Penalising Model Complexity

Ph.D. student: Diego Battagliese

Advisor: Prof. Brunero Liseo

XXXII Cycle

MEMOTEF Department

Sapienza University of Rome

October 2019

Contents

Preface

My interest for the Objective Bayesian analysis starts during the first year of my Ph.D., when my advisor asked me to go whit him to the Objective Bayes Methodology conference, that took place in Austin in 2017. Initially, I felt demotivated as the level of the conference was too difficult for me. Two years later, I still remember the words of one of the speakers of the conference, he said to me that at the first year of his Ph.D. he had the same feelings. During the conference, my advisor gave me some ideas to be developed, then the following thesis ensues. To be honest, the words of the speaker I met in Austin were absolutely true. In fact, just after two years I gave a poster in the next Objective Bayes Methodology conference held in Coventry in July 2019. In just two years I really learnt a lot of things and, in this regard, I would like to thank my advisor for the many concepts he taught me and for the way to think of Statistics. I want to give most of the credit to him, as, in my opinion, he has been fundamental in the course of my Ph.D., and not only for the help he gave me from the academic point of view, but also for his human aspects.
I would like to thank also two other people, Cristiano Villa and Clara Grazian. Cristiano was my supervisor during the six months I spent at the University of Kent in Canterbury from September 2018 to April 2019. Clara was one of my collaborators and she has been very close to me, even if not physically, as she was very far away, but especially during my visiting in Canterbury we had a lot of Skype calls. They both helped me a lot to improve the results I already had when I arrived in Canterbury. Cristiano gave me many notions and suggestions, and not only from an academic point of view, while Clara helped me also with all the computational issues I had. The third chapter of this thesis is a joint work also with them.

Chapter 1

Introduction

In the present thesis, we describe and explore a new method of constructing prior distributions on the additional model components that build up a more flexible model starting from a base model which would not include those components.
First of all, we want to briefly sketch the idea of the PC prior. Suppose to have a certain model that could be rendered more flexible and richer by introducing an extra-component. As a toy example, suppose one wants to make more robust the Gaussian distribution by allowing an additional parameter to control for kurtosis. In this specific case, we are dealing with a Student-t distribution, so a prior for the degrees of freedom ν needs to be specified. The concept of PC prior arises from the distance between the simpler model and the more flexible one. The PC prior is closely related to the Kullback-Leibler divergence (KLD), that is meant to penalise deviations from the base model. Afterwards, the KLD is transformed into a more interpretable distance scale, and such a distance is supposed to be exponentially distributed. The selection of the rate parameter of the exponential distribution is left to the user belief and, finally, by performing a change of variable from the distance scale to the parameter of interest the PC prior is obtained.
As pointed out by **?**, the prior is formulated by means of the penalisation of the distance between two nested models and through the injection of a user-sensible perception about a tail event. The introduction of the belief about the tail event seems to lead to a subjective prior, even though we could control the information to be introduced in the prior by selecting a value for the unique parameter of the penalised complexity prior, PC prior hereafter, in order make it as uninformative as possible. As we will see later on, the distance between the two models is penalised by assigning to it an exponential distribution whose rate parameter constitutes the shrinking parameter that establish the informativeness of the PC prior. Here, we are not claiming that the PC priors are objective, even though they could be more easily seen as weakly informative priors, but we are just saying that they are objective in the sense that they are principled, say they are constructed on the basis of a well known machinery.
Among the main advantages of these priors we can mention the invariance to reparameterisations, that makes these priors close to Jeffreys' priors, they invoke the Occam's razor principle, preferring simplicity over complexity and have good robustness properties. In addition, the connection with the Jeffreys' prior

is not only given by the invariance to reparameterisation, but we will see that for certain values of the shrinking parameter the PC prior approaches the Jeffreys' one. This made us of thinking of PC priors in an objective fashion. Another relevant feature is that in many situations PC priors are invariant to the other parameters lying both in the base model and in the complex model, and in practice only the additional model component matters. This is a direct consequence of the Kullback-Leibler divergence we use to derive the prior as, in most of the cases, it does not depend on other parameters that are not of interest. An example of such an invariance is related to the location-scale models. The latter property is very important because it allows us to derive a separable prior on the additional model component without the need of building a joint prior on the composite parameter vector.
? focus on the class of hierarchical models, where an unobserved latent structure is added by means of a set of model components. Such perspective is the building block which the penalised complexity priors are based on, since it requires the definition of a base model. The choice of the base model is not univocal as it demands the user to define the simplest model for the problem at hand. Nonetheless, the prior mass at the base model should be not zero in order to avoid the prior to overfit. The prior does not overfit when the prior mass at the base model is non zero, otherwise we would incur in posteriors that give no evidence to the base model, even if it is the true one, and as a consequence we would not be able to understand if the evidence for the flexible model come from the data or it is just induced by the prior distribution.
Here, we consider as a base model the model where a particular value of the flexibility parameter makes the component to disappear in the base model. The concept of base model finds natural connection to the hypothesis testing problem. Indeed, when introducing an additive model component we just wonder if such a component should or should not be included in the model. This means that PC priors can play an important role in the Bayesian hypothesis testing, especially when one wants to use objective priors like Jeffreys' priors that most of the times are not integrable and then could lead to the Jeffreys-Lindley's paradox. In fact, PC priors are proper and this admits their use in testing problems.
It is well known that the employment of improper priors is not recommended at all in the Bayesian hypothesis testing context, so, as an alternative, we explore the behaviour of such PC priors in this kind of problem. Nevertheless, the base model, in the viewpoint of the null hypothesis, is not uniquely determined. Users could define different base models on the basis of what they think of the conservative hypothesis. For the sake of clarity, we recall that throughout this thesis we limit ourselves to cases in which the base model ousts the additive component.
In spite of the fact that PC priors are not conceived for testing problems, we don't see how they could not be used in such a framework. To motivate our feeling let's consider the spike-and-slab priors and their closeness to PC priors. Spike-and-slab priors (?, ?) have been mostly used in high-dimensional regression, where they aim to reduce the number of the predictors, a similar purpose as that of the Bayesian Lasso. These priors are made up of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike) as they avoid overfitting and accomplish both prediction and model selection. These priors are two-point mixture and are computationally unpleasant. However, different

approaches have been proposed, as in ?, to overcome the computational issues. We are referring to the Global-local shrinkage priors that are constructed as a scale mixture of normal distributions and the choice of the priors on the variance is made so that to have a similarity to spike-and-slab priors. Global-local priors are not build up with atoms, nevertheless they share the same advantages of spike-and-slab priors, without computational issues. So, why not to use PC priors?!

PC priors are a very flexible tool, because we can easily regulate the scale of these priors by means of a rate parameter, and this could be determinant when we use such PC priors in the Bayesian hypothesis testing context. In addition, we are able to adjust the prior probability on the null and alternative hypotheses simply letting them to depend on the variance of the PC prior that in turn is a function of the rate parameter. We want to remark that this attempt of resolution of the Jeffreys-Lindley's paradox would not take place here, bacause PC priors are proper even though they can approach the Jeffreys' prior for a limiting choice of the parameter controlling the shrinkage. ? proposed to fix the Jeffreys-Lyndley's paradox in the Bayesian sense when using improper, or too diffuse, priors by weighting the prior densities of the two hypotheses through the scale factor of the prior for the parameter of the alternative model in order to calibrate the displacement between a point-null hypothesis versus a diffuse one. It jumps to eyes that there is some theoretical defect due to the fact that the weight on the null hypothesis depends on the prior for the parameter of the alternative model. The correction based on the scale parameter is meant to avoid to incur in the Jeffreys-Lindley's paradox when the prior on the alternative model is improper and the null is pointwise, therefore does not apply to PC priors even though nothing prevent to adoperate such a correction . Calibrating the weight on the hypotheses can be determining in the testing procedure using PC priors as they are higly flexible, but attention must be paid in the selection of the shrinkage parameter because limiting choices of this parameter could end up in completely misleading or meaningless conclusions. Anyhow, what we require to prevent the Bayes Factor to be inconstistent, as in ?, is that the prior does not overfit in the sense of having zero mass at the base model. Nonetheless, it is remarkable to point out that a limiting shrinkage produces Bayes factor giving no evidence for any of the hypotheses and therefore it is nonsense.

It is usual to uphold that priors constructed for predictive models have poor performances when used in Bayesian hypothesis testing. Let us consider for instance the advent of non-local priors (?), they work wonderfully in testing problem as they give approximately zero mass in a neighbourhood of the base model ensuring a balanced convergence of the Bayes factor towards the true models, but not the same is achieved in terms of predictive properties, or, on the opposite hand, the Jeffreys' prior that is practically unbeatable in prediction, but quite useless in hypothesis testing since it is improper most of the times. Anyhow, there is no reason to think that well-behaved priors in prediction should be inadequate in testing. ? in making some comments about the paper from ? have advanced the idea of look at PC priors in a Bayesian decision theory point of view. The principled construction of a PC prior seems to have no decisional aspects to be investigated, even though some principles could be relaxed. In our opinion, PC priors can play a keyrole in model selection and hypothesis testing and they can also be engaged in a decision theory framework.

As we were stating before, PC priors are conceived for hierarchical models and their particularly flexible robustness is an higly desirable feature.
The outline of the remaining chapters is as follows. In Chapter 2, we briefly introduce some concepts on objective Bayesian analysis and the prior distributions involved throughout the thesis. In Chapter 3, we introduce a general strategy to derive a penalised complexity prior for the dependence in a copula setting. The PC prior for the Gaussian copula model is derived and inference and Bayesian hypothesis testing are performed. Finally, the exchangeable model is addressed. In Chapter 4, we deal with models that depart from the Gaussian distribution. The PC prior for the shape parameter of the skew-normal distribution is numerically derived. In addition, inference and Bayesian hypothesis test are performed. Finally, the Student-t case is addressed by exploiting a new convenient formulation of the Kullback-Leibler divergence. In chapter 5, we propose some new methodologies to extend the univariate PC prior to the multivariate case.

Chapter 2

Objective Bayesian Analysis

2.1 What does the term *non-informative* means?

We do not wish to debate the etymological aspects of the term, but we deem worthwhile to spend a few words to clarify the meaning of *non-informative*, when it is referred to prior distributions. For an interesting discussion on the matter, refer to ?.
? pointed out that "there is no prior representing ignorance". Every prior distribution contains some amount of information (although sometimes minimal), in the sense that it depends on the model that has been chosen. In fact, it is commonly agreed that *objectivity* is intended from the moment that the model has been selected in order to represent the quantity of interest. Therefore, when we refer to a prior distribution representing *ignorance*, it has to be understood in the aforementioned sense.
Many terms have been used to label this type of distributions: *conventional*, *default*, *flat*, *formal*, *neutral*, *non-subjective* and *objective* (?). Independently on the term we decide to adopt, when the prior is not elicited by means of expert informations, the contribution of the data to the posterior distribution must be as large as possible. In other words, the data should dominate the posterior.

2.2 Overview of some Objective Priors

2.2.1 A few words on Improper Priors

Objective approach leads in many circumstances to improper priors, in the sense that these priors do not integrate (or sum, in the discrete case) to one. This happens because, as we want to represent as less knowledge as possible about the parameter value, the parametric space is often unbounded.
There are cases where objective priors are proper. For instance, a commonly accepted objective prior distribution for the parameter $\theta \in (0,1)$ of a binomial distribution, representing the probability of success, is $\pi(\theta) = \text{Be}(1/2, 1/2)$, where Be stands for the Beta density; this latter is the Jeffreys' prior for θ. However, a bounded parameter space is not *per se* a sufficient condition for having a proper objective prior. As an example, consider a Negative Binomial distribution with parameters (r, p), where $r > 0$ and $p \in (0,1)$ and where the

usually recommended objective prior for p is $\pi(p) \propto p^{-1}(1-p)^{-1/2}$. This prior, although the parameter space is bounded, turns out to be improper. Another example is to consider the Gaussian copula with correlation parameter ρ. In this case the Jeffreys' prior for ρ is improper even if the correlation parameter space is bounded. Finally, there are scenarios where for an unbounded parameter space it is possible to have proper objective priors, as for the case of the ratio of two multinomial parameters, where the parameter space is $(0, \infty)$, see **?**. Another rare case of proper objective prior over an unbounded parameter space is the Jeffreys' prior for the skewness parameter of the skew-normal distribution, given by **?** and approximated by **?** by means of a $t(0, \pi^2/4; 1/2)$.
Given that inference depends on the posterior, improper priors can be used in practice as long as the posterior is proper, even though they cannot be used in Bayesian hypothesis testing since such priors are defined only up to a constant multiple, and the Bayes factor is itself a multiple of this arbitrary constant. However, improper priors are not probability distributions, and they simply represent positive functions. Their utilisation is just a technical *device* to be used in the Bayes theorem in order to obtain proper posterior distributions (**?**), but it is obvious that, conceptually, Bayes theorem no longer applies.
? give a justification on the adoption of improper priors; if an improper prior $\pi(\theta)$ is adopted, then Bayes theorem does not apply and its use has to be justified, even when the posterior is a proper density. **?** show that the posterior $\pi(\theta|x)$ is a suitable limit of posterior distributions obtained from proper prior distributions. Consider a model $\mathcal{M} = \{p(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$, a strictly positive continuous function $\pi(\theta)$ and an approximating compact sequence $\{\Theta_i\}_{i=1}^{\infty}$ of parameter spaces. The corresponding sequence of posterior distributions $\{\pi_i(\theta|x)\}_{i=1}^{\infty}$, with $\pi_i(\theta|x) \propto p(x|\theta)\pi_i(\theta)$ and where $\pi_i(\theta)$ is a proper prior distribution, is said to be expected logarithmically convergent to the formal posterior $\pi(\theta|x)$ if

$$\lim_{i\to\infty} \int_{\mathcal{X}} \mathrm{KLD}(\pi(\cdot|x) \,\|\, \pi_i(\cdot|x)) p_i(x) dx = 0, \tag{2.1}$$

where $p_i(x) = \int_{\Theta_i} p(x|\theta)\pi_i(\theta)d\theta$. The conclusion is that a prior distribution satisfying condition in (**??**) will yield a posterior distribution that, on average over x, is a good approximation of the proper posterior distribution that would result from restriction to a large compact subset of the parameter space.

2.2.2 Jeffreys' Prior

The main criticism about uniform prior distributions is that they do not represent ignorance. In fact, knowing nothing about θ and knowing that it can take any value with the same probability are two well distinct facts. Mainly, the above criticism to uniform priors has come from the fact that, in general, they are not invariant under one-to-one reparameterizations, and many bayesians consider this property as a must for an objective prior; for a rigorous discussion see **?**, **?** and **?**. In particular, Jaynes asserts that the way a model is parameterized involves subjectivity; as a consequence, a prior distribution influenced by this subjective choice cannot be considered entirely objective. In addition, the state of knowledge about a model does not change by simply rearranging the parameters. Let us better understand the meaning of invariance under one-to-one reparameterizations. Consider a statistical model $p(\cdot|\theta)$ with the prior $\pi(\theta) \propto 1$

being a uniform prior. If we have no knowledge about θ, we will not have any knowledge about $1/\theta$ either. Suppose $\phi = g(\theta) = 1/\theta$, therefore by applying the change-of-variable formula on the one-to-one transformation $g(\theta) = 1/\theta$, we have

$$\pi(\phi) = 1 \cdot \left| \frac{\partial g^{-1}(\phi)}{\partial \phi} \right| = \frac{1}{\phi^2}, \tag{2.2}$$

which is no longer uniform.
In designing an objective approach to derive priors, **?** stressed the importance to have a resulting distribution that is invariant under any one-to-one (differentiable) transformation. Then, he based his method on the Fisher information (**?**), $I(\theta)$, that is a quantification of the amount of information about the unknown parameter θ that is expressed by the model, and that it is invariant under these kind of transformations. The Fisher information is defined as

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right], \tag{2.3}$$

where $\mathbb{E}_\theta$ is the expectation with respect to the model $f(x|\theta)$ and $\log f(x|\theta)$ is the log-likelihood function. In particular, if $\xi = h(\theta)$ and θ are two parameterizations of the same estimation or decision problem, and h is a continuously differentiable function of θ, **?** show that

$$I(\theta) = I(\xi) \cdot [h'(\theta)]^2, \tag{2.4}$$

where $h'(\theta)$ represents the derivative of $h(\theta)$ with respect to θ. It is important to realize that equation (**??**) links the Fisher information of the two parameterizations. Thus, by taking the square root of equation (**??**), we have

$$I^{1/2}(\theta) = I^{1/2}(\xi) \cdot |h'(\theta)|. \tag{2.5}$$

Therefore, the Jeffreys' prior for θ will be linked to the Jeffreys' prior for ξ in the following manner

$$\pi^J(\theta) \propto I^{1/2}(\theta) = \pi^J(\xi) \cdot |h'(\theta)|, \tag{2.6}$$

where on the right-hand-side it is possible to recognize the change-of-variable formula, showing the Jeffreys' prior invariance property.
As an illustration, the Jeffreys' prior for the parameter $\theta \in (0, 1)$ of a Binomial distribution with known n is given by $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, that is a Beta distribution with both shape parameters equal to $1/2$. Moreover, if we consider a Normal distribution with unknown mean μ and known variance σ^2, it can be shown that $\pi^J(\mu) \propto 1$, showing that the uniform prior can still be a valid objective prior, in the sense that it complies with the desiderata of being invariant under one-to-one reparameterizations.
An important limit of this kind of prior, noticed by Jeffreys himself, is that in general it does not lead to acceptable results when applied to a vector of parameters. Let us consider a density function $f(x|\theta)$, where $\theta = [\theta_1, \ldots, \theta_d]^T$ is a vector of d parameters, the Fisher information matrix for this vector of parameters is given by

$$I_{i,j}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\theta) \right]. \tag{2.7}$$

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

$$\pi(\theta) \propto \sqrt{\det(I(\theta))}. \tag{2.8}$$

The prior for the unknown parameters (μ, σ) of a Normal distribution, obtained according to the Jeffreys' rule, is $\pi(\mu, \sigma) \propto 1/\sigma^2$. It has been shown that this prior has poor convergence performance (**?**). To overcome this weakness, Jeffreys suggested to consider the two parameters as independent *a priori*, $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma) \propto 1/\sigma$; this prior has desirable properties. To distinguish between the two priors, we call the first one as *Jeffreys' rule prior*, as it has been obtained by direct application of the Jeffreys' method, whilst the second, assuming independence a priori, is called *Independence Jeffreys' prior*.

2.2.3 Intrinsic Prior

Intrinsic prior distributions have been introduced by **?** to provide a proper Bayesian interpretation for intrinsic Bayes factors.
Suppose we have two models M_j and M_i, then the Bayes factor is defined as

$$B_{ji} = \frac{m_j(x)}{m_i(x)} = \frac{\int f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}{\int f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}, \tag{2.9}$$

where $m_i(x)$ is the marginal density of X under M_i.
Computing B_{ji} requires to specify $\pi_i(\theta_i)$ and $\pi_j(\theta_j)$. In Bayesian analysis it is usual to use noninformative priors π_i^N, where the superscript N stands for noninformative.
Using any of π_i^N in (**??**) would yield

$$B_{ji}^N = \frac{m_j^N(x)}{m_i^N(x)} = \frac{\int f_j(x|\theta_j)\pi_j^N(\theta_j)d\theta_j}{\int f_i(x|\theta_i)\pi_i^N(\theta_i)d\theta_i}. \tag{2.10}$$

The problem here is that the π_i^N are typically improper, and hence defined only up to arbitrary constants c_i. Therefore B_{ji} is defined only up to $\frac{c_j}{c_i}$, which is itself arbitrary.
The solution to this kind of problem is to use part of the data as a training sample. Let $x(\ell)$ denote the part of the data to be used as a training sample, and $x(-\ell)$ represent the remainder of the data. So, $x(\ell)$ will be used to convert the $\pi_i^N(\theta_i)$ to proper posterior distributions

$$\pi_i^N(\theta_i|x(\ell)) = \frac{f_i(x(\ell)|\theta_i)\pi_i^N(\theta_i)}{m_i^N(x(\ell))}, \tag{2.11}$$

where $f_i(x(\ell)|\theta_i)$ is the marginal density of $X(\ell)$ under M_i and

$$m_i^N(x(\ell)) = \int f_i(x(\ell)|\theta_i)\pi_i^N(\theta_i)d\theta_i. \tag{2.12}$$

Then, the idea is to compute the Bayes factor with the remainder of the data, $x(-\ell)$, using the $\pi_i^N(\theta_i|x(\ell))$ as priors. The result is

$$\begin{aligned} B_{ji}(\ell) &= \frac{\int f_j(x(-\ell)|\theta_j, x(\ell))\pi_j^N(\theta_j|x(\ell))d\theta_j}{\int f_i(x(-\ell)|\theta_i, x(\ell))\pi_i^N(\theta_i|x(\ell))d\theta_i} \\ &= B_{ji}^N \cdot B_{ij}^N(x(\ell)), \end{aligned} \tag{2.13}$$

where

$$B_{ij}^N(x(\ell)) = \frac{m_i^N(x(\ell))}{m_j^N(x(\ell))}. \tag{2.14}$$

We may notice that (??) removes the arbitrariness in the choice of constant multiples of the π_i^N. The arbitrary ratio c_j/c_i that multiplies B_{ji}^N would be cancelled by the ratio c_i/c_j that would then multiply $B_{ij}^N(x(\ell))$. Note also that the training sample procedure make sense only if the $m_i^N(x(\ell))$ in (??) is finite. ? showed that the intrinsic prior for the arithmetic intrinsic Bayes factor (?) coincides with the Expected-posterior prior for nested models. The Expected posterior prior is defined as follows

$$\pi^{EPP}(\theta_i) = \int \pi_i^N(\theta_i|y^*)m^*(y^*)dy^*, \tag{2.15}$$

where $\pi_i^N(\theta_i|y^*) \propto f(y^*|\theta_i)\pi^N(\theta_i)$ is the posterior distribution of θ_i under model M_i conditionally on the imaginary data y^* for the baseline prior $\pi^N(\theta_i)$, and

$$m^*(y^*) = \int f(y^*|\theta_0)\pi^N(\theta_0)d\theta_0, \tag{2.16}$$

where θ_0 is the parameter of the simplest model M_0, i.e. the model which is nested in each of the remaining models.
So, the Expected-posterior prior for the parameter under a given model is the expectation of the posterior distribution given imaginary observations y^*, where the expectation is taken with respect to a suitable probability measure $m^*(y^*)$ under reference model M_0. Note that the posterior distribution is computed starting from a typically improper prior.

2.3 Penalised Complexity Prior

Penalised complexity priors (PC priors) have been proposed by ?. The construction of a PC prior is principled, but we cannot say they are default prior. In particular they are weakly informative priors, in the sense that a litte user-defined amount of information need to be introduced into the prior.
? defined four basic principles behind the construction of a PC prior for the generic parameter ξ.

- **Occam's Razor.** Simpler model formulation is preferred until there is enough support for a more complex model. The PC prior is meant to penalise deviations from a base model. Here, one could debate the choice of the base model, for instance choosing as a base model an arbitrary value of ξ in the parameter space Ξ is plausible. We want to remark that the base model should be viewed as the model where the additional component ξ is absent, even though nothing prevent us to choose a different base model coming from our belief about it.
  Here, it is worthwhile to make a digression on the base model. There is a twofold way to conceive the base model. The first one is based on the assumption that the model is built up by building blocks that make the model more and more flexible. The second one consists in choosing as a base model the one referred to a particular value of the parameter of interest ξ. In the latter case, the parameter ξ may appear in both the simpler

and flexible models. We want to remark that both the points of view are correct, but throughout the present work we will refer to the base model as the model where we have a parameter less in the probability density function, so we use the approach of the former definition. In practice, there is a specific value of ξ that makes the component to disappear in the base model.

- **Measure of Complexity.** The increased complexity is measured by the Kullback-Leibler divergence

$$\mathrm{KLD}(f\|g) = \int_{\mathcal{X}} f(x;\xi)\log\left(\frac{f(x;\xi)}{g(x)}\right)dx, \tag{2.17}$$

where $g(x) = f(x;\xi = \xi_0)$, with ξ_0 being a particular of ξ that renders the model simpler.
The Kullback-Leibler divergence is not a metric and to render it a more interpretable distance scale it is transformed into $d(f\|g) = \sqrt{2KLD(f\|g)}$.

