4249/scholarpedia.7085 [revision 91265], p. 7085.

[29] K. Rottmann. *Matematisk formelsamling*. 1st. Spektrum forlag, 2014/2003.

[30] H. Rue. R-INLA. https://github.com/hrue/r-inla/blob/devel/inlaprog/src/pc-priors.c [Accessed: August 18, 2020].

[31] D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors". In: *Statistical Science* 32.1 (2017). https://doi.org/10.1214/16-STS576, pp. 1–28.

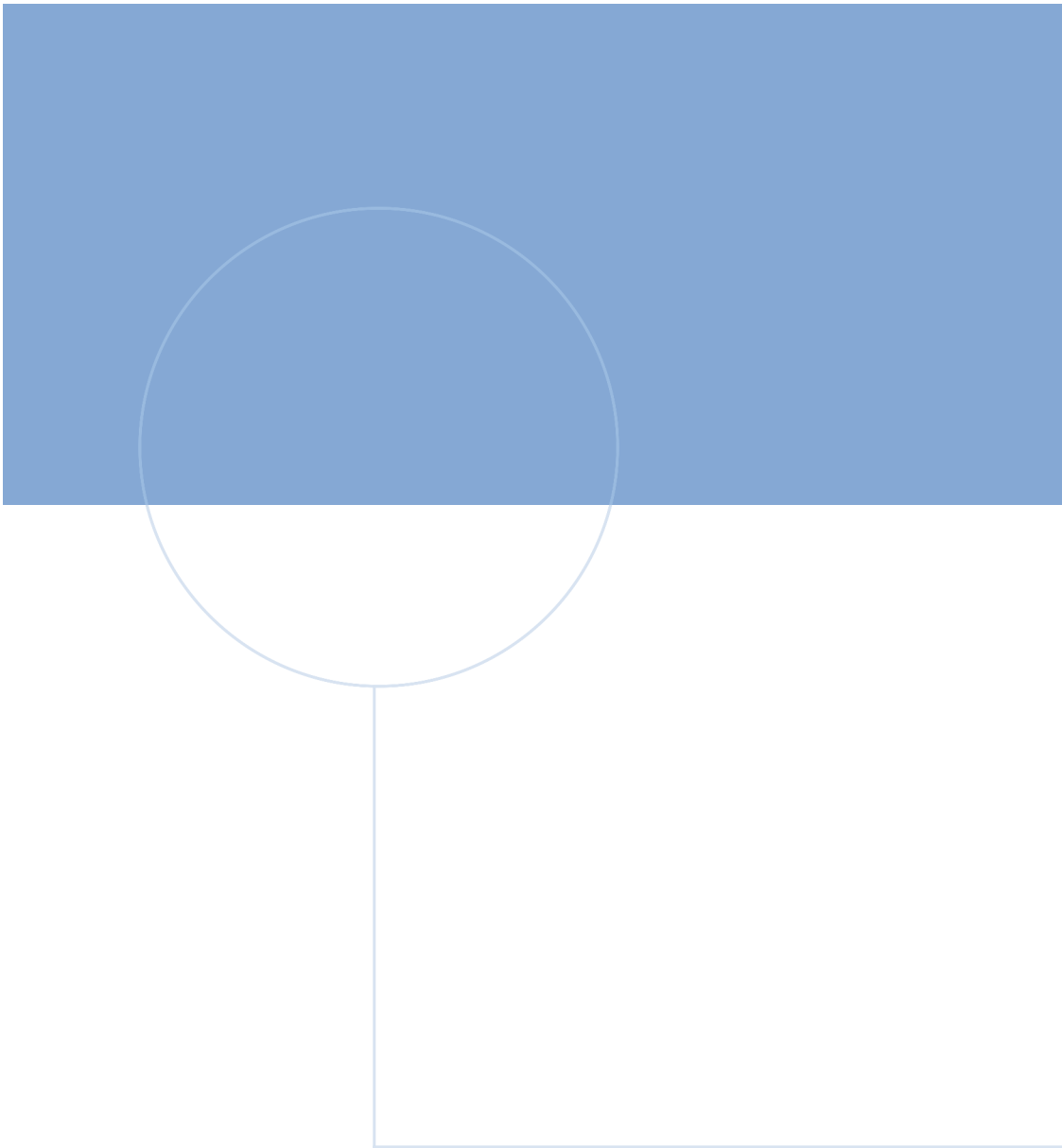
[32] G. Taraldsen and B. H. Lindqvist. "Conditional fiducial models". In: *Journal of Statistical Planning and Inference* 195 (2018). https://doi.org/10.1016/j.jspi.2017.09.007, pp. 141–152.

[33] G. Taraldsen and B. H. Lindqvist. "Fiducial theory and optimal inference". In: *Annals of Statistics* 41.1 (2013). https://doi.org/10.1214/13-AOS1083, pp. 323–341.

[34] G. Taraldsen, J. Tufto, and B. H. Lindqvist. *Statistics with improper posteriors*. https://arxiv.org/abs/1812.01314. 2018.

[35] S. Taylor. "Clustering Financial Return Distributions Using the Fisher Information Metric". In: *Entropy* 21.2 (2019). https://doi.org/10.3390/e21020110, p. 110.

[36] Wikipedia. Prior probability. https://en.wikipedia.org/wiki/Prior_probability [Accessed: August 23, 2020].



NTNU

Norwegian University of
Science and Technology