- **Constant Rate Penalisation.** A constant decay-rate r implies an exponential prior distribution on the distance scale

$$\frac{\pi_d(d+\nu)}{\pi_d(d)} = r^{\nu}, \qquad d, \nu \geq 0 \tag{2.18}$$

where $\pi_d(d) = \theta\exp(-\theta d)$ and $r = \exp(-\theta)$. The constant decay-rate means that we are penaling equally each additional portion of distance in the parameter space; no matter the initial point where we are. This is a reasonable choice in situations where we have not a clear idea about the distance scale. As we said for the base model, nothing prevent us to make a different assumption on the distribution of the distance, but in this case we would drop the constant penalisation rate assumption. At the best of our knowledge, the only continuous distribution to have this property is the exponential distribution, while in the discrete case, the geometric distribution share this property.
We define the PC prior by means of a change of variable

$$\pi(\xi) = \pi_d(d(\xi))\left|\frac{\partial d(\xi)}{\partial \xi}\right|. \tag{2.19}$$

- **User-defined scaling.** The parameter θ can be chosen by making an assumption on a tail event

$$\mathrm{Prob}(Q(\xi) > W) = \alpha. \tag{2.20}$$

This is the crucial point of the principled procedure, since the wrong choice of the rate parameter can affect significantly the estimates. As we will see later, we propose a hierarchical approach where we aim to render the PC prior a sort of objective prior.
Anyhow, the choice of θ is a user task and this render the PC prior similar to a weakly informative prior. Notice that $Q(\xi)$ is a generic tranformation of the parameter ξ, it could be for instance $d(\xi)$ or ξ itself, while W is an upper bound defined by the user and α is the weight we put on the tail event. By changing the prior mass in the tail, we prescribe how informative our prior is.

Chapter 3

PC prior in Copula Models

Copulas are particular multivariate distribution functions with standard uniform univariate margins. Recall that the distribution function H of a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is the function defined by

$$H(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d), \quad \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d. \quad (3.1)$$

Then, the copula of (X_1, X_2) is simply the cumulative distribution function of $(F_1(X_1), F_2(X_2))$. The requirement that the margins of a copula be standard uniform is somewhat arbitrary. For instance, **?** originally considered multivariate distribution functions whose margins were $\mathrm{U}(-1/2, 1/2)$. Another example arises when studying dependence between componentwise block maxima in multivariate extreme-value theory; in that case, it is often more natural to standardise to unit Fréchet margins with distribution function $F(x) = \exp(-1/x),\ x > 0$. The important message is that no matter what continuous univariate distributions the margins are transformed to, it does not alter the philosophy behind the "copula approach" to the study of the dependence. The choice of $\mathrm{U}(0, 1)$ margins turns out to be a natural and convenient one.
In recent years, copulas have turned out to be the subject of a large number of scientific publications; see, for instance, **?** for a bibliometric overview in finance. They were applied in a wide variety of areas such as quantitative risk management, econometric modeling, or environmental modeling, to name a very few; see, for example **?**, **?**, and **?**. The reason for what could be inelegantly called "the copula craze" lies in Sklar's Theorem (**?**). Let $(X_1, \ldots, X_d)$ be a d-dimensional random vector and let $H(\mathbf{x}) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d),\ \mathbf{x} \in \mathbb{R}^d$, be its distribution function. The first part of Sklar's Theorem asserts that every d-dimensional distribution function H can be expressed as

$$H(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d)), \qquad \mathbf{x} \in \mathbb{R}^d, \quad (3.2)$$

in terms of a d-dimensional copula C and the univariate marginal distribution functions $F_1, \ldots, F_d$ obtained from H by $F_j(x_j) = H(\infty, \ldots, \infty, x_j, \infty, \ldots, \infty)$, $x_j \in \mathbb{R}$.
The copula C is thus the function which connects or couples the marginal distribution functions $F_1, \ldots, F_d$ to the multivariate distribution function H, hence the name "copula." For estimation of H from data, this offers a great deal of flexibility as it allows one to model the marginal distribution functions $F_1, \ldots, F_d$

separately from the dependence represented by the copula C, which is often of interest from a statistical and numerical point of view. The second part of Sklar's Theorem provides a converse. Given any copula C and univariate distribution functions $F_1, \ldots, F_d$, a multivariate distribution function H can be composed via (**??**) which then has univariate margins $F_1, \ldots, F_d$ and "dependence structure" C.
Note that from (**??**) we easily derive the joint density function as

$$h(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i) \cdot c(F_1(x_1), \ldots, F_d(x_d)). \tag{3.3}$$

In practice the joint density function is the product of the marginal densities times a copula density function c.
In the next section we will show that the Kullback-Leibler divergence does not depend on the marginal densities and their parameter and this is crucial in deriving PC priors for copulas as we need to take into account only the copula structure.

3.1 Invariance to Marginals

In this section, we consider a multivariate model with independent marginals as a benchmark for a generic multivariate model where the marginals are not independent. The PC prior takes natural place in such a context, as we can include in the model an extra-component taking into account for dependence; this latter is represented by the parameter of a copula density function.
We will see that the PC prior for the copula parameters can be derived regardless of the parameters of the marginal distributions. This is a very neat observation which has important modelling implications and open new venues in the research field of objective Bayesian analysis. In practice, we show that the KLD between a generic multivariate model and the model with independent marginals, does not depend on the marginal densities. This result is achieved by resorting the KLD in terms of copulas.
Here, we provide a summary of the strategy leading to equation (**??**), while the formal theorem and proof are displayed later on.
Suppose we have a joint distribution with dependent marginals. This means we have a model of the form

$$P = \{f_{\mathbf{X};\Theta}(\mathbf{x};\Theta), \mathbf{x} \in \mathbb{R}^k, \Theta \in \mathbb{R}^q\}; \tag{3.4}$$

therefore, according to the Sklar's Theorem, the joint density can be written as

$$f_{\mathbf{X};\Theta}(\mathbf{x};\Theta) = \prod_{j=1}^{k} f_j(x_j;\underline{\theta}_j) c_\psi(F_1(x_1;\underline{\theta}_1), \ldots, (F_k(x_k;\underline{\theta}_k);\psi), \tag{3.5}$$

where $\Theta = \{\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_k, \psi\}$.
Furthermore, let $f_{\mathbf{X};\Theta_0}(\mathbf{x}) = \prod_{j=1}^{k} f_j(x_j, \underline{\theta}_j)$ be the density of $\mathbf{X}$ in the case in which there is independence among the marginals, namely, when the value of ψ returns the independence copula (i.e. $\psi = 0$ for the Gaussian copula).
Given the mapping $u_j = F(x_j)$, the determinant of the Jacobian of the transformation is equal to the inverse of the product of the marginals, therefore the

parts related to the marginals cancel out. In particular, the derivative of the quantile function of U_i, $i = 1, \dots, k$ with respect to u_i is equal to the inverse of the density of X_i. This is the reason why we can drop the marginal distributions from equation (??).
After the change of variable, the Kullback-Leibler divergence can be expressed as

$$\text{KLD}(f_{\mathbf{X};\Theta} \| f_{\mathbf{X};\Theta_0}) = \int_{\mathcal{U}_1} \cdots \int_{\mathcal{U}_k} c(u_1, u_2, \dots, u_k; \psi) \times \log c(u_1, u_2, \dots, u_k; \psi) du_1 \dots du_k, \quad (3.6)$$

where $U_i \sim \text{Unif}(0, 1)$, $i = 1, \dots, k$, so that the integral is meant to be on the unit hypercube. We can notice that the KLD does not depend on the marginal densities of $\mathbf{X}$ and, as a consequence, on the parameters $\underline{\theta}_1, \dots, \underline{\theta}_k$.
This result suggests a general strategy to derive penalised complexity priors in an independent way with respect to the parameters of the marginals.
The finding contained in (??) is an interesting property of the KLD, and since PC priors (?) are derived starting from the KLD, they do not depend on the marginals too.
The strategy consists of considering a statistical model M_0 such that

$$M_0 : \qquad \mathbf{X} \sim f_0(\mathbf{x}; \Lambda) = f_{\mathbf{X};\Lambda}(\mathbf{x}; \Lambda) = \prod_{j=1}^{k} f_j(x_j; \underline{\theta}_j), \quad (3.7)$$

where $\Lambda = \{\theta \in \Theta : \psi = \psi_0\}$ is a subset of Θ. The above model could be made "more complex" by introducing a dependence stucture among the X_j-s, and we can do that by adding a copula function in the Sklar's representation

$$f_\Theta(\mathbf{x}; \Theta) = f_0(\mathbf{x}; \Lambda) c_\psi(F_1(x_1; \underline{\theta}_1), \dots, F_k(x_k; \underline{\theta}_k); \psi). \quad (3.8)$$

Notice that the prior distribution for ψ, as in ?, is obtained by taking a function of the Kullback-Leibler divergence, in order to transform this latter onto a physically interpretable "distance" scale

$$\varphi = d(\psi) = \sqrt{2\text{KLD}(f_\Theta(\psi) \| f_0)}, \quad (3.9)$$

and then by assigning to it an exponential density, which must be turned into the prior for ψ. Since $d(\psi)$ does not depend on $\underline{\theta}_1, \underline{\theta}_2, \dots, \underline{\theta}_k$, it is established a direct correspondence between φ and ψ that allows us to elicit the prior for the copula parameter in an independent manner with respect to the marginal densities. Notice that this result holds for any generic copula density function and also for any dimension. The procedure above make use of the following theorem.

Theorem 1 (Invariance to marginals)**.** *Let $\mathbf{X} \sim f_{\mathbf{X}}(x_1, \dots, x_k)$ be a random vector with density $f_{\mathbf{X}}$ (for the time being, we assume it is absolutely continuous with respect to the Lebesgue measure). Furthermore, let $\mathbf{Y}$ be a random vector with distribution $f_{\mathbf{Y}}(y_1, \dots, y_k) = \prod_{j=1}^{k} f_j(y_j)$ where f_j is the marginal density of X_j and Y_j, then*

$$\text{KLD}(f_{\mathbf{X}} \| f_{\mathbf{Y}}) = \int_{[0,1]^k} c(u_1, u_2, \dots, u_k; \psi) \log c(u_1, u_2, \dots, u_k; \psi) du_1 \dots du_k, \quad (3.10)$$

where $c(u_1, u_2, \dots, u_k)$ represents the copula function associated with the density of $\mathbf{X}$.

Proof. First, recall that the Kullback-Leibler Divergence on the original space is

$$\begin{aligned}\mathrm{KLD}(f_{\mathbf{X}} \| f_{\mathbf{Y}}) = \int_{\mathbf{X}} \prod_{j=1}^{k} f_j(x_j; \underline{\theta}_j) c(F_1(x_1; \underline{\theta}_1), F_2(x_2; \underline{\theta}_2), \dots, F_k(x_k; \underline{\theta}_k); \psi) \\ \times \log c(F_1(x_1; \underline{\theta}_1), F_2(x_2; \underline{\theta}_2), \dots, F_k(x_k; \underline{\theta}_k); \psi) dx_1 \dots dx_k. \quad (3.11)\end{aligned}$$

By applying the following change of variable

$$\Phi(u_1, u_2, \dots, u_k) = \begin{cases} x_1 = F_1^{-1}(u_1) \\ x_2 = F_2^{-1}(u_2) \\ \vdots \\ x_k = F_k^{-1}(u_k) \end{cases} \quad (3.12)$$

one obtains that the integral in (**??**) can be written as

$$\begin{aligned}\int_{[0,1]^k} \prod f_j(F^{-1}(u_j); \theta_j)\ c(u_1, u_2, \dots, u_k; \psi) \log c(u_1, u_2, \dots, u_k; \psi) \\ \times |\det(\mathbf{J}_\Phi(u_1, \dots, u_k))| du_1 \dots du_k, \quad (3.13)\end{aligned}$$

where

$$\mathbf{J}_\Phi(u_1, \dots, u_k) = \begin{bmatrix} \frac{\partial F^{-1}(u_1)}{\partial u_1} & \frac{\partial F^{-1}(u_2)}{\partial u_1} & \cdots & \frac{\partial F^{-1}(u_k)}{\partial u_1} \\ \frac{\partial F^{-1}(u_1)}{\partial u_2} & \frac{\partial F^{-1}(u_2)}{\partial u_2} & \cdots & \frac{\partial F^{-1}(u_k)}{\partial u_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F^{-1}(u_1)}{\partial u_k} & \frac{\partial F^{-1}(u_2)}{\partial u_k} & \cdots & \frac{\partial F^{-1}(u_k)}{\partial u_k} \end{bmatrix} \quad (3.14)$$

and since

$$\begin{aligned}\frac{\partial F_1^{-1}(u_1)}{\partial u_1} = \frac{\partial x_1}{\partial u_1} = \frac{1}{F_1'(x_1)} = \frac{1}{F_1'(F^{-1}(u_1))} = \frac{1}{f_1(F^{-1}(u_1))} \\ \vdots \quad (3.15) \\ \frac{\partial F_k^{-1}(u_k)}{\partial u_k} = \frac{\partial x_k}{\partial u_k} = \frac{1}{F_k'(x_k)} = \frac{1}{F_k'(F^{-1}(u_k))} = \frac{1}{f_k(F^{-1}(u_k))},\end{aligned}$$

the Jacobian matrix takes the following form

$$\mathbf{J}_\Phi(u_1, \dots, u_k) = \begin{bmatrix} \frac{1}{f_1(F^{-1}(u_1))} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_2(F^{-1}(u_2))} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{f_k(F^{-1}(u_k))} \end{bmatrix}, \quad (3.16)$$

where any off-diagonal element is equal to zero.
Therefore, the determinant of the Jacobian matrix is equal to the inverse of the

product of the marginals

$$\det(\mathbf{J}_\Phi(u_1,\dots,u_k)) = \frac{1}{\prod_{j=1}^{k} f_j(F^{-1}(u_j))}. \tag{3.17}$$

In this case, we do not take into account the absolute value because we are dealing with density functions which are positive by definition.
Finally, the Kullback-Leibler divergence can be written as

$$\int_{[0,1]^k} \prod f_j(F^{-1}(u_j);\theta_j)\; c(u_1,u_2,\dots,u_k;\psi) \log c(u_1,u_2,\dots,u_k;\psi) \times \frac{1}{\prod f_j(F^{-1}(u_j);\theta_j)} du_1 \dots du_k, \tag{3.18}$$

where the first and the last factors cancel out, so the proof is completed. □

3.2 The construction of the PC Prior in a generic Bivariate Copula Model

Suppose now to have only two marginal distributions. Notice that the finding above applies to every dimension, but for more than two marginal distributions we need to define a multivariate PC prior, a part from the case in which each pair of marginals share the same correlation.
On the basis of Theorem **??**, we know that in the uniform space the Kullback-Leibler divergence between the joint distribution in the case of dependence among the marginals (see equation (**??**)) and its base model (independence case, i.e. when $c_\psi(u,v) = 1$) does not depend on the marginal distributions and their parameters, so we can write down the KLD between the two models as follow

$$\mathrm{KLD}(f_{\mathbf{X};\Theta} \| f_{\mathbf{X};\Theta_0}) = \int_{\mathcal{U}} \int_{\mathcal{V}} c(u,v;\psi) \log c(u,v;\psi) du dv. \tag{3.19}$$

? recommend to take a function of the KLD, then we obtain the distance as a function of ψ as follows

$$d(\psi) = \sqrt{2\mathrm{KLD}(\psi)}. \tag{3.20}$$

Now, in order to satisfy the constant decay rate condition, we assign an exponential prior distribution to the distance scale, and we select the rate parameter of the exponential density itself by making a probability statement on a tail event.
The last step is to make a simple change of variable in order to get the PC prior for ψ

$$\pi(\psi) = \pi(d(\psi)) \left| \frac{\partial d(\psi)}{\partial \psi} \right|, \tag{3.21}$$

where, for the most of the copula models, the derivative of the distance with respect to ψ must be found numerically. To numerically compute the derivative above we make use of the following formula (Leibnitz's Rule). First, we calculate the derivative of the KLD by exploiting the interchange of the integral and the

derivative

$$\mathrm{KLD}'(\psi) = \frac{d}{d\psi}\int_0^1\int_0^1 c(u,v;\psi)\log c(u,v;\psi)dudv \tag{3.22}$$

$$= \int_0^1\int_0^1 \frac{\partial}{\partial\psi}c(u,v;\psi)\log c(u,v;\psi)dudv, \tag{3.23}$$

then we can compute the derivative of the distance function

$$d'(\psi) = \frac{1}{d(\psi)}\mathrm{KLD}'(\psi). \tag{3.24}$$

In particular, attention must be paid as the distance function could be non-monotone. For instance, in the Gaussian copula model, the distance is symmetric around the base model as it is piecewise monotone according to the sign of ψ.
Then, for the Gaussian copula model (but, in general, for any model whose distance function has two pieces in which it is monotone), we must consider the two branches of the distance separately. So, we should write down the density function of each branch as follows

$$\pi(d_i(\psi)) = \frac{1}{2}\theta\exp(-\theta d_i(\psi)), \qquad i = 1,2, \tag{3.25}$$

where $d_1(\psi)$ and $d_2(\psi)$ are the distances when $-1 < \psi < 0$ and $0 \leq \psi < 1$, respectively. Notice that each branch of the distance function has half an exponential distribution to be assigned, given that we want the PC prior for ψ to be symmetric, in the sense to have the median at the base model; we will give a better explanation of this concept later on.
Anytime the distance function is non-monotone, such as for the Gaussian copula, we must consider a partition of the parameter space, in order to have the distance to be a monotone function on each piece.
In the Gaussian copula example, there are two branches. The first one is for the negative values of ψ corresponding to a monotone $d(\psi)$ on $(-1,0)$, while the second one is for the positive values of ψ for which $d(\psi)$ is monotone on $(0,1)$. Therefore, the resulting prior for ψ is the sum of the two branches

$$\pi(\psi) = \begin{cases} \sum_{i=1}^{2}\pi(d_i(\psi))\left|\frac{\partial d_i(\psi)}{\partial\psi}\right| & \text{if } d(\psi)\in\Theta \\ 0 & \text{otherwise,} \end{cases} \tag{3.26}$$

where Θ is the parameter space of the distance scale.
The last step is to select the rate parameter θ of the density in (??) by considering a probability statement, like $\mathrm{Prob}(d(\psi) > W) = \alpha$. Recall that the exponential C.D.F. is equal to $1 - e^{-\theta d(\psi)}$, while the C.D.F. for the density in (??) is equal to $\frac{1}{2} - \frac{1}{2}e^{-\theta d(\psi)}$, as a consequence the survival function is $\frac{1}{2}e^{-\theta d(\psi)}$.
The probability statement above allows us to equate $\frac{1}{2}e^{-\theta W} = \alpha$, and with some basic algebra we obtain $\theta = -\frac{\log(2\alpha)}{W}$.
Here, it is worthwhile to mention that in the probability statement used to elicit the parameter θ, either considering the distance function or directly the parameter of interest does not make any difference. In fact, there is a direct correspondence between the parameter ψ and the distance $d(\psi)$. In any case, both the definitions of probability statement lead to the same quantity of θ.

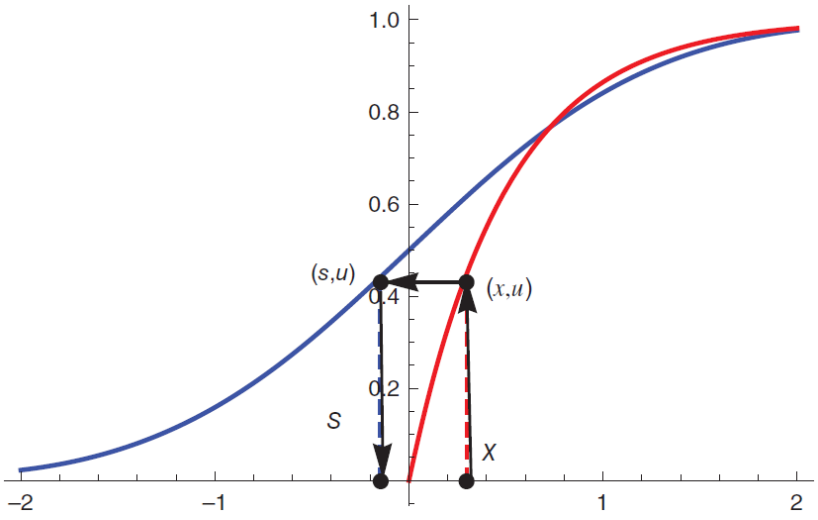

Figure 3.1: Transformation of variables for Gaussian copula.

3.3 The construction of the PC Prior in the Bivariate Gaussian Copula Model

The Gaussian copula and the Student copula are the ones most frequently used in application. Suppose to have two dependent random variables X and Y, for any choice of the distribution functions. For modelling joint distributions we can combine the Gaussian copula with any marginal distribution $u = F(x)$ and $v = G(y)$.
Figure ?? represents the transformation of variables $x \to u \to s$. Similarly one can transform $y \to v \to t$. The idea of the Gaussian copula is to transform the random variables X and Y, with respective cumulative distribution functions F and G, into standard normal variables $S = \Phi^{-1}(F(X))$ and $T = \Phi^{-1}(F(Y))$.
Then, the dependence between X and Y is expressed in terms of the dependence structure of their normal transformations S and T, therefore it can be reduced to linear correlation. With this approach, nonlinear dependence between X and Y is expressed through the linear dependence of their standard normal transforms.
We obtain the Gaussian copula by exploiting the distribution function of the bivariate normal distribution $\Phi_\rho(s,t)$ with zero means, unit variances, and correlation ρ between the components.
Let's $\Phi(\cdot)$ be the standard normal cumulative distribution function. So, we can write the Gaussian copula as

$$C_\rho(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)). \tag{3.27}$$

Let's denote $s = \Phi^{-1}(u)$ and $t = \Phi^{-1}(v)$. Then, the density of the Gaussian copula comes out from the differentiation of equation (??) as follows

$$c_\rho(u,v) = \frac{\partial^2 C_\rho(u,v)}{\partial u \partial v} = \frac{\partial^2 \Phi_\rho(s,t)}{\partial s \partial t} \cdot \frac{\partial s}{\partial u} \cdot \frac{\partial t}{\partial v} = \frac{\phi_\rho(s,t)}{\phi(s)\phi(t)}, \tag{3.28}$$

where $\phi_\rho(s,t)$ is the density function of the bivariate normal distribution with zero means, unit variances, and correlation ρ, while $\phi(\cdot)$ is the density function of the standard normal distribution.
Therefore, the Gaussian copula density turns out to have the following analytical expression

$$c_\rho(u,v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2 s^2 + \rho^2 t^2 - 2\rho st}{2(1-\rho^2)}\right). \tag{3.29}$$

Notice that if our marginals were two normally distributed random variables with zero means and unit variances, we would obtain that $x = s$ and $y = t$.
Unlike more complicated copula densities, in the Gaussian copula model we are able to derive the functional form of the PC prior for the correlation parameter, ρ.
Starting from (??), in order to calculate the Kullback-Leibler divergence, we have to solve the following double integral

$$\begin{aligned}\mathrm{KLD}(\rho) = &\int_0^1 \int_0^1 \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\rho^2\Phi^{-1}(u)^2+\rho^2\Phi^{-1}(v)^2-2\rho\Phi^{-1}(u)\Phi^{-1}(v)\right)} \\ \times &\left[-\frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left(\rho^2\Phi^{-1}(u)^2+\rho^2\Phi^{-1}(v)^2-2\rho\Phi^{-1}(u)\Phi^{-1}(v)\right)\right] du dv,\end{aligned} \tag{3.30}$$

where $\Phi^{-1}(u) = s$ and $\Phi^{-1}(v) = t$. With some algebra the above integral reduces to $\frac{1}{2}\log(1-\rho^2)$, that in fact corresponds to the mutual information of the bivariate normal distribution

Proof. Starting from equation (??) we obtain

$$\begin{aligned}\mathrm{KLD}(\rho) = &-\frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)} \\ &\times \int_0^1 \int_0^1 \left[\rho^2\Phi^{-2}(u) + \rho^2\Phi^{-2}(v) - 2\rho\Phi^{-1}(u)\Phi^{-1}(v)\right] \\ \times &\frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2\Phi^{-2}(u) + \rho^2\Phi^{-2}(v) - 2\rho\Phi^{-1}(u)\Phi^{-1}(v)}{2(1-\rho^2)}\right) du dv.\end{aligned} \tag{3.31}$$

Now we can exploit the following change of variables

$$\begin{cases} u = \Phi(s) \\ v = \Phi(t) \end{cases} \tag{3.32}$$

and, by consequence, $du = \phi(s)ds$ and $dv = \phi(t)dt$.
Therefore (??) can be written as

$$\begin{aligned}\mathrm{KLD}(\rho) = &-\frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\rho^2 s^2 + \rho^2 t^2 - 2\rho st\right] \\ &\times \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2 s^2 + \rho^2 t^2 - 2\rho s}{2(1-\rho^2)}\right) \phi(s)\phi(t) ds dt.\end{aligned} \tag{3.33}$$

The integral in (??) can be splitted into three pieces.
The first piece is

$$\begin{aligned}
&\frac{\rho^2}{\sqrt{1-\rho^2}}\int_{\mathbb{R}^2} s^2\frac{1}{2\pi}e^{-\frac{1}{2}s^2}e^{-\frac{1}{2}t^2}e^{-\frac{1}{2(1-\rho^2)}\rho^2 s^2}e^{-\frac{1}{2(1-\rho^2)}\rho^2 t^2}e^{\frac{1}{2(1-\rho^2)}2\rho st}dsdt\\
=&\rho^2\int_{\mathbb{R}^2} s^2\frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[s^2+t^2-2\rho st\right]}dsdt\\
=&\rho^2\mathbb{E}S^2\\
=&\rho^2\left[(\mathbb{E}S)^2+\mathrm{Var}[S]\right]\\
=&\rho^2.
\end{aligned}$$

In the same way we obtain the second piece of the integral

$$\begin{aligned}
&\frac{\rho^2}{\sqrt{1-\rho^2}}\int_{\mathbb{R}^2} t^2\frac{1}{2\pi}e^{-\frac{1}{2}s^2}e^{-\frac{1}{2}t^2}e^{-\frac{1}{2(1-\rho^2)}\rho^2 s^2}e^{-\frac{1}{2(1-\rho^2)}\rho^2 t^2}e^{\frac{1}{2(1-\rho^2)}2\rho st}dsdt\\
=&\rho^2\int_{\mathbb{R}^2} t^2\frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[s^2+t^2-2\rho st\right]}dsdt\\
=&\rho^2\mathbb{E}T^2\\
=&\rho^2\left[(\mathbb{E}T)^2+\mathrm{Var}[T]\right]\\
=&\rho^2.
\end{aligned}$$

Finally, the third piece of the integral is

$$\begin{aligned}
&\frac{-2\rho}{\sqrt{1-\rho^2}}\int_{\mathbb{R}^2} st\frac{1}{2\pi}e^{-\frac{1}{2}s^2}e^{-\frac{1}{2}t^2}e^{-\frac{1}{2(1-\rho^2)}\rho^2 s^2}e^{-\frac{1}{2(1-\rho^2)}\rho^2 t^2}e^{\frac{1}{2(1-\rho^2)}2\rho st}dsdt\\
=&-2\rho\int_{\mathbb{R}^2} st\frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[s^2+t^2-2\rho st\right]}dsdt\\
=&-2\rho\big[\mathbb{E}[ST]\big]\\
=&-2\rho\big[\mathbb{E}[ST]-\mathbb{E}[S]\mathbb{E}[T]\big]\\
=&-2\rho\big[\rho\big]\\
=&-2\rho^2,
\end{aligned}$$

where the fourth equality comes out from the fact that $\mathbb{E}(S)=\mathbb{E}(T)=0$.
Therefore, the KLD is

$$\begin{aligned}
\mathrm{KLD}(\rho)&=-\frac{1}{2}\log(1-\rho^2)-\frac{1}{2(1-\rho^2)}\left[\rho^2+\rho^2-2\rho^2\right]\\
&=-\frac{1}{2}\log(1-\rho^2).
\end{aligned}\tag{3.34}$$

□

Once we have found the Kullback-Leibler divergence, the computation of the prior for ρ is straightforward. Initially we transform the KLD into the more interpretable distance function and then we place an exponential prior on it. The distance measure must be viewed separately for negative and positive

correlations

$$d(\rho) = \sqrt{-\log(1-\rho^2)} = \begin{cases} d_1(\rho), & -1 < \rho < 0 \\ d_2(\rho), & 0 \leq \rho < 1 \end{cases}, \tag{3.35}$$

where $d(\rho)$ is symmetric around 0 given that ρ is raised to the square power. Therefore, $d(\rho) = d(-\rho)$ for any ρ. By using (**??**) we have that

$$\left|\dot{d}(\rho)\right| = \frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}. \tag{3.36}$$

By using (**??**) the resulting PC prior for ρ is

$$\begin{aligned} \pi(\rho) = \frac{1}{2}\theta e^{-\theta\sqrt{-\log(1-\rho_-^2)}} \frac{|\rho_-|}{(1-\rho_-^2)\sqrt{-\log(1-\rho_-^2)}} \\ + \frac{1}{2}\theta e^{-\theta\sqrt{-\log(1-\rho_+^2)}} \frac{|\rho_+|}{(1-\rho_+^2)\sqrt{-\log(1-\rho_+^2)}}, \end{aligned} \tag{3.37}$$

where ρ_- and ρ_+ denote the negative and positive domain of ρ.
Over the whole domain of ρ we obtain

$$\pi(\rho) = \frac{\theta}{2}\exp\Big(-\theta\sqrt{-\log(1-\rho^2)}\Big)\frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}. \tag{3.38}$$

The latter is a proper prior and this is evident as the PC prior is a tranformation of the exponential prior assigned to the distance function. Anyhow, it can be numerically proved that

$$\int_{-1}^{1} \frac{\theta}{2}\exp\Big(-\theta\sqrt{-\log(1-\rho^2)}\Big)\frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}d\rho = 1. \tag{3.39}$$

Another remarkable fact is that the prior is clearly symmetric as it depends on ρ only through the square and the absolute value, but keep in mind that this depends on the equal weight assigned to the positive and negative correlations in the parameter space. In addition, such as the standard normal distribution, any odd moment is equal to zero

$$\int_{-1}^{1} \rho^k \frac{\theta}{2}\exp\Big(-\theta\sqrt{-\log(1-\rho^2)}\Big)\frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}d\rho = 0, \tag{3.40}$$

for k being any odd natural number.

3.4 Derivation of the Jeffreys' Prior for ρ

In this section, we derive the Jeffreys' prior for ρ and then we will use it to make a comparison with our PC prior, so that we can understand how close to objectivity PC prior can be.
The Jeffreys' prior encapsulate the amount of information that a random variable following a Gaussian copula distribution carries about the parameter ρ. In

practice, it is the variance of the score function, or, in other words, the expected value of the observed information, namely the negative of the second derivative of the log-likelihood.
The Jeffreys' prior is the square root of the Fisher information for ρ, where this latter is defined as

$$I(\rho) = -\mathbb{E}\left[\frac{\partial^2}{\partial\rho^2}\log f(x|\rho)\middle|\rho\right], \tag{3.41}$$

where $\log f(x|\rho)$ is the log-likelihood function and $f(x|\rho)$ is the Gaussian copula density function.
Let's start from the log-likelihood function

$$\log f(x|\rho) = -\frac{1}{2}\left(\log(1-\rho^2) + \frac{s^2\rho^2 + t^2\rho^2 - 2\rho st}{1-\rho^2}\right), \tag{3.42}$$

where s and t are defined above.
Then, the first derivative is

$$\frac{\partial \log f(x|\rho)}{\partial\rho} = \frac{-s^2\rho + st(1+\rho^2) - \rho(t^2+\rho^2-1)}{(1-\rho^2)^2}, \tag{3.43}$$

while the second derivative is

$$\frac{\partial^2 \log f(x|\rho)}{\partial\rho^2} = -\frac{s^2(3\rho^2+1) - 2\rho st(\rho^2+3) + t^2(3\rho^2+1) + \rho^4 - 1}{(1-\rho^2)^3}. \tag{3.44}$$

Then, we can calculate the Fisher information as follow

$$\begin{aligned} I(\rho) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\rho^2}\log f(x|\rho)\middle|\rho\right] \\ &= \frac{\mathbb{E}(s^2)(3\rho^2+1) - 2\mathbb{E}(st)\rho(\rho^2+3) + \mathbb{E}(t^2)(3\rho^2+1) + \rho^4 - 1}{(1-\rho^2)^3} \\ &= \frac{(3\rho^2+1) - 2\rho^2(\rho^2+3) + (3\rho^2+1) + \rho^4 - 1}{(1-\rho^2)^3} \\ &= \frac{1-\rho^4}{(1-\rho^2)^3} = \frac{(1-\rho^2)(1+\rho^2)}{(1-\rho^2)^3} \\ &= \frac{1+\rho^2}{(1-\rho^2)^2}, \end{aligned} \tag{3.45}$$

where the third equality comes out from the fact that the second moment of the random variables S and T is equal to 1, while the covariance is equal to ρ.
Therefore, the Jeffreys' prior turns out to be

$$\pi^J(\rho) = \frac{\sqrt{1+\rho^2}}{1-\rho^2}. \tag{3.46}$$

Figure **??** shows the PC prior for varying θ along with the Jeffreys' prior. We can see that the Jeffreys' prior is approached by the PC prior with a smaller and smaller value of θ. Furthermore, notice that the velocity by which the PC prior goes to infinity is greater than the one for the Jeffreys' prior. In the former case, the velocity is exponential, while the Jeffreys' prior goes to infinity with a square root rate, therefore slower than the PC prior.

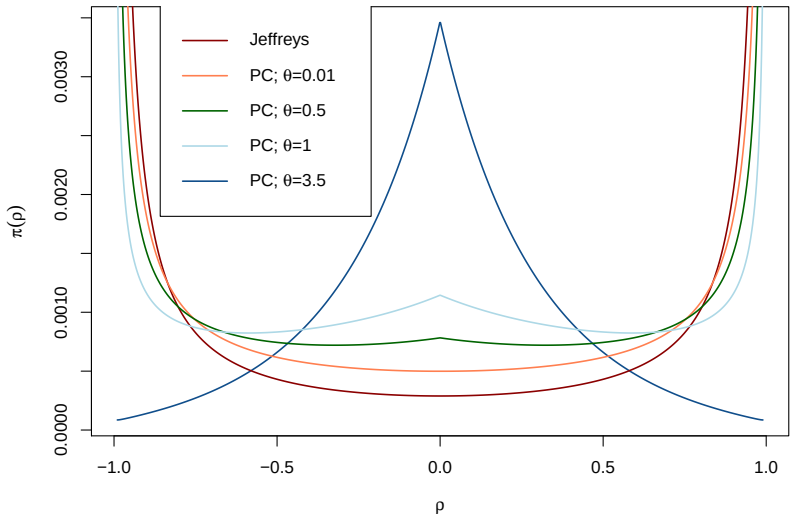

Figure 3.2: The PC prior for varying θ and the Jeffreys' prior.

3.5 Generalised Beta distribution

In order to be able to compare our PC prior with the objective prior distributions for ρ existing in the literature, we just propose an alternative to the Jeffreys' prior above. In particular, we try to construct an objective version of the Generalised Beta distribution. Unlike the Jeffreys' prior, the Generalised Beta distribution is proper and this means it could be used in Bayesian hypothesis testing problems.
This versatile version of the Beta distribution has the following density

$$\pi(\rho) = \frac{\Gamma(u+v)}{(b-a)^{u+v-1}\Gamma(u)\Gamma(v)}(\rho-a)^{u-1}(b-\rho)^{v-1}; \qquad x \in (a=-1, b=1), \tag{3.47}$$

where, as for the usual Beta density, the parameters $u, v > 0$.
Given that we are taking a prior over the correlation, the boundaries a and b are respectively equal to -1 and 1, but, in general, we could consider any interval of interest.
Our proposal to render the Generalised Beta distribution a sort of objective prior consists of selecting the hyperparameters u and v in the following manner. We could set the mean and the variance of such a Generalised Beta density equal to the sample correlation and the maximum of the variance, respectively. In practice, we have a sistem of two equations in two unknows of the following form

$$\begin{cases} \mathbb{E}[\rho] = \frac{u}{u+v}(b-a) + a = \hat{\rho} \\ \text{Var}[\rho] = \frac{(u+1)u}{(u+v+1)(v+u)}(b-a)^2 + \frac{u^2}{(u+v)^2}(b-a)^2 = 8 \end{cases} \tag{3.48}$$

where 8 is the maximum variance that this particular Generalise Beta density can attain.

By solving the system we retrieve the hyperparameters in an objective fashion. The only thing one could criticize is that we use the data twice (once in the construction of the prior and then in the derivation of the posterior), even though it has been showed that the Empirical Bayes methods have good asymptotic properties (**?**).

3.6 Deriving the hyperparameter of the PC Prior

Unlike the Jeffreys' prior, the PC prior depends on the parameter θ that need to be specified. This latter plays a fundamental role as it influences how fast the prior shrinks towards the base model, i.e. the independence among marginals. To select θ, **?** propose to use a probability statement based on a tail event. In our opinion, a reasonable choice in this case would be to define the tail event such that very large or zero correlations should be not much likely. Anyhow, it turns out that the choice of θ is a crucial point and as a consequence it will have a considerable impact on the estimates.

3.6.1 Jeffreys' Prior for θ

Instead of using the probability statement, another stategy we could consider is the hierarchical approach consisting in assigning a prior distribution to the rate parameter θ. The rate parameter θ lies in $(0, \infty)$, so any prior distribution with the same support works. For instance, we could use a Gamma or an Exponential prior, but in this case we should define the hyperparameter of these distributions.
As we have seen for the Jeffreys' prior for ρ, firstly we need to calculate the Fisher information for θ

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \rho^2} \log f(\rho|\theta) \middle| \theta\right], \tag{3.49}$$

where $f(\rho|\theta) = \frac{\theta}{2} \exp\Big(-\theta\sqrt{-\log(1-\rho^2)}\Big) \frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}$ is the PC prior for ρ.
Therefore

$$\log f(\rho|\theta) = \log(\theta) - \log(2) - \theta\sqrt{-\log(1-\rho^2)} + \log(|\rho|) - \log(1-\rho^2) - \log(\sqrt{-\log(1-\rho^2)}). \tag{3.50}$$

Now, let's compute the first and the second derivatives

$$\frac{\partial \log f(\rho|\theta)}{\partial \theta} = \frac{1}{\theta} - \sqrt{-\log(1-\rho^2)}, \tag{3.51}$$

$$\frac{\partial^2 \log f(\rho|\theta)}{\partial \theta^2} = -\frac{1}{\theta^2}. \tag{3.52}$$

Then

$$-\mathbb{E}\left[-\frac{1}{\theta^2}\right] = \frac{1}{\theta^2}. \tag{3.53}$$

It is immediate to see that the Jeffreys' prior for θ is improper. In fact,

$$\pi^J(\theta) = \frac{1}{\theta} \tag{3.54}$$

is not integrable in the positive real line and therefore it does not guarantee the posterior to be proper. Anyhow, we will see in the next section that this result turns out to be useful for the calculation of the intrinsic prior for θ.

3.6.2 Intrinsic Prior for θ

The Jeffreys' prior for θ is obviously an objective choice but we are not sure about the properness of the posterior distribution for ρ.
As an alternative, from an objective point of view, we can calculate the intrinsic prior for the rate parameter θ and then we can specify the hyperparameter of such an intrinsic prior distribution by maximizing the variance of the PC prior for ρ where an intrinsic prior is put on θ. The procedure to derive the intrinsic prior is borrowed from **?** and is described below.
Let $M_1 = \{f_{\rho;\Theta}(\rho;\theta), \rho \in [-1,1], \theta \in \mathbb{R}^+\}$, where $f_{\rho;\Theta}(\rho;\theta)$ is the penalised complexity prior for ρ, and $M_0 = \{f_{\rho;\Theta_0}(\rho;\theta_0), \rho \in [-1,1], \theta_0 \in \mathbb{R}^+\}$, where $f_{\rho;\Theta_0}(\rho;\theta_0) = f_{\rho;\Theta}(\rho;\theta)|_{\theta=\theta_0}$. Suppose we want to test the hypothesis $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. If $\pi^N(\theta) = \frac{1}{\theta}$ (that is the Jeffreys' prior for θ), the intrisic prior for θ is given by

$$\pi^I(\theta) = \int_{-1}^{1} \pi(\theta|\rho_\ell) f(\rho_\ell|H_0) d\rho_\ell \tag{3.55}$$

where

$$f(\rho_\ell|\theta_0) = \frac{\theta_0}{2} \exp\Big(-\theta_0\sqrt{-\log(1-\rho_\ell^2)}\Big) \frac{|\rho_\ell|}{(1-\rho_\ell^2)\sqrt{-\log(1-\rho_\ell^2)}} \tag{3.56}$$

and

$$\begin{aligned}
\pi(\theta|\rho_\ell) &= \frac{\pi^N(\theta) f(\rho_\ell|\theta)}{m^N(\rho_\ell)} \\
&= \frac{\frac{1}{2}\exp\Big(-\theta\sqrt{-\log(1-\rho_\ell^2)}\Big) \frac{|\rho_\ell|}{(1-\rho_\ell^2)\sqrt{-\log(1-\rho_\ell^2)}}}{\int_0^\infty \frac{1}{2}\exp\Big(-\theta\sqrt{-\log(1-\rho_\ell^2)}\Big) \frac{|\rho_\ell|}{(1-\rho_\ell^2)\sqrt{-\log(1-\rho_\ell^2)}} d\theta} \\
&= \exp\Big(-\theta\sqrt{-\log(1-\rho_\ell^2)}\Big)\sqrt{-\log(1-\rho_\ell^2)}
\end{aligned} \tag{3.57}$$

where ρ_ℓ represents the training sample. Note that if there is no subset of ρ_ℓ for which $0 < m^N(\rho_\ell) < \infty$, then ρ_ℓ is called *minimal training sample*. **?** showed that often it will simply be a sample of size $\max(\dim(\theta))$. It can be a smaller sample, however, especially if the π^N is proper in some variables. Recall that π^N is a starting distribution and it is typically improper, like in our case where it is the Jeffreys' prior for θ.
Therefore

$$\pi^I(\theta) = \frac{\theta_0}{(\theta+\theta_0)^2}. \tag{3.58}$$

The procedure by which we construct the intrinsic prior ensures that we obtain a proper prior, indeed we can see that $\int_0^\infty \frac{\theta_0}{(\theta+\theta_0)^2} d\theta = 1$.
In order to give an objective interpratation to our PC prior, we can use the intrinsic prior derived above. The only thing we have to do now is to set the

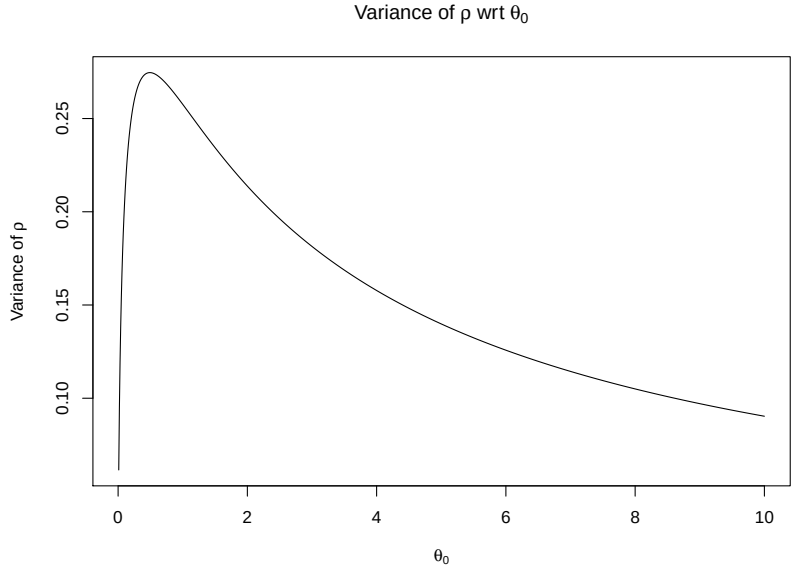

Figure 3.3: The variance of ρ as a function of θ_0.

hyperparameter θ_0 in an objective fashion and a way to do that is to consider the value of θ_0 that maximizes the variance.
In practice, the new prior for ρ (depending now on the hyperparameter θ_0) comes out from

$$\pi(\rho|\theta_0) = \int_0^\infty \pi^{PC}(\rho|\theta)\pi^I(\theta|\theta_0)d\theta. \tag{3.59}$$

We want to maximize the variance which, in practice, corresponds to the second moment given that the mean is equal to zero

$$\int_{-1}^{1}\int_0^\infty \rho\pi^{PC}(\rho|\theta)\pi^I(\theta|\theta_0)d\theta d\rho = 0. \tag{3.60}$$

Then, the variance of the new prior for ρ can be written as

$$\int_{-1}^{1}\int_0^\infty \rho^2\pi^{PC}(\rho|\theta)\pi^I(\theta|\theta_0)d\theta d\rho, \tag{3.61}$$

where the double integral is numerically computed. The maximizer is $\theta_0 = 0.491525$. Recall that this value of θ_0 can be viewed as an objective choice, given that it render the prior as flat as possible.
Figure **??** shows the variance of ρ as a function of θ_0. Notice that, in this case, the hierarchical model stabilise the variance of the PC prior for ρ, in fact, whithout considering the intrinsic prior for θ, the variance is hard to compute, even numerically and especially for value of θ between 0 and 1, given that for these value of θ the variance of the PC prior for ρ increases as the prior spreads out towards the alternative model.
The maximization problem is numerically solved by means of two different methods and both of them return the same value of θ_0. The first one is the Brent's

method, while the second one is the Golden section method that constricts more and more the interval around the maximum until it reach the true value of the maximum itself.

3.6.3 Subjective approach

We can also introduce our belief into the prior by assigning to the rate parameter θ a given prior distribution and then specifying the hyperparameters of such a distribution. By adding a level of hierarchy we have the advantage to stabilise the estimates but we must be careful, because we could force the probability mass to shrink towards the base model.
Consider, for instance, to assign to θ a Gamma prior distribution, then the new prior for ρ depends on the hyperparameters of the Gamma distribution as follows

$$\begin{aligned}
\pi(\rho|\alpha,\beta) &= \int_0^\infty \pi^{PC}(\rho|\theta)\cdot \text{Gamma}(\theta|\alpha,\beta)d\theta \\
&= \int_0^\infty \frac{\theta}{2} e^{-\theta\sqrt{-\log(1-\rho^2)}} \frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}} \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}d\theta \\
&= \frac{|\rho|}{2(1-\rho^2)\sqrt{-\log(1-\rho^2)}} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^\alpha e^{-\theta\left(\beta+\sqrt{-\log(1-\rho^2)}\right)}d\theta \\
&= \frac{|\rho|}{2(1-\rho^2)\sqrt{-\log(1-\rho^2)}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\left(\beta+\sqrt{-\log(1-\rho^2)}\right)^{\alpha+1}}, \qquad (3.62)
\end{aligned}$$

and for α being a positive integer we obtain

$$\pi(\rho|\alpha,\beta) = \frac{|\rho|}{2(1-\rho^2)\sqrt{-\log(1-\rho^2)}} \frac{\alpha\beta^\alpha}{\left(\beta+\sqrt{-\log(1-\rho^2)}\right)^{\alpha+1}}, \qquad (3.63)$$

where α and β are respectively the shape and rate parameter of the Gamma distribution.
Figure **??** shows the resulting PC prior after having put a Gamma prior on θ. In particular, the hyperparameters are $\alpha = 1$ and $\beta = 0.3$, making the Gamma prior distribution boiling down to an Exponential prior distribution with rate parameter equal to 0.3. Figures **??** and **??** represent different choices of the hyperparameters α and β. It is worth to notice that for both $\alpha, \beta \to 0$ the PC prior shows a spike at the base model, but still preserves some probability mass away from zero. This is similar, in a sense, to what happens with the intrinsic prior above. In practice, the hierarchical approach render our PC prior even more flexible. On the other hand, if we do not consider the hierarchical approach, by simply changing the value of θ we concentrate the probability mass at the base model or we spread it out at the boundaries.

As we have done above for the Generalised Beta distribution, we could render the Gamma prior distribution close to an objective prior by equating the mean and the variance of the density (**??**) to the sample correlation and the maximum variance, respectively. This allows us to find the hyperparameters α and β without introducing some specific user belief.

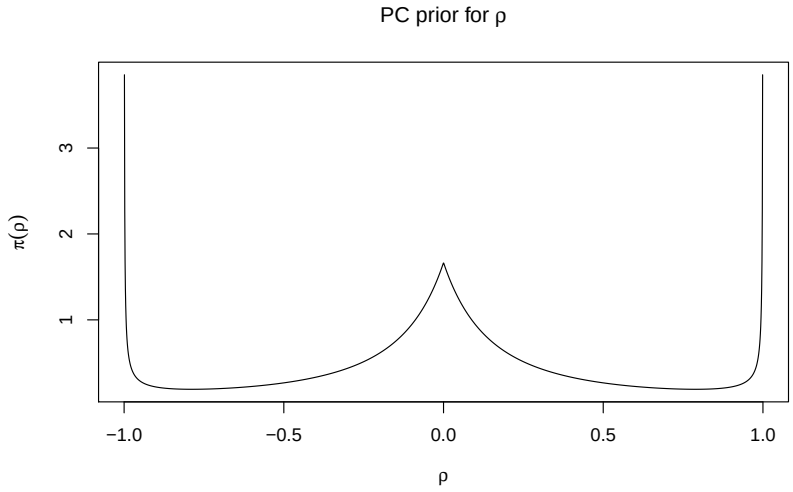

Figure 3.4: The PC prior for ρ where a Gamma$(1, 0.3)$ is assigned to θ.

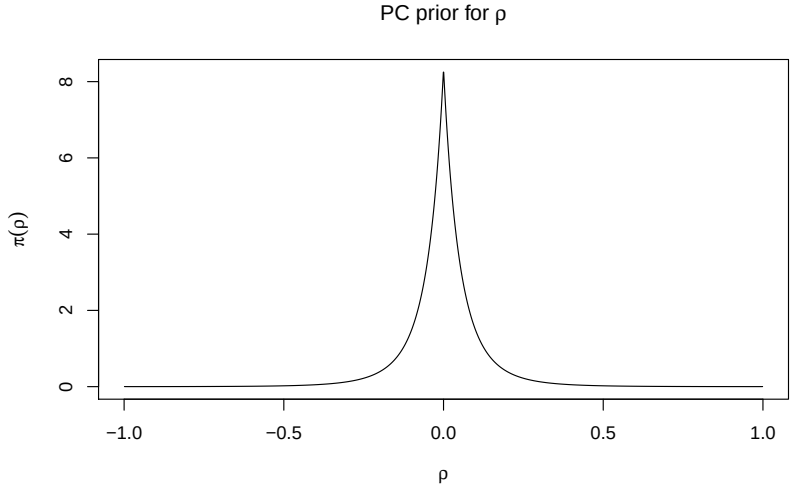

Figure 3.5: The PC prior for ρ where a Gamma$(5, 0.3)$ is assigned to θ.

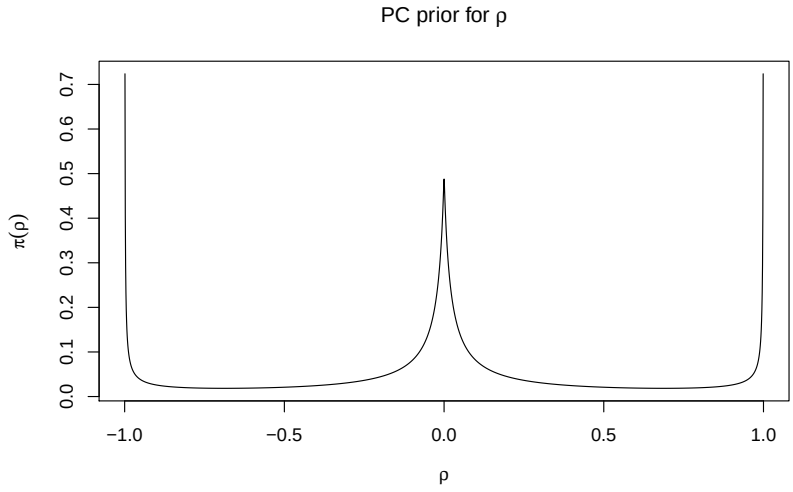

Figure 3.6: The PC prior for ρ where a Gamma$(0.02, 0.02)$ is assigned to θ.

3.7 The Arc-sine Prior for ρ

In this section we will present a proper alternative to the Jeffreys' prior for ρ. In fact, as we will see in the next section, given the improperness of the Jeffreys' prior for ρ, we have no default priors to be used in Bayesian hypothesis testing problems. Fortunately, we have an alternative that still preserves the nature of objective prior but it is proper. We are talking about the Arc-sine prior.
? described the improper prior for the correlation in bivariate normal data, conditional on the variances, as follows

$$\pi^J(\rho) \propto \frac{\sqrt{1+\rho^2}}{1-\rho^2}. \tag{3.64}$$

As we said above this prior cannot be used for Bayesian hypothesis testing. Then, **?** noted that the arc-sine prior,

$$\pi_{\text{arc-sine}}(\rho) = \frac{1}{\pi}\frac{1}{\sqrt{1-\rho^2}}, \tag{3.65}$$

is similar to the Jeffreys prior, but integrable on $[-1, 1]$.
Unlike the Jeffreys' prior, the arc-sine prior is not invariant to reparameterisation. Recall that also the PC prior has this desirable property of invariance to reparameterisation.
Let us consider, for instance, the reparameterisation in terms of the Kendall's Tau, τ, i.e. a measure of the dependence alternative to ρ. **?** has shown that for the bivariate normal model with correlation r

$$\tau = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)dH(x,y) - 1 = \frac{2}{\pi}\arcsin(r), \tag{3.66}$$

and by simply rearranging the double integral we can check that this relation is still valid for the Gaussian copula density.

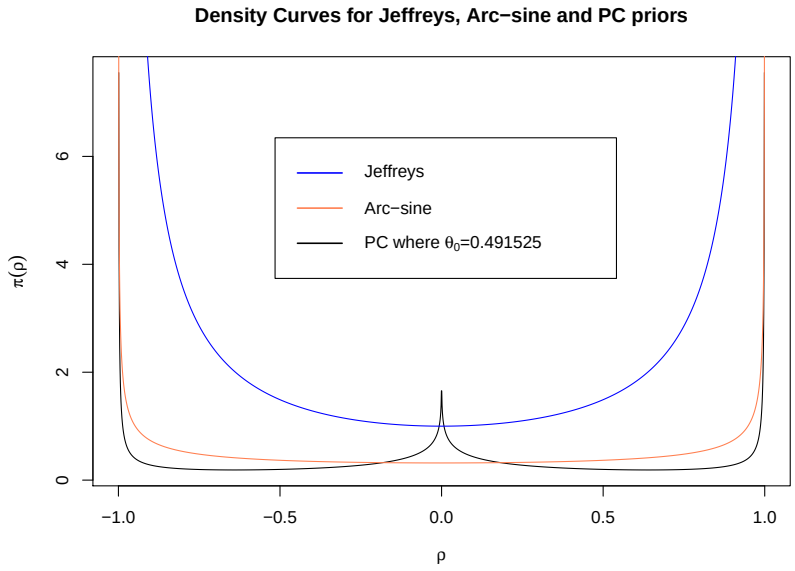

Figure 3.7: The Jeffreys', Arc-sine and PC priors for ρ, this latter has an intrinsic prior on θ where $\theta_0 = 0.491525$.

Even though the Jeffreys' and Arc-sine priors for ρ are quite similar, the corresponding priors for τ are completely different. Condider the Jeffreys' prior for ρ, then the Jeffreys' prior for τ is

$$\begin{aligned}\pi^J(\tau) &= \pi^J(\rho)\left|\frac{\partial\rho}{\partial\tau}\right| \\ &= \frac{\sqrt{1+\sin^2(\frac{\pi\tau}{2})}}{1-\sin^2(\frac{\pi\tau}{2})}\cos\left(\frac{\pi\tau}{2}\right)\frac{\pi}{2} \\ &= \frac{\pi}{2}\frac{\sqrt{1+\sin^2(\frac{\pi\tau}{2})}}{\cos(\frac{\pi\tau}{2})}.\end{aligned} \tag{3.67}$$

On the other hand, the Arc-sine prior for τ is an uniform distribution

$$\pi_{\text{arc-sine}}(\tau) = \frac{1}{2}. \tag{3.68}$$

So, even if the Jeffreys' and Arc-sine priors are very similar when computed for ρ, as we may see from Figure **??**, they look very different for τ.

3.8 Inference

We check out the frequentist performance of our PC prior, where an intrinsic prior is put on θ, via a simulation study. For each true value of the parameter ρ^* $(-0.95, -0.5, 0, 0.05, 0.5, 0.95, 0.999)$ and for each fixed sample size (n =

$5, 30, 100, 1000$) we have generated 200 independent samples from the Gaussian copula and for each of them we have calculated the posterior mean, the 95% credible interval and the Bayes Factor.
In addition, we compute the $\text{MSE}\big(\hat{\rho}(x)\big) = \mathbb{E}_{\hat{\rho}(x)}\Big[\big(\hat{\rho}(x) - \rho^*\big)^2\Big]$, where $\hat{\rho}(x) = \int_{-1}^{1} \rho\pi(\rho|x)d\rho$. Recall that we assign an intrinsic prior to the rate parameter of the PC prior for ρ and we maximize the variance by selecting $\theta_0 = 0.491525$.
The posterior distribution $\pi(\rho|x)$ is calculated by means of a sort of Metropolis-Hastings algorithm within Gibbs. We refer to the algorithm as a sort of Gibbs sampling because we do not deal with two full conditional distributions; in fact only the PC prior for ρ depends on θ. The acceptance rate, a, of the Metropolis-Hastings step for ρ is based on the posterior distribution as follows

$$a = \frac{L(\rho^P)\pi^{PC}(\rho^P|\theta^C)\pi^I(\theta^C)}{L(\rho^C)\pi^{PC}(\rho^C|\theta^C)\pi^I(\theta^C)}, \tag{3.69}$$

where the $\pi^I(\theta^C)$ at the numerator and at the denominator cancel out. ρ^P stands for the proposed ρ coming from the proposal density, while ρ^C stands for the current value. Notice also that before accepting or not a new ρ, a value of θ has been already accepted, so this is the meaning of θ^C. Obviously, in the Metropolis-Hastings step for ρ, the value of θ both at the numerator and the denominator is fixed, but the dependence on the value of θ still remains in the PC prior for ρ given θ, $\pi^{PC}(\rho^i|\theta^C)$, $i = P, C$.
In the estimation problem, we use the Jeffreys' prior as a term of comparison, since for inference we do not need the prior to be proper. The Jeffreys' prior is notoriously one of the best priors for estimation problems, therefore we use it as a benchmark for our prior, in order to check out objectivity.
What emerges from the simulation study is that our prior have a smaller Mean Squared Error than the Jeffreys' prior for true values of ρ close to zero. Figure **??** shows what we have just said. In fact, by looking at the boxplots we can easily notice how the distributions of the posterior mean, computed by using the PC prior, are more concentrated around the true value of ρ, and especially for moderate sample sizes. For the sake of completeness, we want to remark that the Jeffreys' prior has a smaller Mean Squared Error for intermediate correlations.
As one can expect, for small ρ^*, our objective PC prior is superior in terms of MSE to the Jeffreys' prior; this is because of the little spike of the objective PC prior in correspondence of the base model. If we want to compare the two priors overall, we should compute an overall MSE (overall values of ρ^*). This will result in a single number for both the priors. We may also give a graphical representation of the MSE by looking at the distributions of the squares of the posterior means minus the true values ρ^* (Figure **??**). This is possible as in the normal model the location and scale parameters are independent, but this is not true in general; let's think for instance of the Poisson model where the mean and the variance are the same. In addition, the variance of the proposal density in the MCMC algorithm is left equal for any value of ρ^* in a specific sample size scenario, therefore such a graphical representation makes sense.
In Figure **??**, the dashed line represents the mean of the squares and therefore it stands for the overall MSE. We can appreciate how the objective PC prior has a better behaviour overall, apart from the case of $n = 1000$, where the MSEs are basically the same. For instance, when $n = 100$, the overall MSE is smaller for the PC prior than the Jeffreys' prior, as they are respectively equal to 0.0034232

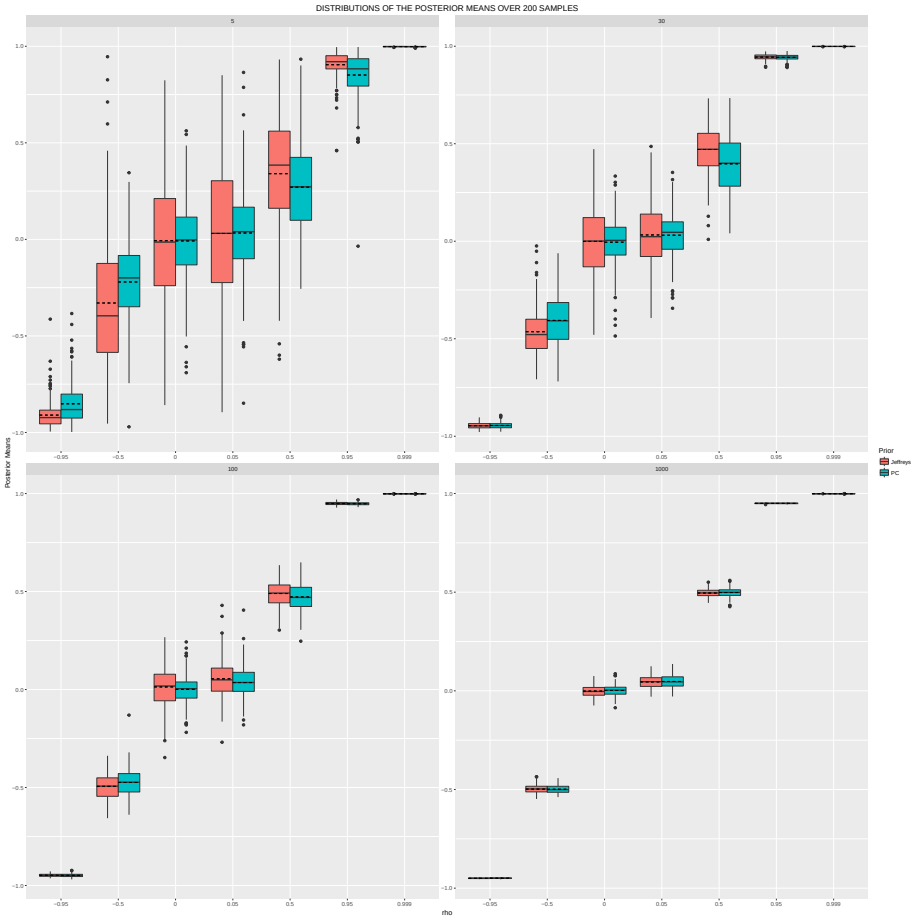

Figure 3.8: Distributions of the posterior mean computed over 200 samples. The blue boxplots represent the PC prior, while the red ones represent the Jeffreys' prior.

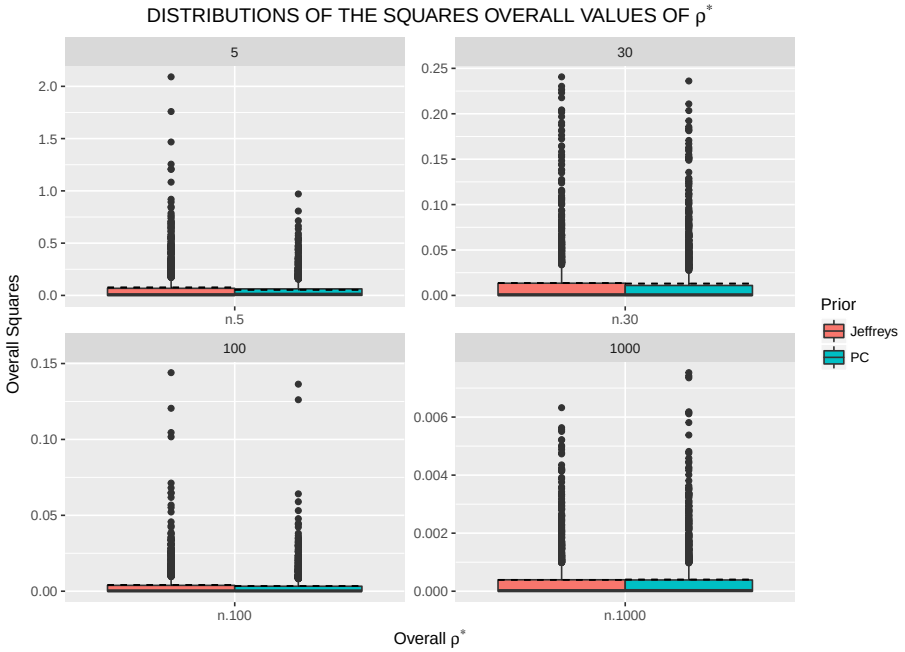

Figure 3.9: Distributions of the squares computed overall values of ρ^*. The blue boxplots stand for the PC prior, while the red ones stand for the Jeffreys' prior.

and 0.0041509.
It could be interesting to make a comparison between the objective version of the PC prior and the subjective one. For this purpose, we ran a simulation study using three different choices of θ, i.e. 0.1, 1 and 5. The results are resumed in Table **??**.
By looking at Table **??** and Table **??** simultaneously, we may appreciate the good behaviour of our objective PC prior. In fact, if we look at the scenario with $n = 100$ and $\rho = 0.95$, the objective PC prior is very close to the subjective PC prior with $\theta = 0.1$, even though the true value of the correlation is very large. Now, if we look at the same scenario but for $n = 30$, the two versions of the PC prior achieve basically the same MSE. It is obvious that the objective PC prior is well behaved in the vicinity of the base model; in fact, for $n = 100$, this latter is very similar in terms of MSE even to the subjective PC prior with a strong shrinkage, i.e. $\theta = 5$.
Now, let's look at the overall MSE for some scenario. For instance, when $n = 30$, the subjective PC priors with $\theta = 0.1$, $\theta = 1$ and $\theta = 5$ give overall MSEs respectively equal to 0.012117, 0.013042, 0.014732, whilst the objective PC prior gives an overall MSE equal to 0.013038. On the other hand, for $n = 100$, the subjective PC priors return overall MSEs equal to 0.0044079, 0.0037849 and 0.0037605, when θ is 0.1, 1 and 5, respectively, while the objective PC prior gives an overall MSE equal to 0.0034232, highlighting a better behaviour than the subjective PC prior for any choice of θ.
Finally, we want to conclude by saying that the hierarchical approach described above provides very good estimates and renders the PC prior a flexible tool both

True value	**MSE Posterior Mean**		
	PC		
	$\theta = 0.1$	$\theta = 1$	$\theta = 5$
n=5			
$\rho = -0.95$	0.009604	0.023882	0.329455
$\rho = -0.5$	0.104175	0.099008	0.160471
$\rho = 0$	0.104020	0.073082	0.010342
$\rho = 0.05$	0.101032	0.080895	0.013012
$\rho = 0.5$	0.128447	0.109913	0.157483
$\rho = 0.95$	0.010571	0.022708	0.332673
$\rho = 0.999$	1.80e-06	2.89e-06	0.008966
n=30			
$\rho = -0.95$	0.000199	0.000223	0.000451
$\rho = -0.5$	0.020827	0.021912	0.039885
$\rho = 0$	0.022266	0.027240	0.009467
$\rho = 0.05$	0.023473	0.022475	0.010823
$\rho = 0.5$	0.014371	0.019133	0.042093
$\rho = 0.95$	0.000247	0.000306	0.000403
$\rho = 0.999$	0.003439	1.02e-07	1.06e-07
n=100			
$\rho = -0.95$	5.21e-05	5.34e-05	6.17e-05
$\rho = -0.5$	0.004794	0.004485	0.007131
$\rho = 0$	0.010610	0.008295	0.005112
$\rho = 0.05$	0.009945	0.007657	0.006381
$\rho = 0.5$	0.005407	0.005948	0.007565
$\rho = 0.95$	4.78e-05	5.46e-05	7.23e-05
$\rho = 0.999$	2.03e-08	2.38e-08	2.46e-08

Table 3.1: Mean Squared Error computed over the posterior mean for different choices of the parameter θ

in estimation problems and in Bayesian Hypothesis testing.

3.9 Bayesian Hypothesis Testing

A natural way to select amongst models from a Bayesian point of view is to compute the Bayes factor. As pointed out by **?**, for a point null hypothesis testing, there may be a concern with the objective Bayesian approach. In fact, if the prior for the parameter of the alternative model has infinite variance, then the Bayes factor will alwalys select the null model, regardless the observed data. The aforementioned issue discourages Bayesians to use standard noninformative priors in the Bayesian hypothesis testing context. Then, in our specific case, the use of the Jeffreys' prior for ρ is not an adequate choice.
Therefore, many Bayesians use conventional proper priors in the calculation of the Bayes Factor. However, we know that for small values of θ, the penalised complexity prior approaches the Jeffreys' prior, but it is still proper.
So, our aim is to use our PC prior for ρ, where the intrisc prior with $\theta_0 = 0.491525$ is put on θ, in order to test the dependence among random variables, and to make a comparison we can consider only proper prior distributions; in particular we use the Arc-sine prior for ρ, that is a proper version of the Jeffreys' prior.
Suppose we wish to test the hypotheses

$$H_0 : \rho = 0 \qquad \text{vs} \qquad H_1 : \rho \neq 0$$

for the Gaussian copula model. Let $\gamma_0 = P(H_0)$ be the prior probability assigned to the null hypothesis and let $\pi^{PC}(\rho|\theta_0)$, where $\theta_0 = 0.491525$, be the prior distribution for ρ under the alternative model.
The invariance with respect to marginals of theorem **??** makes the joint prior, over the parameters of the marginals and the correlation parameter, a prior with separable components, and as a consequence allows us to write the Bayes Factor as follows

$$B_{01} = \frac{c_\rho(u, v; \rho)|_{\rho=0}}{\int c_\rho(u, v; \rho)\pi^{PC}(\rho|\theta_0 = 0.491525)d\rho},$$

which represents the evidence in favour of the null hypothesis with respect to the alternative. The decision on whether one rejects H_0 in favour of H_1 is based on the posterior probability, given by

$$P(H_0|x) = \left[1 + \frac{1 - P(H_0)}{P(H_0)} \frac{1}{B_{01}}\right]^{-1}.$$

Notice that we can also calibrate γ_0 by letting it depend on the variance of the PC prior, as the variance is a function of the rate parameter θ.
Here, it is worthwhile to make a digression. One could say that a sort of numerical integration of the PC prior in the vicinity of the base model should give the same information on γ_0. Actually, in this case, we think it is not recommendable to follow the aforementioned strategy, as the base model, here, has a particular meaning and we are not simply testing different values of the correlation parameter. In addition, if we perform hypothesis testing by using $P(H_0|x)$ directly, instead of computing the Bayes factor, we may lost the information intrinsic to

the Bayes factor, that is related to the asymmetric rates of convergence of B_{01} and B_{10}.
As we will see more in detail in the next chapter, the choice of a small value of θ will produce Bayes factors that support the alternative model, even when the true model is the null; on the contrary, by choosing a large value of θ, the Bayes factor will contract to one, for any true value of ρ, therefore no evidence for a specific model is provided.
Recall that now we are not considering as the shrinkage parameter the rate parameter θ, but we have assigned to it an intrinsic prior and we have fixed θ_0 in order to maximize the variance of the PC prior for ρ. Then, the issues above are mentioned just for the sake of completeness.
In Table ?? is shown the Mean Squared Error, the coverage probabilities and the Bayes Factor for each combination of (ρ^*, n). Notice that the MSE has been also shown in a different perspective in Figure ??. Now let us concentrate only on the Bayes factor.
The Bayes factor is computed by means of a simple Monte Carlo method; in particular a value of θ is drawn from the intrinsic prior, then it is plugged into the PC prior for ρ, then we simulate from this latter and we substitute the simulated values of ρ into the Gaussian copula density.
As we may see from Figure ?? the hierarchical approach with the choice of $\theta_0 = 0.491525$ makes the prior flat enough to be compared with the Arc-sine prior, but it still preserves some probability mass in correspondence of the base model. This seems to be a better stategy than setting an arbitrary value of θ, that, as we said, could lead to misleading conclusions about the dependence among marginals in the Bayesian hypothesis testing context.
The hierarchical approach has also the advantage to stabilize the variance of the PC prior, but, on the other hand, the main inconvenience could be to make too much shrinkage towards the base model, even though the objective solution aimed to maximize the variance of the PC prior for ρ prevent us to incur in such an issue.
The choice of consider the relative frequency of times that the Bayes factor is less than or equal to 0.5 (Table ??) is totally arbitrary and it is borrowed from ?.
It is worthwhile to recall that the BF_{01} goes to infinity when the true model is the independence one, whilst it goes to zero when the true model deviates from the independence assumption.
As we may see from Table ??, the frequency of times that $\text{BF}_{01} \leq 0.5$ is basically smaller for the PC prior compared to the Arc-sine prior, when the true model is $\rho = 0$. This is a good result for our prior, since for $\rho = 0$ the BF_{01} should go to infinity and as a consequence the frequency of times that $\text{BF}_{01} \leq 0.5$ should be small. On the other hand, we can notice that for slight correlations, like $\rho = 0.05$, the frequency of times that $\text{BF}_{01} \leq 0.5$ is bigger for the PC prior when compared to the Arc-sine prior. Once again this is a point in favour of our prior. This kind of behaviour has an obvious interpretation. For slight correlations, the PC prior with an intrinsic prior on θ is more able than the Arc-sine prior to catch the dependence among variables; this is because of the little spike of the PC prior in correspondence of the base model (see Figure ??). Even though, the objectivity we want to reach by maximizing the variance of the PC prior tends to spread out the probability mass of the PC prior, the hierarchical approach produces some shrinkage towards the base model, and this allows us to preserve the nature of the PC prior.

True value	MSE		Coverage		$BF_{01} < 0.5$	
	PC	Jeffreys	PC	Jeffreys	PC	Arc-sine
n=5						
$\rho = -0.95$	0.020897	0.006453	0.930	0.940	0.995	0.990
$\rho = -0.5$	0.120055	0.145870	0.930	0.940	0.140	0.155
$\rho = 0$	0.048071	0.129006	0.995	0.940	0.045	0.055
$\rho = 0.05$	0.056025	0.124518	0.980	0.940	0.070	0.060
$\rho = 0.5$	0.107815	0.118223	0.905	0.915	0.160	0.140
$\rho = 0.95$	0.024423	0.007614	0.945	0.935	0.990	0.985
$\rho = 0.999$	2.99e-06	1.77e-06	0.885	0.855	1.000	1.000
n=30						
$\rho = -0.95$	0.000282	0.000227	0.910	0.935	0.995	1.000
$\rho = -0.5$	0.026641	0.016513	0.930	0.930	0.780	0.765
$\rho = 0$	0.015009	0.032376	0.980	0.955	0.045	0.030
$\rho = 0.05$	0.014757	0.029801	0.980	0.950	0.020	0.020
$\rho = 0.5$	0.0343211	0.016206	0.880	0.960	0.770	0.845
$\rho = 0.95$	0.000254	0.000249	0.940	0.950	0.995	1.000
$\rho = 0.999$	9.33e-08	9.46e-08	0.870	0.850	1.000	1.000
n=100						
$\rho = -0.95$	6.09e-05	5.37e-05	0.945	0.955	0.995	1.000
$\rho = -0.5$	0.005481	0.004635	0.965	0.970	1.000	1.000
$\rho = 0$	0.005589	0.010714	0.985	0.950	0.005	0.030
$\rho = 0.05$	0.006555	0.009007	0.960	0.960	0.055	0.030
$\rho = 0.5$	0.006219	0.004583	0.930	0.960	1.000	1.000
$\rho = 0.95$	5.55e-05	6.09e-05	0.960	0.915	1.000	1.000
$\rho = 0.999$	2.25e-08	2.27e-08	0.890	0.890	1.000	1.000
n=1000						
$\rho = -0.95$	5.75e-06	4.50e-06	0.935	0.960	1.000	1.000
$\rho = -0.5$	0.000413	0.000487	0.980	0.950	1.000	1.000
$\rho = 0$	0.000820	0.000915	0.975	0.935	0.005	0.000
$\rho = 0.05$	0.001051	0.000919	0.925	0.955	0.145	0.090
$\rho = 0.5$	0.000511	0.000421	0.930	0.950	1.000	1.000
$\rho = 0.95$	4.65e-06	4.05e-06	0.960	0.970	1.000	1.000
$\rho = 0.999$	2.38e-09	1.81e-09	0.865	0.900	1.000	1.000

Table 3.2: The MSE, the coverage probabilities and the frequency of times that $BF_{01} < 0.5$ over 200 samples.

To conclude, we remark once again that the PC prior is more sensible than the Arc-sine prior in capturing small dependences and this is still valid as the sample size grows, as we may see from Table **??**.

3.10 Defining Objective Priors on the Models

Here, we propose a general strategy that could be used in order to avoid to incur in the Jeffreys-Lindley paradox. This latter arises when one tests a point null hypothesis and an objective prior is used for the alternative hypothesis. **?** proposed a solution to the paradox consisting of calibrate the prior probabilities assigned to each hypothesis with respect to the variance of the prior involved in the calculation of the Bayes Factor. In particular, the variance of our PC prior is regulated by the parameter θ. Recall that the PC prior is not an objective prior, but, as we have seen in the previous sections, we want it to be as objective as possible. For the sake of clarity we want to recall once again that our prior is proper, but nothing prevent us to adjust the prior probabilities on the models. To do that, we use the well known asymptotic Bayesian property that, if a model is misspecified, the posterior accumulates at the model which is the nearest, in terms of KLD, to the true model (Berk, 1966). As such, the divergence $\text{KLD}(c_\rho(u,v;\rho)\|c_{\rho_0}(u,v;\rho_0))$ represents the loss we would incur if model M_1 is removed and it is the true model. We can compute the expected loss as

$$\begin{aligned}
&\int_{-1}^{1} \text{KLD}\big(c_\rho(u,v;\rho)\|c_{\rho_0}(u,v;\rho_0)\big)\pi^{PC}(\rho|\theta)d\rho \\
=&\int_{-1}^{1} -\frac{1}{2}\log(1-\rho^2)\frac{\theta}{2}e^{-\theta\sqrt{-\log(1-\rho^2)}}\frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}d\rho \\
=&\frac{1}{\theta^2}.
\end{aligned} \tag{3.70}$$

The model prior is determined by means of the *self-information* loss function, which represents the loss connected to a probability statement. By equating the self-information with the expected loss we have that the prior on the alternative model is

$$1-\gamma_0(\theta) \propto e^{\frac{1}{\theta^2}}. \tag{3.71}$$

Note that the prior for the null hypothesis is $\gamma_0(\theta) \propto 1$, and so we have

$$\gamma_0 = \frac{1}{1+\exp\big(\frac{1}{\theta^2}\big)}, \tag{3.72}$$

and

$$1-\gamma_0 = \frac{\exp\big(\frac{1}{\theta^2}\big)}{1+\exp\big(\frac{1}{\theta^2}\big)}. \tag{3.73}$$

3.11 Danube data set

The `danube` data set contains ranks of base flow observations from the Global River Discharge project of the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC), a NASA data center. The measurements are monthly average flow rate for two stations situated at Scharding (Austria)

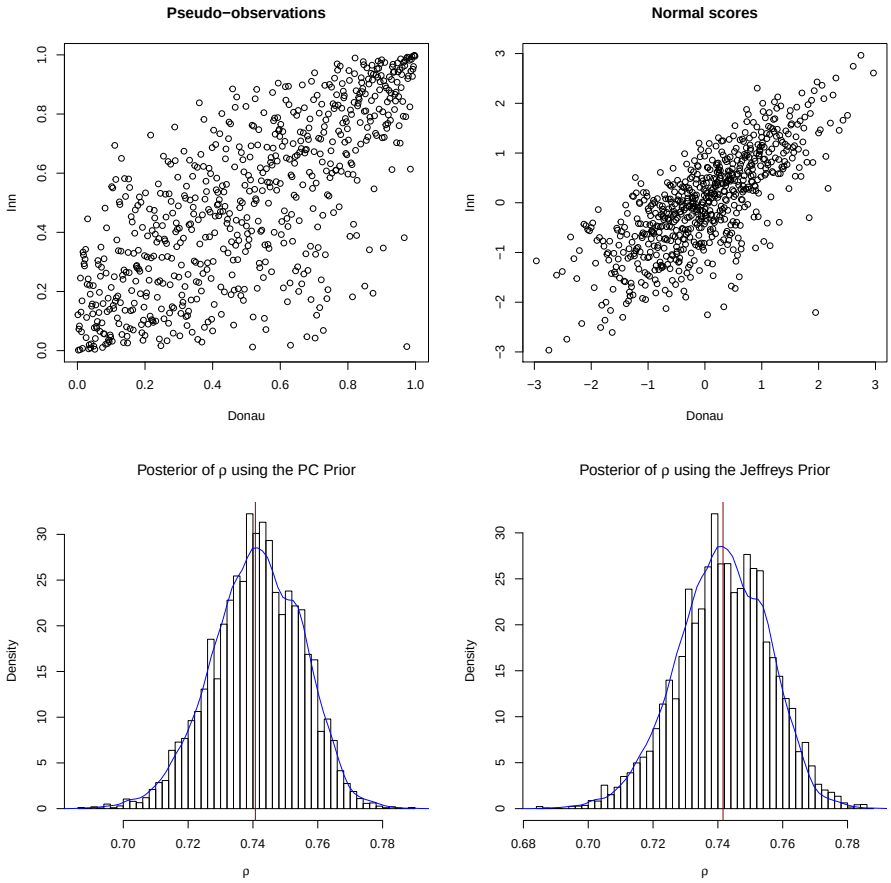

Figure 3.10: Pseudo-observations (top-left) and normal scores (top-right) of the danube data set. Posterior of ρ obtained by means of the PC prior with $\theta_0 = 0.491525$ (bottom-left) and by means of the Jeffreys' prior (bottom-right).

on the Inn river and at Nagymaros (Hungary) on the Danube.
The data have been pre-processed to remove any time trend. Specifically, **?** extracted the raw data, and obtain the fast Fourier transformed centered observations. The negative spectrum is retained and a linear time series model with 12 seasonal components is fitted. Then residuals are extracted.
Notice that the correlation between time series should be computed over the residuals and not over the original series, because in this latter case we would carry back correlation within the series. Figure **??** shows the pseudo-observations and the normal scores of the `danube` dataset, along with the posterior distribution of ρ derived by using both the PC prior with θ_0 maximizing the variance and the Jeffreys' prior. The vertical red line represents the posterior mean, while the estimated density is the blue curve, that is basically the same for both the priors.
From a frequentist point of view, we can carry out a dependence test based on the Spearmn's rank correlation, that is equal to 0.7374098; notice that it

is very close to the simulated posterior mean. Let us check out now if there is correspondence between the frequentist and the Bayesian approach. From the frequentist side, under the null hypothesis $H_0 : C = \Pi$ of independence between two random variables X and Y (where Π is the independence copula), the distribution of ρ_n, i.e. the Spearman's Rho, is close to normal with zero mean and variance $1/(n-1)$, so that one may reject H_0 at approximate level $\alpha = 5\%$, for instance, if $\sqrt{n-1}|\rho_n| > z_{\alpha/2} = 1.96$. So, the frequentist test based on the Spearman's rank correlation rejects the null hypothesis, as $n = 659$ and $\rho_n = 0.7374098$.
On the other hand, the Bayes factor computed by using both the PC prior with $\theta_0 = 0.491525$ and the Arc-sine prior is equal to $9.69112e-112$ and $6.87657e-112$, respectively, showing off a very strong evidence for H_1. So, there is concordance between the frequentist and the Bayesian approach, and this is because we are using proper priors; on the other hand, if we had used improper priors we would have incurred in the Jeffreys-Lindley's paradox.

3.12 Exchangeable Model

Suppose that the joint distribution given by the product of marginal distributions times a Gaussian copula, $\mathbf{X} = (X_1, \ldots, X_p)$ does not change under permutation of the indexes $1, \ldots, p$, i.e. each marginal is equally correlated with each other. In this case we have an exchangeable model characterised by an equicorrelation matrix:

$$\mathbf{R}(\rho) = \begin{bmatrix} 1 & \rho & \ldots & \ldots & \rho \\ \rho & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & \rho \\ \rho & \ldots & \ldots & \rho & 1 \end{bmatrix}, \tag{3.74}$$

where the main diagonal is a vector of ones, just because the Gaussian copula transforms random variables into standard normal ones, and the out-diagonal elements are all equal to ρ, supporting the same correlation among the variables. Notice that, even in this case, for $\rho = 0$ the Gaussian copula boils down to the independence copula. Recall that the Gaussian copula with correlation matrix $\mathbf{R}$ can be written as

$$C_{\mathbf{R}}(\mathbf{u}) = \Phi_{\mathbf{R}}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)), \tag{3.75}$$

where Φ^{-1} is the quantile function of a standard normal, $\Phi_{\mathbf{R}}$ is the CDF of a multivariate normal with zero mean vector and covariance matrix equal to the equicorrelation matrix $\mathbf{R}$, and p is the dimension of $\mathbf{R}$. The density function of a Gaussian copula can be written as

$$c_{\mathbf{R}}(\mathbf{u}) = \frac{1}{\sqrt{\det \mathbf{R}}} \exp\left(-\tfrac{1}{2} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_p) \end{pmatrix}^T \left(\mathbf{R}^{-1} - \mathbf{I}\right) \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_p) \end{pmatrix} \right), \tag{3.76}$$

where $\mathbf{I}$ is the identity matrix. It is clear that $\mathbf{R} = \mathbf{I}$ when $\rho = 0$ and by consequence $c_{\mathbf{R}}(\mathbf{u}) = 1$. In other words, for $\rho = 0$ the exchangeable model boils

down to the base model characterised by independence among the variables.
Let $M_1 = \{f_{\mathbf{X};\mathbf{R}}(\mathbf{x};\rho), \mathbf{x} \in \mathbb{R}^p\}$ be the complex model where the variables are mutually correlated, with correlation parameter equal to ρ, and let $M_0 = \{f_{\mathbf{X};\mathbf{I}}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^p\}$ be the base model, where with abuse of notation we omit the parameters of the marginal densities since we have demonstrated that the prior for ρ is invariant to these parameters. Let f_1 be the density of M_1 and f_0 be the density of the base model. Then, we can write down the Kullback-Leibler divergence as

$$\mathrm{KLD}(f_1 \| f_0) = \frac{1}{2}\left[\mathrm{trace}\big(\mathbf{I}^{-1}\mathbf{R}\big) - p - \log\left(\frac{|\mathbf{R}|}{|\mathbf{I}|}\right)\right], \tag{3.77}$$

where notation $|\cdot|$ stands for the matrix determinant. The distance from the simpler model is expressed as $\sqrt{2\mathrm{KLD}(f_1 \| f_0)}$ and it is given by

$$d(\rho) = \sqrt{-\log(|\mathbf{R}|)}. \tag{3.78}$$

In order to comply with the constant decay-rate assumption, we assign the exponential distribution in (**??**) to the disatnce scale in (**??**), where θ plays a key role as it controls the shrinkage of the prior towards the base model. Even in this case, the distance has two branches, but now they are no longer symmetric in the shape, as we will see later on.
For the exchangeable model, we derive a general formula for the distance in (**??**) for any dimension p of the equicorrelation matrix

$$d(\rho) = \sqrt{-\log(\mathrm{sgn}(p)(\rho-1)^{p-1}((p-1)\rho+1))}, \tag{3.79}$$

where

$$\mathrm{sgn}(p) = \begin{cases} -1 & \text{if } p \text{ is even} \\ +1 & \text{if } p \text{ is odd} \end{cases}. \tag{3.80}$$

The distance function above is very similar to the one obtained by **?** for the residuals of the one-factor mixed models.
For instance, the distance from the base model for three equicorrelated marginals ($p = 3$) is equal to:

$$d(\rho) = \sqrt{-\log(2\rho^3 - 3\rho^2 + 1)}, \tag{3.81}$$

and this latter is defined only for $\rho \in [-1/2, 1]$. This is because the equicorrelation matrix $\mathbf{R}$ is positive-definite if and only if $-\frac{1}{p-1} < \rho < 1$. The three random variables with equal correlation ρ can sure have correlation equal to $+1$, but obviously not equal to -1. Consider three random variables X, Y, Z where $\rho_{XY} = \rho_{XZ} = -1$, then ρ_{YZ} cannot be equal to -1, but in fact it is equal to $+1$. The minimum value of the common correlation that the random variables can achieve is $-\frac{1}{2}$. In general, the smallest value of common correlation of p random variables is $-\frac{1}{p-1}$, and we can think of this situation as a simplex of dimension $p-1$, in the p-dimensional space, where the top vertices are -1 and the bottom vertex is $+1$.
A trivial proof can be given by looking at the variance of p unit variance random

variables X_i. We have that

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{p} X_i\right) &= \sum_{i=1}^{p} \text{Var}(X_i) + \sum_{i=1}^{p}\sum_{j\neq i}^{p} \text{Cov}(X_i, X_j) \\ &= p + \sum_{i=1}^{p}\sum_{j\neq i}^{p} \rho_{X_i,X_j} \\ &= p + p(p-1)\bar{\rho}, \end{aligned} \tag{3.82}$$

where the second equality comes from the fact that the variables have unit variance. Obviously, $\text{Var}\left(\sum_{i=1}^{p} X_i\right)$ must be ≥ 0, so we can write

$$\bar{\rho} \geq -\frac{1}{p-1}, \tag{3.83}$$

but if all the variables have correlation ρ, the average value $\bar{\rho}$ also equals ρ, so we obtain

$$\rho \geq -\frac{1}{p-1}. \tag{3.84}$$

Even if it may seem counterintuitive, the random variables can achieve common negative correlation equal to $-\frac{1}{p-1}$.
Then, the resulting PC prior for $p = 3$ is

$$\pi(\rho|\theta) = \frac{\theta}{2} e^{-\theta\sqrt{-\log(2\rho^3-3\rho^2+1)}} \frac{3|\rho|}{(1-\rho)(2\rho+1)\sqrt{-\log(2\rho^3-3\rho^2+1)}}, \tag{3.85}$$

where $\rho \in [-\frac{1}{2}, 1]$. In this case the parameter space of the negative hand side is halved but the prior mass associated to negative correlations is still the fifty percent. This is a consequence of the fact that we assign half an exponential to each branch of the distance function.
More generally, we can write down the PC prior density function for a generic dimension p, that is

$$\begin{aligned}\pi(\rho|\theta) = &\frac{\theta}{2} \exp(-\theta\sqrt{-\log((1-\rho)^{p-1}(1+(p-1)\rho)}) \\ &\cdot \frac{p(p-1)}{2} \frac{|\rho|}{(1-\rho)(1+(p-1)\rho)\sqrt{-\log(1-\rho)^{p-1}(1+(p-1)\rho)}}, \end{aligned} \tag{3.86}$$

where $\rho \in [-\frac{1}{p-1}, 1]$. It is important to notice that even if the prior is not symmetric in the shape, $\text{Prob}(\rho < 0) = \text{Prob}(\rho > 0) = 0.5$, so it is symmetric in the sense that the median is at the base model. It is also worth to mention that the limit of the prior towards the base model for a general dimension p is

$$\lim_{\rho\to 0} \pi(\rho|\theta) = \theta\sqrt{\frac{p(p-1)/4}{2}}, \tag{3.87}$$

and it is straightforward to see that for the bivariate case the limit is $\frac{\theta}{2}$.
We can appreciate from Figures ?? and ?? how the lower bound moves towards zero, in particular it equals $-\frac{1}{4}$ for 5 equicorrelated marginals. As we said above, despite of this movement towards zero, the prior mass remains the same for both

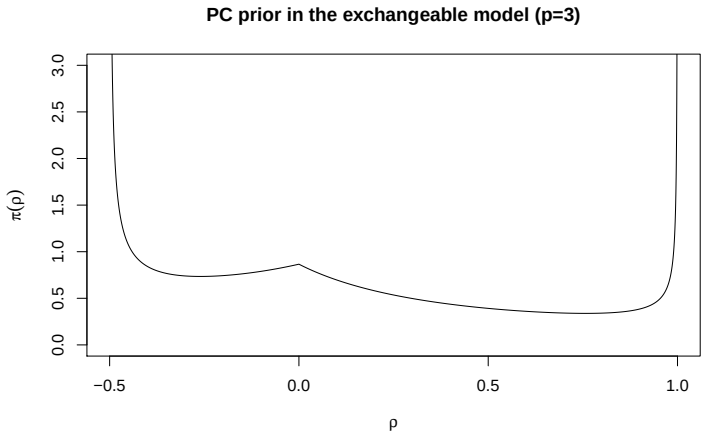

Figure 3.11: PC prior for ρ in the exchangeable model with 3 marginals and scaling parameter $\theta = 1$.

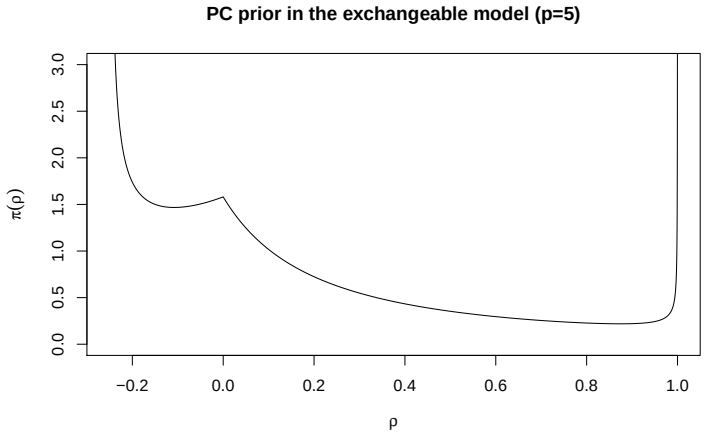

Figure 3.12: PC prior for ρ in the exchangeable model with 5 marginals and scaling parameter $\theta = 1$.

negative and positive correlations.
As a final comment, I would like to make a clarification. One could debate the reason why we always consider as the base model the model where $\rho = 0$. Actually, other base models may be taken into account. For instance, **?** consider for the autoregressive model both the base model of no dependency in time ($\rho = 0$) and the base model of no change in time ($\rho = 1$), while **?** still consider the case of $\rho = 1$ for varying coefficient models. Actually, in a Gaussian copula framework, by setting the value of $\rho = 1$, we would not have the extra-component to disappear in the base model. Apart from that, in the latter case, the Gaussian copula density function will be zero, and as a consequence the multivariate density at the base model will be zero.
Anyhow, it could be interesting to check for other copulas if it is possible to consider limiting choices of the association parameter different from the case of independence.

Chapter 4

PC Priors with Gaussian Base Model

The Gaussian distribution has ever been the most popular and usable device in the field of statistics. Even in the context of the penalised complexity priors, the normal density has a particular meaning, especially because we can consider it as an elastic base model which could be extended both in terms of tail thickness and skewness. As claimed by **?**, PC priors have been constructed for hierarchical models. So, when deviating from normality, a wider model has to be taken into account and a prior distribution on the additional model component has to be specified. The choice to employ a principled prior that penalizes the distance between two nested models looks natural in the hierarchical modelling.
Although there is no subjective aspect to be found in the construction of a PC prior, one could debate the choice of the distance measure, even though the KLD is an appropriate measure based on a relative entropy concept and with a lot of attractive properties, or, even more interestingly, one could criticise the assignment of an exponential distribution to the distance scale, that seems to be a reasonable choice, especially for mathematical purposes.
The exponential choice ensures an equal penalisation rate of the distance, without taking into account the position we are in the parameter space. For instance, one could have a different "belief" of the penalisation of the distance, so that the prior turns out to reflect the subjective perception of what a penalisation idea means. To the best of our knowledge the exponential distribution is the only continuous distribution to have the property of the constant-decay rate, and in absence of any other specific knowledge on the distance scale, it seems to be an obvious although not objective choice.
The base model being a normal distribution has a lot of possible extensions. The first one is in terms of location-scale. For instance, if we want to calculate the PC prior for the location parameter, we just need to compute the distance between the two models.
Here, it is worthwhile to make a clarification. Sure, the mean is a property of the Gaussian distribution, therefore it is present in any case, even when it is equal to zero, but on the basis of our view of the base model, when the parameter $\mu = 0$ the density at the denominator of the KLD will have a parameter less. In our vision, the essence of the building blocks construction is coherent

with the assumption that the extra-component is not present in the base model. Obviously the base model could be intended even for a value of μ different from zero.
In the location-scale example, the Kullback-Leibler divergence has an analytical expression, say equal to $\mu^2/2$, then the PC prior is easily computed. It turns out that for the location parameter, the PC prior is simply a double exponential distribution.
We think it is worth to mention that the penalisation of the positive and negative locations would be different if we adopted different rate parameters in the exponential distribution assigned to the distance. Even for the scale parameter, the prior could be penalised differently, it depends on what we believe of small variances and large variances. The PC prior for the scale parameter σ is clearly not symmetric by looking at its shape, given that the base model is not in the middle of the parameter space, but, as we will see, a different concept of symmetry is embedded in the PC prior.
Our definition of symmetric PC prior relies on the assumption to give the same weight to each of the two chunks of the parameter space where the distance function is monotone. So, from this point of view, the PC prior is symmetric in the sense that the median is at the base model. In practice, this means that we assign the same probability mass to different parts of the parameter space.

Definition 1 (Symmetry of the PC prior)**.** A PC prior whose base model is at the interior of the parameter space is said to be symmetric if the median occurs at the base model.

We will explain later that there exists also an asymmetric interpretation of the PC prior whose base models are not at the boundaries of the parameter space.
A different problem occurs when we want to compute the prior both for the location and scale parameters, since, in this case, we must specify a joint PC prior as the Kullback-Leibler divergence is a function of both the parameters. Nevertheless, the normal base model needs not to be the standardised one, generally as each component can be added one-at-the-time and, in this case, the location and scale parameters are invariant between each other.
Moreover, the location-scale extension is not the only one that could be taken into account. Let us think, for instance, of a skew-normal distribution that is constructed by perturbating the symmetry of a normal density, and therefore a skewness parameter is added, or, on the other hand, a Student-t distribution, that extends the normal density by allowing for different amount of kurtosis. In these particular cases the Kullback-Leibler divergence turns out to be invariant with respect to the location and scale parameters. We are claiming that, even considering a non-standard Gaussian distribution as a base model, the distance depending on the additive model component, which could be a shape as well as a degree of freedom parameter, is not a function of the location and scale parameters. This is a remarkable fact as it allows us to derive the prior on the additive model component in an independent way with respect to the location-scale structure.
The more interesting thing is that, in practice, the PC prior on the additional model component does not mix with the joint prior for μ and σ, then we can look at them in a separate way. As we will see, the latter property plays a key role in the use of PC priors in Bayesian hypothesis testing, as the Bayes

factor can be computed without considering the prior for the location and scale parameters.
Related to the joint PC prior concept, let us recall that the introduction of either location and scale parameters one-at-the-time (let's think of something evoking a conditional PC prior) does not make the distance a bivariate function, while a different problem is to add the two components in one block.

4.1 Extension of the standard Gaussian Base Model

As we said, the standard Gaussian base model is very flexible, in the sense that it can be extended in multiple directions. The true interesting thing in this case is that, for this particular base model, the Kulback-Leibler divergence has a specific representation in terms of the entropy and the second moment of the more complex model. We will see later how this alternative specification of the KLD has been helpful in the derivation of the PC prior for the degrees of freedom of a Student-t distribution.
In particular, it is well-known that the Kullback-Leibler divergence between two normal distributions has a closed-form expression, but in cases where the integral is hard to solve, we can take advantage of the following alternative specification of the KLD.

Theorem 2 (Alternative KLD for the Gaussian base model)**.** *Suppose to have a standard normal variate whose density function is f, and a random variable, Y, with a more flexible distribution, g. Then, the KLD between any model that is built up by adding a component to the standard normal base model and the standard normal distribution itself can be expressed as*

$$\mathrm{KLD}(g\|f) = -H(Y) + \frac{1}{2}\left(\mathbb{E}(Y^2) + \log(2\pi)\right), \tag{4.1}$$

where $H(\cdot)$ stands for the entropy.

Proof. The KLD between g and f is defined as

$$\begin{aligned}
\mathrm{KLD}(g\|f) &= \int g \log\left(\frac{g}{f}\right) dy \\
&= \int g \log g \, dy - \int g \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)\right) dy \\
&= -H(Y) - \left(-\frac{1}{2}\int y^2 g \, dy + \log\left(\frac{1}{\sqrt{2\pi}}\right)\right) \\
&= -H(Y) + \frac{1}{2}\mathbb{E}(Y^2) + \log(\sqrt{2\pi}),
\end{aligned} \tag{4.2}$$

and by simply rearranging the last term, the proof is completed. □

Notice that the rule also holds for the non-standardised normal base model, but, in this case, arrangements inherent to location and scale parameters must be taken into account.
Now, let us show some examples. Henceforth, consider the well-known Kullback-Leibler divergence between two normal distributions

$$\mathrm{KLD}(N_1\|N_0) = \frac{1}{2}\left(\left(\frac{\sigma_1}{\sigma_0}\right)^2 + \frac{(\mu_0-\mu_1)^2}{\sigma_0^2} - 1 + 2\log\left(\frac{\sigma_0}{\sigma_1}\right)\right). \tag{4.3}$$

For instance, suppose we want to calculate the PC prior for μ (where the base model is the zero mean and unit variance normal density), then the KLD is equal to $\frac{\mu^2}{2}$ and the distance, $d(\mu) = \mu$. It is obvious to see that the PC prior for μ is a double exponential distribution, but as we have noticed above, this relies on the fact that we are giving the same weight to both the positive and negative parts of the parameter space where the distance function is monotone. The latter concept can be seen as a further principle in the construction of penalised complexity priors as it allows the user to introduce a different type of information, this time based on the skewness reflected into the PC prior distribution.
So, it is possible to build up skew-PC prior by simply giving different weights to the pieces of the parameter space where the KLD is monotone. There are situations where it makes sense to give different weights to the exponential densities that we use to penalise each branch of the distance function. We stress the fact that the double exponential prior for μ comes up from the choice to give the same weight, say $1/2$, to the positive and negative locations in the parameter space. Notice that this is not the problem of a different penalisation, that is related instead to the rate parameter of the exponential distribution and could occur in any case, even for skew-PC priors.
Let's go more deeply into the concept above. Suppose to give weight ω to the negative locations and, as a consequence, $1-\omega$ to the positive ones. Let θ and λ be the rate parameters of the exponential distributions assigned to the negative and positive locations, respectively. The limits towards the base model would be $\omega\theta$ for the negative locations and $(1-\omega)\lambda$ for the positive ones. Now, if $\omega = (1-\omega) = 1/2$, the left-hand limit will coincide with the right-hand limit only if $\theta = \lambda$, otherwise they would be different. Anyhow, if $\omega \neq (1-\omega)$, the left-hand limit will coincide with the right-hand one only if $\omega = \frac{\lambda}{\theta+\lambda}$. So, we could introduce some skewness in our PC prior by simply regulating the mass we assign to the different monotone branches of the distance function in the parameter space, and certainly it remain the possibility to penalise differently each branch, also for this skewed version of the PC prior.
In the present work, we will not consider this asymmetric version of the PC prior, nonetheless we think is worth to mention that the construction of a PC prior, as in **?**, encloses an implicit principle based on the symmetry of the PC prior distribution to be derived. We repeat once again that our definition of symmetry means that the median is at the base model and not that the prior has a symmetric shape, which could also happen. Notice also that by assigning different weights to different parts of the parameter space, we will no longer have the median at the base model.
We have seen that the PC prior for μ is a double exponential distribution, but we want to check now that the same KLD can be derived by using formula (**??**). Firstly, let us have a look at the normal entropy, that is the maximum entropy across the families of probability density functions. The normal entropy is obviously not depending on the location parameter μ, and it is equal to $\frac{1}{2}\log\left(2\pi e\sigma^2\right)$, or equivalently, to $\frac{1}{2}\log\sigma^2 + \frac{1}{2}\left(1+\log\left(2\pi\right)\right)$.
Then, we can write down the KLD by means of theorem **??** in the following way

$$\mathrm{KLD}(N_1\|N_0) = -\frac{1}{2}\log\left(2\pi e\sigma^2\right) + \frac{1}{2}\left(\sigma^2+\mu^2\right) + \frac{1}{2}\log\left(2\pi\right). \qquad (4.4)$$

It turns out that, for the additional component being just μ, so that σ^2 is left

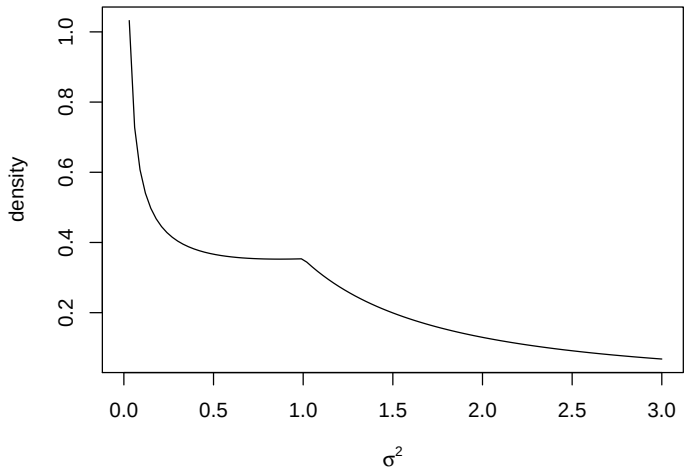

Figure 4.1: The PC prior for σ^2 with $\theta = 1$.

to be equal to 1, the KLD, as expected, is $\frac{\mu^2}{2}$, and consequently $d(\mu) = \mu$.
On the other hand, if we want to calculate the PC prior for σ^2, it is immediate to see that for $\mu = 0$, the KLD is equal to $\frac{1}{2}\left(\sigma^2 - \log\sigma^2 - 1\right)$, and obviously the same comes out from (??).
Given the KLD, we can easily compute the PC prior for σ^2, whose density function is

$$\pi^{PC}(\sigma^2) = \frac{\theta}{2} e^{-\theta\sqrt{\sigma^2 - \log\sigma^2 - 1}} \frac{|\sigma^2 - 1|}{2\sigma^2\sqrt{\sigma^2 - \log\sigma^2 - 1}}. \tag{4.5}$$

Returning back to the symmetry of the PC prior, we recall once again that the probability density function in (??) is symmetric in the sense that the median is at the base model. Nonetheless, it is evident how the shape of the prior distribution is not symmetric (see figure ??). As we said many times, this is a consequence of the fact that we assign half an exponential distribution to each branch of the distance function being monotone, even though, in the left-hand side, σ^2 is in $(0, 1)$, while, in the right-hand side, is in $(1, \infty)$.
Obviously, one could relax this condition and, as we said before, we would end up in an asymmetric version of the PC prior. The asymmetric PC prior can be useful for instance in certain long memory processes where the user has a prior belief in favour of a persistent or an anti-persistent process, like in the fractional Gaussian noise (see ?).

Figure ?? resumes what we said above. The shape is not symmetric at all, even though, considering our concept of symmetry, the distribution is symmetric, given that the median is at the base model. Figure ?? shows that for an increasing θ the shape of the prior distribution becomes more and more symmetric, as the shrinkage parameter makes all the mass going towards zero.
The KLD construction according to theorem ?? obviously holds also for the

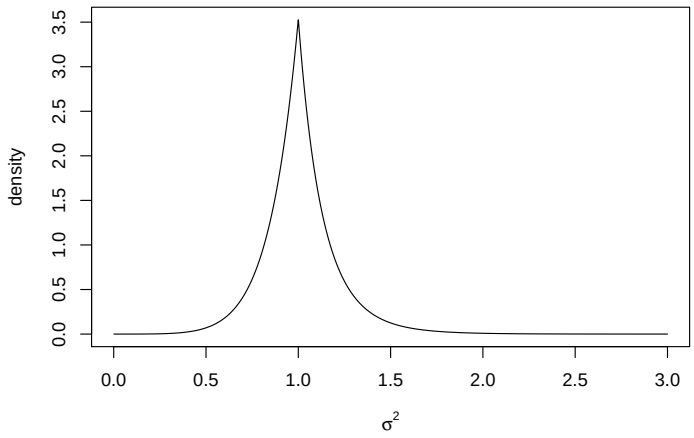

Figure 4.2: The PC prior for σ^2 with $\theta = 10$.

skew-normal model as it allows us to write the KLD in the following manner

$$\begin{aligned} \text{KLD}(\delta) &= -H_{SN(\delta)} + \frac{1}{2}\left(1 - 2\frac{\delta^2}{\pi} + \delta^2\frac{2}{\pi}\right) + \frac{1}{2}\log(2\pi) \\ &= -H_{SN(\delta)} + \frac{1}{2}\left(1 + \log(2\pi)\right), \end{aligned} \tag{4.6}$$

where $H_{SN}(\delta)$ is the skew-normal entropy, and it is only function of the shape parameter δ. So, for the sake of simplicity, we consider the skew-normal entropy where the location and scale parameters are respectively equal to 0 and 1, as the PC prior for the shape parameter of a skew-normal variate is invariant to the location-scale structure.
In **?** it is shown that the skew-normal entropy can be built up starting from the normal entropy as follows

$$H_{SN(\mu,\sigma^2,\alpha)} = H_{N(\mu,\sigma^2)} - E_{X_0}\Big[\log\left(2\Phi\left(\alpha X_0\right)\right)\Big], \tag{4.7}$$

where $X_0 \sim SN(\alpha)$, $H_{N(\mu,\sigma^2)} = \frac{1}{2}\log\sigma^2 + \frac{1}{2}\left(1 + \log(2\pi)\right)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. By using this parameterization, α represents the shape parameter taking value over all the real line, but there exists also an alternative parameterization, in δ, which bounds the support of the skewness parameter between -1 and 1.
Given the invariance to location and scale parameters, the computation of the KLD according to theorem (**??**), can be done by using the standard version of the entropy

$$H_{SN(0,1,\alpha)} = H_{N(0,1)} - E_{X_0}\Big[\log\left(2\Phi\left(\alpha X_0\right)\right)\Big] \tag{4.8}$$

where $H_{N(0,1)} = \frac{1}{2}\left(1 + \log(2\pi)\right)$. So, if $\alpha = 0$, the skew-normal and normal entropies coincide. Then, the KLD (either in terms of δ and α) is equal to zero as the two terms in (**??**) cancel out. On the other hand, the KLD is

equal to $E_{X_0}\big[\log\left(2\Phi\left(\alpha X_0\right)\right)\big]$ when $\alpha \neq 0$, and this can be simply checked by plugging (??) in equation (??). We will see in the next session that the same result comes out by solving the integral in the canonical way, i.e. without using the KLD provided in theorem ??.

4.2 The PC Prior in the Skew-Normal Model

In many practical statistical works, datasets reveals departures from symmetry, hence something more flexible than the normal model is needed. The skew-normal distribution extends the normal one by introducing in the cumulative distribution function a perturbation parameter that accounts for skewness. The probability density function of a scalar skew-normal random variable X is of the form

$$f(x;\mu,\sigma,\lambda) = \frac{2}{\sigma}\,\phi\Big(\frac{x-\mu}{\sigma}\Big)\,\Phi\Big(\lambda\frac{x-\mu}{\sigma}\Big), \quad x \in \mathbb{R}, \quad \lambda \in (-\infty,+\infty), \tag{4.9}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the standard Gaussian pdf and CDF.
We can look at the skew-normal model as a flexible version of the normal one, where the latter represents the base model. In fact, for a particular value of λ, i.e. $\lambda = 0$, the density (??) boils down to the normal density, since $\Phi(0) = 1/2$.
A peculiar feature of the skew-normal distribution is that sometimes it does not have the maximum likelihood estimator of the shape parameter in the interior of the parameter space, especially in small to moderate sample sizes and when the absolute value of λ is large. Therefore, other approaches are needed and the use of Bayesian methods has flourished. ? has analysed the problem of the boundary estimates of the MLE in the case in which there is only the shape parameter, i.e. a random sample $z = (z_1, \ldots, z_n)$ is drawn from a random variable $Y \sim \mathrm{SN}(0,1,\lambda)$. So, when the sample size is small and λ deviates enough from zero, the log-likelihood function of the skew-normal distribution have a large probability to have the maximum at the boundaries of the parameter space.
The likelihood function

$$L(\lambda) = \text{constant} \times \Phi(\lambda z_1) \times \cdots \times \Phi(\lambda z_n) \tag{4.10}$$

is a strictly monotonic function if all the sample draws have the same sign. In this case the MLE is $\hat{\lambda} = \pm\infty$. The probability of ending up in such a sample with the pattern discussed above is

$$p_{n,\lambda} = P(Z_1 < 0)^n + P(Z_1 > 0)^n = \left(\frac{1}{2} - \frac{\arctan\lambda}{\pi}\right)^n + \left(\frac{1}{2} + \frac{\arctan\lambda}{\pi}\right)^n, \tag{4.11}$$

which goes to 0 as $n \to \infty$, provided $|\lambda| < \infty$, even though this probability is appreciable for small sample sizes. If location and scale parameters are included in the model, the log-likelihood function is a three-parameters function and the boundary MLE can still occur but in this case we don't know what is the probability leading to such a sample.
The frequentist approach has several problem in making inference about the parameters of the SN distribution. Mainly, there are three problems

- The MLE for λ can be infinite;

- the Fisher information matrix is singular when $\lambda = 0$;
- there can be local maximum in the ML function.

Trying to solve the second problem, **?** proposed to use a different parameterisation (centered parameterisation), based on the centered parameters that are represented by the mean, the variance and the skewness index of the SN distribution. Although the solution proposed by **?** is suitable for solving the second problem, it doesn't work for the first problem. Therefore, **?** proposed a method to avoid boundary estimates, based on a modification of the likelihood equation, on the footstep of **?** who advanced a general bias-reduction technique. In fact, the occurrence of $|\hat{\lambda}| = \infty$ with non-null probability produces maximal bias. **?** proved that the estimator based on a modified likelihood equation is always finite and it is the solution of the equation

$$S(\tilde{\lambda}) + M(\tilde{\lambda}) = 0, \tag{4.12}$$

where S is the usual score function and

$$M(\lambda) = -\frac{\lambda}{2}\frac{a_4(\lambda)}{a_2(\lambda)}, \tag{4.13}$$

where, in turn, $a_k = E\left[Z^k \left(\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\right)^2\right]$, $k = 2, 4$ and the expected values are calculated with respect to the standard skew-normal distribution and must be numerically computed.
In practice, the Firth's method replaces the usual score function $S(\lambda)$ by $S^*(\lambda) = S(\lambda) - I(\lambda)\, b(\lambda)$, and solve for $S^*(\lambda) = 0$. In this specific case, λ is meant to be any parameter whose MLE is biased. In fact, $b(\lambda)$ is the bias of the MLE and $I(\lambda)$ is the expected Fisher information.
? applied this scheme to a random sample $z = (z_1, \ldots, z_n)$ from $\mathrm{SN}(0, 1, \lambda)$, then the equation to solve takes the form

$$\sum_{i=1}^{n} \zeta_1(\lambda z_i) z_i - M(\lambda) = 0, \tag{4.14}$$

where $\zeta_1(\lambda z_i) = \frac{\phi(\lambda z_i)}{\Phi(\lambda z_i)}$ and $M(\lambda) = I(\lambda) b(\lambda)$ is defined in (**??**). Sartori has proved that (**??**) admits a finite solution, although he didn't give a proof of the uniqueness of the solution. A remarkable fact is that, when dealing with full exponential families, the correction term of the score function is equal to the term produced by the employment of the Jeffreys' prior distribution for λ in a Bayesian context, but this is not the case.
As we said above, the frequentist approach based on the Maximum Likelihood Estimation may fail when one wants to estimate the skewness parameter of the skew-normal distribution, therefore Bayesian methods have been employed in the recent years.
? proposed a noninformative Bayesian analysis of the SN model under the reference prior, on the basis of the method proposed by **?** (BB reference prior). For the uniparametric model, the BB reference prior agrees with the Jeffreys' prior. Then, the BB reference prior, or Jeffreys' prior, for the skew-normal model is given by

$$\pi^J(\lambda) \propto \sqrt{\int_0^\infty 2x^2 \phi(x) \frac{\phi^2(\lambda x)}{\Phi(\lambda x)(1 - \Phi(\lambda x))} dx}. \tag{4.15}$$

? also showed that

- $\pi^J(\lambda)$ is symmetric around 0 and decreasing in $|\lambda|$;
- the tails of $\pi^J(\lambda)$ are of order $O(\lambda^{-\frac{3}{2}})$;
- the posterior maximum is finite;
- the posterior mean is infinite for the samples for which the MLE is infinite.

The density in (??) is a rare case of proper Jeffreys' prior, even though the support of λ is unbounded. Typically, Jeffreys' priors are improper even in bounded parametric spaces, so this is a peculiar exception that allows us to perform default Bayesian hypothesis testing by using such a prior distribution. The prior distribution in (??) is obtained from the Fisher information, i.e. $\pi^J(\lambda) \propto I^{\frac{1}{2}}(\lambda)$, and it cannot be written in closed form as also the Fisher information cannot (see ?). Despite of the latter inconvenience the Jeffreys' prior for λ of the SN model can be approximated very well by a Student-t distribution centered in zero with scale $\pi/2$ and one half degree of freedom, $t(x; 0, \frac{\pi}{2}, \frac{1}{2})$.
? noticed that

$$\frac{\phi(x)}{\sqrt{\Phi(x)(1-\Phi(x))}} \approx 2\phi(2x/\pi), \tag{4.16}$$

and using this result they obtained an approximation of the Fisher information as follows

$$I(\lambda) \approx \frac{2}{\pi}\left(1 + \frac{2\lambda^2}{\pi^2/4}\right)^{-\frac{3}{2}}. \tag{4.17}$$

It ensues that the prior in (??) can be approximated by

$$\pi^J(\lambda) \approx \sqrt{\frac{2}{\pi}}\left(1 + \frac{2\lambda^2}{\pi^2/4}\right)^{-\frac{3}{4}}, \tag{4.18}$$

that is actually a $t(0, \pi/2, 1/2)$.
? extended the previous results to the shape parameter of a skew-t distribution. In practice, the framework is the same; we only need to replace in (??) the normal pdf and CDF with the ones of the Student-t distribution, i.e. $\phi(x)$ and $\Phi(x)$ becomes respectively $t_\nu(x|\nu)$ and $T_\nu(x|\nu)$, where ν denotes the degrees of freedom of the Student-t distribution. Even in this case the approximation works very well, but in general it becomes more and more accurate when the value of ν increases.
In the following sections we will derive the PC prior for λ, based on the penalization of the skew-normal model associated with the introduction of a perturbation parameter in the Gaussian density, and then we will perform Bayesian inference and Hypothesis testing by using such a PC prior .

4.2.1 Invariance of the PC Prior wrt location and scale parameters

Before going on, it is worthwhile to introduce a very useful concept. An important feature of the PC prior for λ is the invariance with respect to the location and scale parameter values. Roughly speaking, although the location and scale parameters could take any value, they do not affect the KLD when only the

skewness component is added to the more complex model. Suppose we introduce μ and σ in a standard skew-normal distribution, and then we compute the divergence between (??) and $\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$, where the latter represents the base model, or in other words a non standardised Gaussian distribution. Then, we can state the following proposition.

Proposition 1 (Invariance wrt the location and scale parameters)**.** Let $X_1 \sim SN(\mu, \sigma^2, \lambda)$ and $Y_1 \sim N(\mu, \sigma^2)$ be the skew-normal and normal densities, respectively, with the same location and scale parameters. Furthermore, let $X_2 \sim SN(0, 1, \delta)$ and $Y_2 \sim N(0, 1)$ be the standard versions of the above densities. The Kullback-Leibler divergence between X_1 and Y_1 does not differ from the one between X_2 and Y_2. In other words, the resulting PC prior for λ does not depend on μ and σ.
Indeed, with a change of variable, the integral

$$\int_{\mathcal{X}} \frac{2}{\sigma}\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big)\Phi\Big(\lambda\frac{x-\mu}{\sigma}\Big)\log\Big(2\Phi\Big(\lambda\frac{x-\mu}{\sigma}\Big)\Big)dx, \tag{4.19}$$

can be written as

$$\int_{\mathcal{T}} 2\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\Phi\Big(\lambda t\Big)\log\Big(2\Phi\Big(\lambda t\Big)\Big)dt, \tag{4.20}$$

where $t = \frac{x-\mu}{\sigma}$ and $dt = \frac{dx}{\sigma}$.

Thus, the distance between skew-normal and normal densities with the same location and scale parameters is just a function of the skewness parameter.
It seems redundant to mention that the same result holds in the case in which we specify the skew-normal density in terms of δ, instead of λ, but, for the sake of completeness, we remark that

$$\int_{\mathcal{X}} \frac{2}{\sigma}\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big)\Phi\Big(\lambda(\delta)\frac{x-\mu}{\sigma}\Big)\log\Big(2\Phi\Big(\lambda(\delta)\frac{x-\mu}{\sigma}\Big)\Big)dx, \tag{4.21}$$

is rewritable as

$$\int_{\mathcal{T}} 2\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\Phi\Big(\lambda(\delta)t\Big)\log\Big(2\Phi\Big(\lambda(\delta)t\Big)\Big)dt, \tag{4.22}$$

where $\lambda(\delta) = \frac{\delta}{\sqrt{1-\delta^2}}$.

4.2.2 Construction of the PC Prior for the shape parameter

Given the aforementioned result, we can deal simply with a standard skew-normal distribution, without taking into account μ and σ.
Suppose we have a standard skew-normal distribution, i.e. where the location and the scale parameters are equal to 0 and 1 respectively. In other words, we have a density of the form

$$f_X(x; 0, 1, \lambda) = 2\ \phi(x)\ \Phi(\lambda x) \tag{4.23}$$

where for $\lambda = 0$ we retrieve the symmetric counterpart that is represented by the standard normal density $\phi(x)$.

Recall that the Kullback-Leibler divergence is just a measure of the information lost when a base model is used to approximate a more flexible model and therefore it is not a metric.
The KLD between densities (**??**) and the standard normal $\phi(x)$ is

$$\text{KLD}(f_{X;\lambda} \| f_{X;0}) = \int_{\mathcal{X}} 2\ \phi(x)\ \Phi(\lambda x) \log(2\ \Phi(\lambda x))\ dx. \tag{4.24}$$

Equation (**??**) can be seen as the expected value wrt a standard skew-normal density of the quantity, $\log(2\Phi(\lambda x))$. This integral has not a closed form, but approximations can be performed. Equation (**??**) is a function of λ, even considering values of μ and σ different from 0 and 1 respectively.
As already said, equation (**??**) cannot be considered as a measure of a distance metric, so we make use of the unidirectional measure $d(SN(0,1,\lambda)\|N(0,1)) = \sqrt{2\text{KLD}(SN(0,1,\lambda)\|N(0,1))}$ to obtain the distance between the two distributions. We remark that in **?** the square root is justified by considering the square power of the KLD, while the 2 is add for convenience.
Then, we uniformly penalise any portion of the parameter space having the same length. Roughly speaking, any additional quantity of distance is penalised equally and no matter where we are in the parameter space.
This implies a constant decay-rate r that corresponds to an exponential prior on the distance scale, in order to ensure the following condition to be satisfied

$$\frac{\pi_d(d+\gamma)}{\pi_d(d)} = r^{\gamma}, \qquad d, \gamma \geq 0 \tag{4.25}$$

where $\pi_d(d) = \theta \exp(-\theta d(\lambda))$ and $r = \exp(-\theta)$. Now, it is evident how an addtional amount of distance γ is penalised equally, without taking into account the position in the parameter space of d.
? suggest to select the parameter θ by making an assumption on a tail event. In our case, the condition is of the form

$$\text{Prob}(d(\lambda) > W) = \alpha \tag{4.26}$$

where W is an upper bound and α is the probability mass we put on this tail event.
To select θ we use the distance $d(\lambda)$, saying that such a distance is greater than a given threshold with a small probabilty. Generally, any transformation of λ can be used. The convenience of using the distance $d(\lambda)$ lies in the fact that we know its distribution function, so that we can easily calculate the parameter θ. Based on the choice of θ one decides how much informative the prior should be. In a sense, if we shrink most of the mass towards the base model or, on the other hand, towards the alternative one, we are being very informative, so a non informative choice lays between these extreme choices. In the former case, we are assuming a large value of θ, whilst, in the latter, a small one.
The procedure to find the parameter θ is straightforward. The survival function of the exponential distribution is equated to the probability mass that we put on the tail event so that

$$\begin{aligned} e^{-\theta W} &= 2\alpha \\ -\theta W &= \log(2\alpha) \\ \theta &= \frac{-\log(2\alpha)}{W}. \end{aligned} \tag{4.27}$$

It could appear counterintuitive the reason why we put an extra 2 next to α, but we must keep in mind that the PC prior for λ is symmetric around 0, then the distance from the base model is the same either for negative and positive values of λ and, as a consequence, to each part of the parameter space (positive and negative) is assigned just half an exponential distribution. Anyhow, if we omit the 2 we should consider it encapsulated in α.
Finally, we can define our PC prior by means of the well-known change of variable

$$\pi(\lambda) = \pi(d(\lambda))\left|\frac{\partial d(\lambda)}{\partial \lambda}\right|, \tag{4.28}$$

where, in our case, the derivative must be found numerically.
The skew-normal density can also be reparameterised in terms of the shape parameter δ, instead of λ, so that the asymmetry is even more evident. The framework above does not change, and we just need to apply the reparameterisation to any of the functions we start from (i.e. the KLD or the distance), during the principled procedure. For instance, we can easily compute the KLD as a function of δ by replacing λ with $\lambda(\delta)$ in (**??**), where

$$\lambda = \lambda(\delta) = \frac{\delta}{\sqrt{1-\delta^2}}, \qquad \lambda \in (-\infty, +\infty), \tag{4.29}$$

$$\delta = \delta(\lambda) = \frac{\lambda}{\sqrt{1+\lambda^2}}, \qquad \delta \in (-1, 1). \tag{4.30}$$

Then, if we reiterate the procedure above we obtain the PC prior for δ. Similarly, one could pass from the PC prior for λ to the one for δ by simply applying the change of variable formula.
The distance expressed as a function of δ is

$$d(\delta) = \sqrt{2\mathrm{KLD}(\delta)} = \sqrt{2\int_{\mathcal{X}} 2\ \phi(x)\ \Phi(\lambda(\delta)x)\log(2\ \Phi(\lambda(\delta)x))\ dx}, \tag{4.31}$$

where $\mathrm{KLD}(\delta)$ comes out from the substitution of (**??**) into equation (**??**).
The distance function in (**??**) is symmetric around 0 (Figure **??**), as well as the KLD; this reflects the fact that the shape parameter affect symmetrically the base density, either in the case of positive and negative δ, by simply slanting the distribution in opposite directions. The minimum is at 0, where $d(0) = 0$, while the maximum is attained at the boundary values.
Then, we assign an exponential prior to such a distance scale (see Figure **??**), but we have to pay attention to the symmetry of the distance measure; indeed the prior is calculated separately for the two distances with positive and negative values of δ, and in practice half an exponential is assigned to each branch of the distance function.
This implies that

$$\pi(d_i(\delta)) = \frac{1}{2}\theta\exp(-\theta d_i(\delta)), \qquad i = 1, 2, \tag{4.32}$$

where $d_1(\delta)$ and $d_2(\delta)$ are the distances when $-1 < \delta < 0$ and $< 0 \leq \delta < 1$. Given the symmetry of the distance scale, i.e. $d(\delta) = d(-\delta)$, it is a natural choice to put half an exponential density on each of the two pieces of the parameter space where the distance is monotone. Notice that if we do not divide

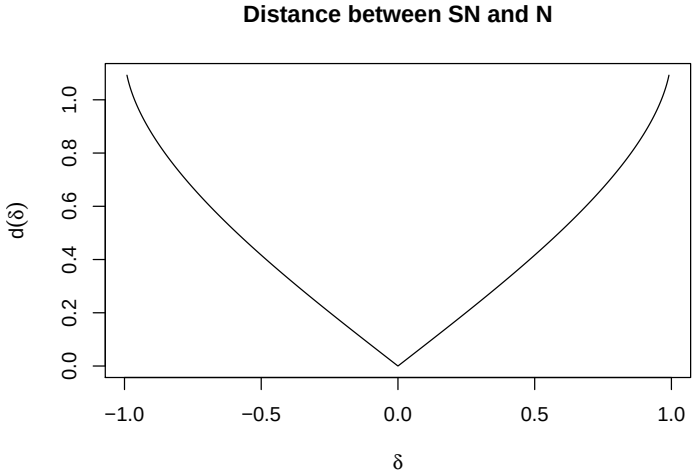

Figure 4.3: The distance measure $d(\delta)$ of a skew-normal distribution from a Gaussian base model

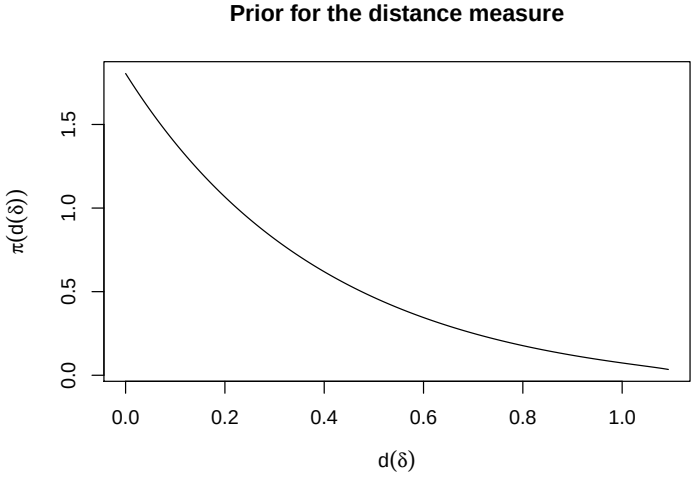

Figure 4.4: The exponential prior on the distance measure for positive δ. The graph is the same in the case of negative δ. The rate parameter $\theta = 3.62$ is inferred by means of an user-defined criterion based on a tail event.

by 2, the resulting PC prior will not integrate to one.
To better understand the concepts above, we can look at the distance in Figure ?? and notice that it takes the same values for negative δ. This is the reason why we divide by 2 each of the exponential densities in (??).
The crucial point is to select the rate parameter θ, and to do that we use the usual user-defined criterion, which in turn can be used to modify the informativeness of the prior.
In the principled construction outlined by ?, the rate parameter θ is selected by means of a statement on a tail event, that, in our case, is of the form, $\text{Prob}(d(\delta) > 1.08) = 0.01$, where $W = d(0.99) = 1.08$ and $\alpha = 0.01$. This is equivalent to say that $\text{Prob}(\delta > 0.99) = 0.01$, since there exists a direct correspondence between δ and $d(\delta)$. Notice that the latter probability statement requires to compute the survival fuction of the PC prior.
The sentences above implies $\theta = -\log(2\alpha)/W = 3.62$,where the 2 at the numerator is due to the fact that the probability statement must be considered twice, either for the positive and negative values of δ, according to the symmetry of the distance scale and to our assumption to keep the median of the PC prior at the base model.
When assigning a weight on a tail event we are intoducing our belief in the prior. Indeed, a small probability means that in our perception the boundary values are unlikely, on the contrary, we express belief for the alternative model when this probability is high and then we believe less and less in the base model.
In our opinion, in order to be noninformative or at least weakly informative, given an extreme value of W, the probability level α should not be neither too high, giving a lot of evidence for the flexible model, nor too small, corroborating the hypothesis of the base model.
In Figure ?? it is shown the PC prior for δ associated to different levels of shrinkage, controlled by the parameter θ. In our example, α is set equal to $0.005, 0.05$ and 0.1, where the two last cases reflect an increased prior belief in skewness versus symmetry. Using these tail probabilities, the corresponding values for θ are approximately equal to $4.26, 2.13$ and 1.49.
The tails of these PC priors look like truncated but, in practice, this is a computational issue deriving from the fact that these priors are calculated numerically, therefore stability problems at the boundary values may occur. Anyhow, like the Jeffreys' prior, the PC prior for δ is suitable for catching extreme values of δ, given its well-known robustness property.
In order to complete the principled construction of the prior for δ, let us make the ordinary change of variable transformation

$$\pi_i(\delta) = \pi(d_i(\delta))\left|\frac{\partial d_i(\delta)}{\partial \delta}\right|, \qquad i = 1, 2, \tag{4.33}$$

where we make use of the Leibniz's Rule to numerically compute the derivative of the Kullback-Leibler divergence, because if we would numerically compute the derivative wrt δ of the integral in (??) we would get a 0, as the KLD is a function of only δ once x is integrated out. In other words, we calculate

$$\frac{\partial d(\delta)}{\partial \delta} = \frac{1}{d(\delta)}\text{KLD}'(\delta) \tag{4.34}$$

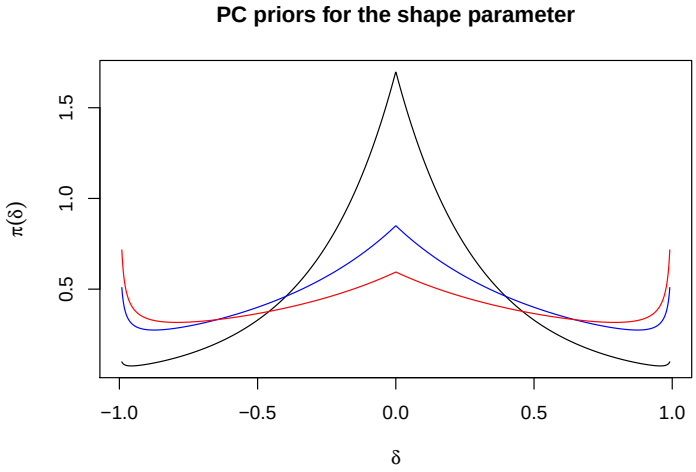

Figure 4.5: The PC prior for δ using three different scalings, in which θ is set equal to 4.26 (black), 2.13 (blue) and 1.49 (red).

where

$$\mathrm{KLD}'(\delta) = \frac{d}{d\delta}\int_a^b 2\ \phi(x)\ \Phi(\lambda(\delta)x)\log(2\ \Phi(\lambda(\delta)x))\ dx \tag{4.35}$$

$$= \int_a^b \frac{\partial}{\partial\delta} 2\ \phi(x)\ \Phi(\lambda(\delta)x)\log(2\ \Phi(\lambda(\delta)x))\ dx. \tag{4.36}$$

Notice that if a and b are constants, rather than functions of δ, the above formula holds. Thus, if we want to differentiate an integral over a finite space, it suffices to integrate the derivative of the integrand, provided that the integrand is a differentiable function over that finite space. Note also that the interchange of the derivative and the integral equates a partial derivative with an ordinary derivative. Formally, this is the case, since equation (??) is a function of δ only, while the integrand of equation (??) is a function of both δ and x. The integral in (??) must be solved numerically as the integrand in (??) is

$$\begin{aligned}
&\frac{\partial 2\phi(x)\Phi(\lambda(\delta)x)\log(2\Phi(\lambda(\delta)x))}{\partial\delta}\\
=&2\phi(x)\phi(\lambda(\delta)x)x\cdot\log(2\Phi(\lambda(\delta)x)) + 2\phi(x)\Phi(\lambda(\delta)x)\cdot\frac{\phi(\lambda(\delta)x)x}{\Phi(\lambda(\delta)x)}\\
=&2\phi(x)\phi(\lambda(\delta)x)x\big[1+\log(2\Phi(\lambda(\delta)x)\big].
\end{aligned} \tag{4.37}$$

Finally, we must be careful in making the change of variable of the final step, because we have to handle each monotone curve separately.
Suppose there exists a partition of the parameter space Δ (that is the parameter space of δ) such that $d(\delta)$ is monotone on each chunk. Note that, in this case, each $d_i(\delta)$ is a one-to-one transformation from any piece of the parameter space Δ onto the space Θ (i.e. the space of $d(\delta)$). Notice that the set $\Theta = (0, \infty)$. In particular, the function $d(\delta)$ is monotone on $(-1, 0)$ and on $(0, 1)$.

Then, the pdf for δ is

$$\pi(\delta) = \begin{cases} \sum_{i=1}^{2} \pi(d_i(\delta)) \left| \frac{\partial d_i(\delta)}{\partial \delta} \right| & \text{if } d(\delta) \in \Theta \\ 0 & \text{otherwise.} \end{cases} \tag{4.38}$$

The procedure that we have depicted above allows us to get the prior numerically as the KLD has not a closed form. To this aim we could try to perform an approximation of the KLD, based on the moments of the skew-normal distribution, in order to obtain a closed-form density. Such an approximation is quite good, but it works not very well on the tails, especially when the parameter θ is small; this is because the probability mass spreads out away from the base model.

4.2.3 Approximation of the KLD

Here, we propose a way to approximate the KLD in equation (**??**). This can be done by simply approximating the logarithm of the normal CDF by means of a polynomial regression of degree five. The amazing fact is that the intercept α get closer and closer to $-\log 2$ as we increase the degree of the polynomial regression, and this is crucial to have the $\text{KLD}(\lambda = 0) = 0$. We could also increase the degree of the polynomial regression, but, in practice, this is not convenient as the quintic regression seems to work very well.
Then, the KLD can be written as

$$\begin{aligned} \mathbb{E}_Y[\log(2\Phi(\lambda Y))] = \log 2 + \\ \mathbb{E}_Y(\alpha + \beta\lambda Y + \xi\lambda^2 Y^2 + \gamma\lambda^3 Y^3 + \epsilon\lambda^4 Y^4 + \eta\lambda^5 Y^5), \end{aligned} \tag{4.39}$$

where α, β, ξ, γ, ϵ and η are the coefficients of the polynomial regression, and $Y \sim SN(\lambda)$.
So, the KLD can be approximated by the first five moments of the skew-normal distribution as follows

$$\begin{aligned} \log 2 + \alpha + \beta\lambda\sqrt{\frac{2}{\pi}}\delta + \xi\lambda^2 + \gamma\lambda^3\sqrt{\frac{2}{\pi}}(3\delta - \delta^3) + \\ 3\epsilon\lambda^4 + \eta\lambda^5\sqrt{\frac{2}{\pi}}(15\delta - 10\delta^3 + 3\delta^5), \end{aligned} \tag{4.40}$$

where

$$\delta = \delta(\lambda) = \frac{\lambda}{\sqrt{1+\lambda^2}}. \tag{4.41}$$

In this way, we would be able to derive an analytical expression of the PC prior for λ.

4.3 Bayesian Inference for the Skew-Normal Model

In this section, we are going to check the frequentist properties of the PC prior, the one numerically computed, and we want to compare it to the Jeffreys' prior in **?**, in order to see if there could be a certain value of the parameter θ that

can be interpreted as objective.
Notice that if such a value would exist, this would not mean that the probability statement on the tail event as in (??) is univocal, since there exist many possible combinantions of α and W giving the same value of θ. So, this is a different point of view, no longer based on the probability statement eliciting the parameter θ, but directly on θ and, in a sense, we get rid of the user information entry.
The PC prior for the shape parameter, λ or δ, is certainly proper as it is a transformation of the exponential prior assigned to the distance scale, and also the Jeffreys' prior is proper even if it has an unbounded support, unlike one can expect. So, the comparison between the PC and the Jeffreys' priors is suitable also in the Bayesian Hypothesis testing problem.
The PC prior for δ has a no closed-form density, but we can still write it down as follows

$$\pi^{PC}(\delta) = \frac{\theta}{2} e^{-\theta\sqrt{2\mathrm{KLD}(\delta)}} \frac{|\mathrm{KLD}'(\delta)|}{\sqrt{2\mathrm{KLD}(\delta)}}, \tag{4.42}$$

where $\mathrm{KLD}(\delta) = \int_{-1}^{1} 2\phi(x)\Phi(\lambda(\delta)x)\log(2\Phi(\lambda(\delta)x))$ and $\mathrm{KLD}'(\delta)$ is given in equation (??). Note that the denominator of (??) does not need to be considered in absolute value as the KLD is always positive.
We perform a simulation study for different values of the shape parameter, $\lambda^*(0, 0.5, 1, 2, 3, 5)$, for various sample sizes, $n(10, 30, 100)$, and for several choices of the shrinkage parameter, $\theta(0.1, .5, 1, 1.5, 2, 5)$. For each combination of the aforementioned vectors we simulate 200 samples, where each of them has $1.5 \cdot 10^4$ observations, after a burn-in of $5 \cdot 10^3$ observations. We use a Random Walk Metropolis-Hastings algorithm to simulate from the posterior distribution, where the proposal distribution was chosen to be a truncated normal.
The acceptance probability at each iteration, i, is given by

$$a_{k,j,i} = \frac{\prod_{i=1}^{n_k} 2\phi(x_i)\Phi(\lambda^N x_i) \cdot \pi^{PC}(\lambda^N|\theta_j)}{\prod_{i=1}^{n_k} 2\phi(x_i)\Phi(\lambda^C x_i) \cdot \pi^{PC}(\lambda^C|\theta_j)}, \tag{4.43}$$

where λ^N stands for the proposed value of λ, while λ^C stands for the current one.
The case of small sample size, n, is strongly affected by the value of the parameter θ, especially when it leads to opposite conclusion. To clarify the aforementioned statement, consider a true model with a significant departure from symmetry and a very large value of θ, shrinking the probability mass towards the base model.
Tipically, the functions implemented in the `R` packages are conceived to simulate from a skew-normal expressed in terms of λ, but we can easily transform a λ value into a δ one by means of equation (??).
For any combination of λ^*, n and θ we calculate

- the MSE of the posterior mean;
- the MSE of the posterior median;
- the MSE of the posterior mode;
- the coverage probabilities.

The value of the MSE of the posterior median and mode is useful to catch skewness in the posterior distribution when compared to the MSE of the posterior

True value	**MSE Posterior Mean**						
	PC						**Jeffreys**
	$\theta = 0.1$	$\theta = 0.5$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 5$	
n=10							
$\lambda = 0$	0.089867	0.082730	0.069666	0.072587	0.059750	0.028800	0.118437
$\lambda = 0.5$	0.088862	0.087516	0.088894	0.077097	0.092027	0.092135	0.085726
$\lambda = 1$	0.066566	0.088980	0.078523	0.094467	0.112976	0.171337	0.075261
$\lambda = 2$	0.039076	0.033467	0.047420	0.061341	0.078876	0.204050	0.017891
$\lambda = 3$	0.023296	0.024840	0.032520	0.039368	0.055860	0.176597	0.009958
$\lambda = 5$	0.014325	0.018413	0.023931	0.029809	0.039666	0.164643	0.006689
n=30							
$\lambda = 0$	0.041333	0.043351	0.032179	0.037871	0.029039	0.018809	0.045931
$\lambda = 0.5$	0.034232	0.030510	0.039603	0.036132	0.034367	0.054545	0.033327
$\lambda = 1$	0.019050	0.021026	0.023711	0.021800	0.027198	0.051911	0.017541
$\lambda = 2$	0.007867	0.006249	0.008385	0.007593	0.008000	0.018115	0.003866
$\lambda = 3$	0.002404	0.003762	0.003927	0.003847	0.004464	0.010296	0.002202
$\lambda = 5$	0.001179	0.001254	0.001838	0.002076	0.001888	0.004029	0.000972
n=100							
$\lambda = 0$	0.016266	0.013199	0.012840	0.013109	0.011853	0.007854	0.014722
$\lambda = 0.5$	0.008994	0.011007	0.011841	0.010943	0.011454	0.014646	0.011108
$\lambda = 1$	0.005352	0.005173	0.006052	0.005972	0.006058	0.007594	0.005109
$\lambda = 2$	0.001386	0.001274	0.001309	0.001432	0.001277	0.001660	0.001194
$\lambda = 3$	0.000663	0.000576	0.000623	0.000588	0.000607	0.000834	0.000522
$\lambda = 5$	0.000120	0.000143	0.000164	0.000129	0.000167	0.000193	0.000111

Table 4.1: Mean Squared Error computed over the posterior mean and expressed in δ

mean. Actually, the difference is quite small, given also that the MSE are computed with respect to δ, nonetheless the simulated posterior distributions are often skewed. For instance, for the mode, $E[(Mode(Y_s) - \delta(\lambda))^2]$, where Y_s is the sampling distribution for $s = 1, \ldots, 200$ and converted into δ, $\delta(\lambda) = \lambda/\sqrt{1+\lambda^2}$ is the true asymmetry, and the expected value is taken over the sampling distribution. Exactly the same holds for the median and the mean of the simulated sampling distribution.
Posterior median and mode are more reasonable choices, especially for samples where the MLE is infinite, because this entails the non finiteness of the posterior mean. Therefore, based on the posterior median and mode, credible intervals can be constructed, and they constitutes a valid alternative to the frequentist approach, where the confidence intervals are based on the likelihood function, especially when the MLE is infinite (**?**).
The coverage probabilities are calculated over the 95% two-sided credible intervals. All the summaries are compared with the Jeffreys' prior (actually its approximation (**?**)).

What emerges from the simulation study is, as expected, that large values of θ are less significant, in the sense that they produce more biased estimates, especially in samples where the true $\lambda \neq 0$, and this obviously happens for all the summaries of the posterior distribution, namely the mean, the median and the mode. The only case in which a large value of θ works well is when the true

True value	**MSE Posterior Median**						
	PC						**Jeffreys**
	$\theta = 0.1$	$\theta = 0.5$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 5$	
n=10							
$\lambda = 0$	0.111301	0.101081	0.082363	0.083720	0.067048	0.025711	0.148128
$\lambda = 0.5$	0.099046	0.097222	0.097966	0.084717	0.097952	0.101058	0.098773
$\lambda = 1$	0.059164	0.080351	0.070268	0.087710	0.110060	0.182589	0.073065
$\lambda = 2$	0.024342	0.019423	0.030741	0.042326	0.057033	0.197228	0.010022
$\lambda = 3$	0.010805	0.011202	0.015784	0.019809	0.032503	0.154329	0.003790
$\lambda = 5$	0.004303	0.006103	0.008874	0.010493	0.016610	0.131761	0.001609
n=30							
$\lambda = 0$	0.045657	0.047195	0.034525	0.040574	0.030281	0.017896	0.051107
$\lambda = 0.5$	0.035107	0.031635	0.040318	0.037144	0.035672	0.057147	0.034567
$\lambda = 1$	0.017312	0.018835	0.021671	0.018932	0.024056	0.047485	0.016307
$\lambda = 2$	0.005960	0.004548	0.006016	0.005392	0.005466	0.012975	0.003052
$\lambda = 3$	0.001479	0.002294	0.002394	0.002347	0.002715	0.006185	0.001476
$\lambda = 5$	0.000513	0.000554	0.000890	0.001018	0.000812	0.001787	0.000459
n=100							
$\lambda = 0$	0.016873	0.013632	0.013131	0.013298	0.011943	0.007434	0.015329
$\lambda = 0.5$	0.008977	0.011061	0.011779	0.010845	0.011247	0.014328	0.011149
$\lambda = 1$	0.005142	0.004777	0.005720	0.005519	0.005671	0.006969	0.005002
$\lambda = 2$	0.001253	0.001117	0.001148	0.001245	0.001069	0.001404	0.001120
$\lambda = 3$	0.000574	0.000470	0.000511	0.000473	0.000500	0.000665	0.000459
$\lambda = 5$	8.26e-05	9.62e-05	0.000112	8.63e-05	0.000114	0.000126	7.81e-05

Table 4.2: Mean Squared Error computed over the posterior median and expressed in δ

True value	**MSE Posterior Mode**						
	PC						**Jeffreys**
	$\theta = 0.1$	$\theta = 0.5$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 5$	
n=10							
$\lambda = 0$	0.123219	0.103710	0.078503	0.069142	0.050591	0.014757	0.134979
$\lambda = 0.5$	0.089478	0.100381	0.111881	0.103866	0.117811	0.142752	0.098465
$\lambda = 1$	0.096399	0.101293	0.099884	0.153311	0.162491	0.284020	0.092151
$\lambda = 2$	0.062502	0.063836	0.077122	0.113732	0.131884	0.430525	0.025890
$\lambda = 3$	0.048700	0.047134	0.059764	0.066059	0.115667	0.398507	0.021196
$\lambda = 5$	0.024354	0.035619	0.052295	0.060260	0.087452	0.404376	0.013272
n=30							
$\lambda = 0$	0.047357	0.052124	0.036384	0.042368	0.028650	0.017498	0.056356
$\lambda = 0.5$	0.038507	0.038416	0.044747	0.049277	0.039885	0.075968	0.040942
$\lambda = 1$	0.023308	0.022653	0.027589	0.025490	0.034549	0.071869	0.020645
$\lambda = 2$	0.009375	0.009091	0.011502	0.009257	0.011511	0.025423	0.004415
$\lambda = 3$	0.003939	0.004097	0.006329	0.004789	0.006684	0.014068	0.003122
$\lambda = 5$	0.002106	0.002204	0.003714	0.003707	0.004135	0.007200	0.003631
n=100							
$\lambda = 0$	0.018426	0.016572	0.018525	0.014511	0.013671	0.008085	0.019010
$\lambda = 0.5$	0.011318	0.013871	0.013864	0.014241	0.012805	0.015396	0.013266
$\lambda = 1$	0.006714	0.005615	0.007054	0.007287	0.006926	0.008962	0.005633
$\lambda = 2$	0.001590	0.001515	0.001538	0.001740	0.001656	0.002240	0.001329
$\lambda = 3$	0.000783	0.000655	0.000713	0.000602	0.000762	0.000967	0.000550
$\lambda = 5$	0.000186	0.000189	0.000211	0.000207	0.000205	0.000313	0.000145

Table 4.3: Mean Squared Error computed over the posterior mode and expressed in δ

True value	**Coverage**						
	PC						**Jeffreys**
	$\theta = 0.1$	$\theta = 0.5$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 5$	
n=10							
$\lambda = 0$	0.945	0.985	0.985	0.960	0.985	1.000	0.940
$\lambda = 0.5$	0.945	0.955	0.945	0.965	0.945	0.905	0.960
$\lambda = 1$	0.955	0.930	0.945	0.935	0.900	0.805	0.960
$\lambda = 2$	0.940	0.970	0.945	0.925	0.895	0.715	0.975
$\lambda = 3$	0.935	0.925	0.930	0.925	0.885	0.750	0.975
$\lambda = 5$	0.945	0.940	0.895	0.910	0.870	0.590	0.980
n=30							
$\lambda = 0$	0.935	0.960	0.965	0.950	0.975	0.985	0.950
$\lambda = 0.5$	0.950	0.960	0.935	0.960	0.950	0.880	0.945
$\lambda = 1$	0.935	0.955	0.930	0.965	0.925	0.880	0.950
$\lambda = 2$	0.955	0.965	0.950	0.960	0.960	0.935	0.945
$\lambda = 3$	0.980	0.940	0.935	0.960	0.950	0.910	0.950
$\lambda = 5$	0.945	0.965	0.900	0.930	0.965	0.920	0.930
n=100							
$\lambda = 0$	0.920	0.950	0.965	0.955	0.970	0.980	0.945
$\lambda = 0.5$	0.975	0.960	0.940	0.955	0.945	0.945	0.935
$\lambda = 1$	0.965	0.950	0.940	0.945	0.935	0.895	0.925
$\lambda = 2$	0.935	0.970	0.950	0.940	0.985	0.935	0.940
$\lambda = 3$	0.905	0.945	0.920	0.940	0.935	0.915	0.940
$\lambda = 5$	0.975	0.955	0.970	0.975	0.935	0.955	0.960

Table 4.4: Coverage probabilities over the 95% credible interval of the simulated posterior distribution expressed in δ

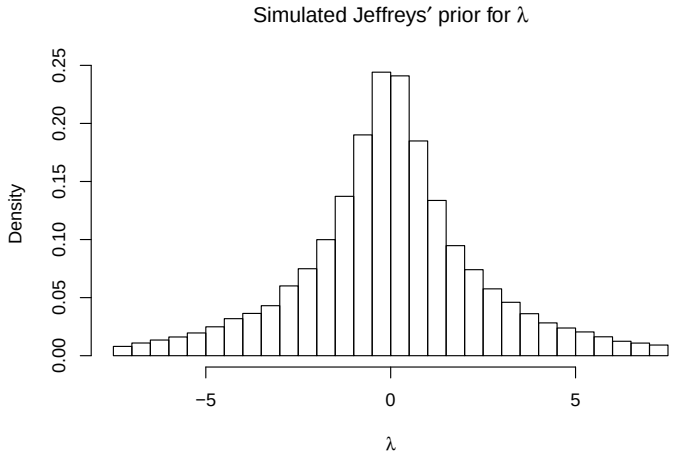

Figure 4.6: Simulation of the Jeffreys' prior for λ.

parameter $\lambda = 0$, and therefore the skew-normal distribution boils down to the normal one. This kind of behaviour is obvious, since large values of θ produce more and more shrinkage towards the base model.
So, when the true model is Gaussian, the employment of a large value of θ produces more accurate estimates for λ, even though the gap wrt small values of θ vanishes for large sample sizes. On the contrary, when the asymmetry is strong, the value of θ that seems to capture more adequately such an asymmetry is a small one. In the middle there are all the other combinations.
In practice, particular values of θ can be better to catch specific values of λ. So, the choice of θ depends on the problem at hand. The tendency is that limiting choices of θ work better for extreme values of the asymmetry (0 included).
Nevertheless, in the current work we are interested to the identification of a particular θ that can be interpreted as objective, preventing us to choose a particular value of θ for each given amount of asymmetry, λ.
Notice that even being a Student-t distribution, the Jeffreys' prior looks U-shaped when it is plotted in terms of δ. It is straightforward to see that it produces more precise estimates when a true strong asymmetry is taken into account (tables above). Indeed, column 8 of each of the three tables (**??**, **??**, **??**) has a decreasing tendency as the absolute value of λ increases. This is an evident fact, since the Jeffreys' prior for δ puts most of the mass on the tails and, in a sense, it is assimilable to a PC prior distribution where the parameter θ is close to 0 (see figures **??** and **??**).

In Figure **??** we show a simulated PC prior expressed in terms of λ. The PC prior, expressed in terms of δ, for values of θ small enough is very close to the Jeffreys' prior of Figure **??**. Indeed, at a glance, it seems to come out from the tables above that a PC prior with the parameter θ approximately between 0.1 and 0.5 approaches the estimates produced by the Jeffreys' prior.
Notice that, unlike the Jeffreys' prior, the PC prior having $\theta = 0.1$ (figure **??**) has a minimal spike at 0, therefore the estimates can be more precise in the cases where the true value of λ is zero and especially for small sample sizes.

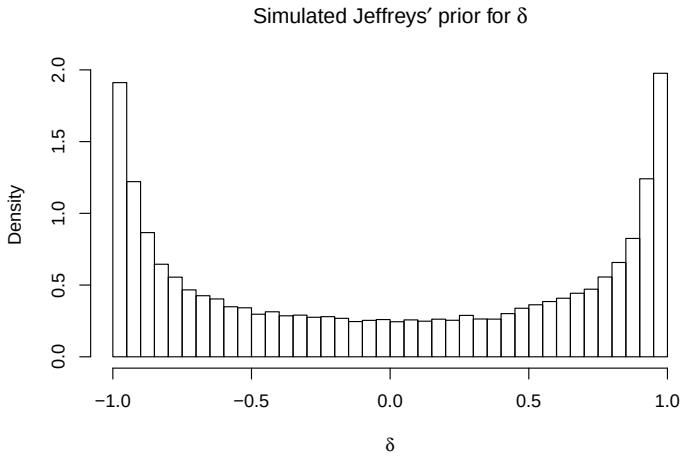

Figure 4.7: Simulation of the Jeffreys' prior for δ.

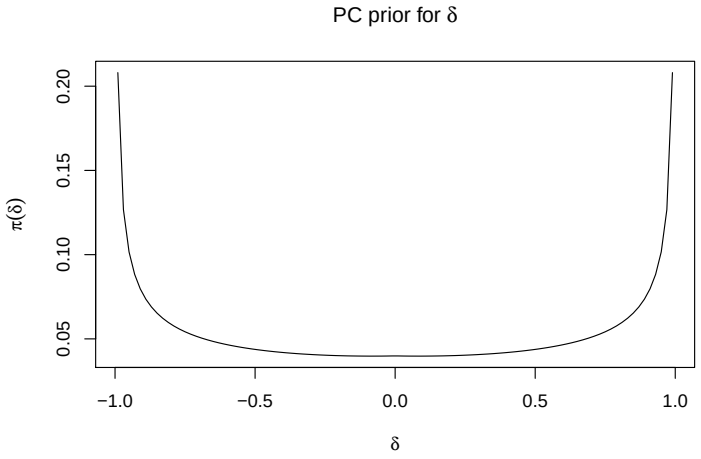

Figure 4.8: PC prior for δ with $\theta = 0.1$.

Now, let us consider Table **??**, the one reporting the value of the MSE over the posterior median. We can observe how close to column 8 is the second column, i.e. the column corresponding to the PC prior with $\theta = 0.1$. So, if we choose a noninformative value for θ we would say one between 0.1 and 0.5.
Table **??** seems to show better results than tables **??** and **??**, probably beacause the posterior median is one of the most reliable estimators. Most of the difference is visible for small values of n. The posterior median and mode are good estimators as they are always finite, on the other hand, the finiteness of the posterior mean depends on the finiteness of the Maximum Likelihood Estimator.
However, an interesting feature is that the general behaviour (in terms of MSE) of the PC prior for almost any value of θ is decreasing when λ grows, and this is a characteristics shared also by the Jeffreys' prior. The aforementioned trait explains the good properties of both the priors in terms of robustness, although issues arise for the PC prior whose parameter θ takes too large values. Notice that when $\theta \to \infty$ the prior distribution becomes, in practice, a Dirac distribution at the base model.
Results of Table **??** confirm how misleading an high value of θ can be. In fact, especially for small sample sizes, the coverage probabilities behave very bad when the parameter θ of the PC prior is sufficiently large. This is a further proof of the inadequacy of a too strong shrinkage. Also the Bayes factor, as we will see in the following section, suffers a large value of θ, as it becomes inconsistent for the PC prior approaching the Dirac distribution.

4.4 Bayesian Hypothesis Testing

In this section we will use a PC prior for a Bayesian hypotheses test and we will compare it to the Jeffreys' prior in **?**, successively approximated by **?** as a Student-t distribution, $t(\lambda|\mu = 0, \sigma = \pi/2, \nu = 1/2)$. The comparison with the Jeffreys' prior makes sense given its properness, as a special case of proper Jeffreys' prior over an unbounded domain.
The proposition stated in section **??** is very important, especially because it allows us to write the Bayes factor in a simplified manner, i.e. without considering the joint prior distribution over the location and scale parameters, that is separable with respect to the PC prior build up on the skewness component introduced in the model. According to the proposition above, the Bayes factor for testing

$$H_0 : \lambda = 0 \qquad \text{vs} \qquad H_1 : \lambda \neq 0$$

can be written as

$$\text{BF}_{01}(x) = \frac{\prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)|_{\lambda=0}}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)\pi^{PC}(\lambda|\theta)d\lambda}, \tag{4.44}$$

where the marginal likelihood $\int_{-\infty}^{\infty} \prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)\pi^{PC}(\lambda|\theta)d\lambda$ is computed by means of an importance sampling approximation.
Then, the Bayes factor is computed as follows

$$\text{BF}_{01}(x) = \frac{\prod_{i=1}^{n} \phi(x_i)}{E_g[\prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)\pi^{PC}(\lambda|\theta)/g(\lambda)]}, \tag{4.45}$$

where $g(\lambda)$ is an auxiliary distribution. The advantage to use an importance sampling solution instead of a standard Monte Carlo approximation is that we

can reduce the variance of the importance sampling estimates depending on the choice of the importance distributions. In general, if we pick auxiliary distributions close enough to the posterior density we have good estimates in terms of the variance of the importance estimates.
We use, as the auxiliary distribution, a truncated uniform density that expresses no idea about the value of λ. By using a standard Monte Carlo we would draw directly from the PC prior for λ and consequently we could obtain samples that produce negligible values of the likelihood function, for instance like if we use a very small parameter θ and the asymmetry is not so strong.
However, as we may notice either from Table **??** and **??**, the choice of $\theta = 0.1$ is suboptimal, in terms of the convergence of the Bayes factor towards the true model. In fact, a part from the case of no asymmetry, where the Bayes factor goes to infinity and the frequency of times being less than 0.5 and greater than 2 should be small in the former case and large in the latter one, the employment of a PC prior with $\theta = 0.1$ is suboptimal. The choice to use the aforementioned thresholds, i.e. 0.5 and 2, has been borrewed from **?**; in addition, especially for the lower threshold, such a choice has a natural answer in scheme (**??**), as, for a PC prior with parameter θ small enough, the BF_{01} has the tendency to take either very large values or values close to 0.5.
One can notice, for $\theta = 0.1$ and the true model having a skewness different from zero, that the frequency of times that the $\text{BF}_{01} < 0.5$ is not acceptable, even compared to the Jeffreys' prior and especially when the true λ is small, although the situation improves for large values of λ and n.
The same can be seen by looking at the frequency of times that $\text{BF}_{01} > 2$. The reason of such a behaviour lies in the fact that small values of θ tend to produce too large Bayes factors, even in the case of $\lambda \neq 0$. Let us understand why. If we draw from a PC prior with a parameter θ too small, we are supporting the idea of a skewed model. So, it is more likely to sample large value of λ, or equivalently, values of δ close to 1. Then, by plugging these values, drawn from the PC prior distribution with a small θ, into the likelihood function will produce a marginal likelihood close to 0, as long as the distance between the likelihood function and the prior distribution is large. In practice, it is like if we were taking into account negligible values of the likelihood.
Now, to better understand the concepts above, let us see the analytical behaviour at the boundaries of the marginal likelihood as in equation (**??**).
Suppose to draw values of λ from a PC prior with $\theta \to 0$, then for a generic x_i

$$\text{if} \begin{cases} \lambda \to \infty \begin{cases} x_i & \text{is positive} \implies \text{BF}_{01} \approx 0.5 \\ x_i & \text{is negative} \implies \text{BF}_{01} \approx \infty \end{cases} \\ \lambda \to -\infty \begin{cases} x_i & \text{is positive} \implies \text{BF}_{01} \approx \infty \\ x_i & \text{is negative} \implies \text{BF}_{01} \approx 0.5 \end{cases} \end{cases} . \tag{4.46}$$

What emerges from the scheme above is that if the likelihood function is not concentrated at the boundaries of the parameter space and therefore a generic observation is likely to have the opposite sign of an extreme value of λ drawn from the PC prior with a small θ, then the marginal likelihood goes to zero. In fact, $\prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)$ goes to zero when computed for an extreme value of λ and for x_i having the opposite sign. As a consequence the Bayes factor goes to ∞. Otherwhise, it goes to 1/2 when the argument of $\Phi(\cdot)$ goes to ∞ and consequently $\Phi(\cdot)$ goes to one. The sign of the observations, x_i, results essential

in determining the limiting behaviour of the Bayes factor.
On the other hand, the Bayes factor becomes inconsistent for large values of θ that lead to values of λ concentrated around the base model. This latter has the effect to reduce the marginal likelihood of a standard skew-normal distribution to a likelihood function of a standard normal distribution.
As we can see from tables ?? and ??, the more we increase the value of θ, the more the Bayes factor goes faster towards the true model, but at a certain point it stops to converge, i.e. for $\theta = 5$. Let us try to understand why. PC priors with too large values of θ produce Bayes factor whose values are close to 1, giving no more than a bare mention in favour of one of the two models. It doesn't matter what is the true model, the Bayes factor will bring no evidence in favour of the true model, and for $\theta \to \infty$ it will be exactly equal to 1. When using a PC prior with a very strong shrinkage, the prior mass contract to the base model and at the limit it looks like a Dirac distribution at the base model.
So, what happens analitically is that the integral of the likelihood at the denominator boils down to the one at the numerator, as the prior on the alternative model will have all the mass concentrated at zero and no mass otherwise. In practice, we will have

$$\mathrm{BF}_{01}(x) = \frac{f(x|\lambda)|_{\lambda=0}}{\int_{-\infty}^{\infty} f(x|\lambda)\mathbb{I}_{\{\lambda=0\}}d\lambda} = 1, \tag{4.47}$$

where $\mathbb{I}_{\{\lambda=0\}}$ denotes the Dirac distribution and $f(x|\lambda)$ is the likelihood function. The property above discourages the employment of large values of θ in the context of Bayesian hypothesis testing, even though these values may work properly in some estimation problems.

We can appreciate the behaviour described above by looking at tables ?? and ??. In fact, we can notice in Table ?? that for the true model being a skew-normal, the number of times in which the Bayes factor is smaller than 0.5 has the tendency to increase gradually up to the value of $\theta = 2$. Then, for $\theta = 5$ it starts to decrease as a proof of the fact that the Bayes factor contracts to 1. The same tendency can be seen by looking at Table ??, indeed the frequency of times that the Bayes factor is larger than 2 shrinks to 0 for $\theta = 5$, even for the true model having $\lambda = 0$.
Then, a calibrated choice for θ in the context of the Bayesian hypothesis testing seems to be $\theta = 2$. As discussed above, a too large value θ produce Bayes factors equal to one, on the other hand, too small values will be overconfident in favour of the null model. Anyhow, the PC prior with $\theta = 2$ beats surely the Jeffreys' prior, that, such as the PC prior with small θ, is overconfident in favour of no skewness.
Let us see the graphical behaviour of the Bayes factor for different values of θ. In Figure ?? we show the histograms of the Bayes factor for different values of θ, where there are 30 observations and the true model has no asymmetry, so it is a standard normal.
As we can see from the histograms, the best choice in this case would be $\theta = 0.1$, but as discussed above, this is only a fictitious good choice, as such a value of θ provides very large Bayes factors in any case, even when the true model has a certain asymmetry.
Finally, we can observe how a large value of θ, i.e. $\theta = 5$, makes the Bayes factor inconsistent, in the sense that it concentrates to 1, providing no evidence

True value	**Frequency of times $BF_{01} < 0.5$**						
	PC						**Jeffreys**
	$\theta = 0.1$	$\theta = 0.5$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 5$	
n=10							
$\lambda = 0$	0.010	0.015	0.025	0.040	0.070	0.030	0.045
$\lambda = 0.5$	0.025	0.095	0.195	0.190	0.215	0.155	0.120
$\lambda = 1$	0.160	0.315	0.415	0.420	0.510	0.350	0.445
$\lambda = 2$	0.405	0.710	0.800	0.820	0.865	0.760	0.880
$\lambda = 3$	0.560	0.875	0.950	0.960	0.980	0.870	0.945
$\lambda = 5$	0.765	0.980	0.995	0.985	0.990	0.945	1.000
n=30							
$\lambda = 0$	0.000	0.015	0.020	0.020	0.045	0.040	0.005
$\lambda = 0.5$	0.125	0.330	0.410	0.465	0.495	0.530	0.380
$\lambda = 1$	0.675	0.875	0.900	0.905	0.960	0.945	0.840
$\lambda = 2$	0.995	1.000	1.000	1.000	1.000	1.000	1.000
$\lambda = 3$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\lambda = 5$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
n=100							
$\lambda = 0$	0.000	0.010	0.010	0.015	0.015	0.060	0.010
$\lambda = 0.5$	0.700	0.885	0.905	0.920	0.935	0.970	0.890
$\lambda = 1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\lambda = 2$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\lambda = 3$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\lambda = 5$	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.5: The frequency of times that $BF_{01} < 0.5$ over 200 samples.

True value	**Frequency of times $BF_{01} > 2$**						
	PC						**Jeffreys**
	$\theta = 0.1$	$\theta = 0.5$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 5$	
n=10							
$\lambda = 0$	0.960	0.875	0.705	0.545	0.180	0.000	0.775
$\lambda = 0.5$	0.900	0.725	0.445	0.225	0.045	0.000	0.590
$\lambda = 1$	0.670	0.335	0.130	0.040	0.005	0.000	0.185
$\lambda = 2$	0.240	0.030	0.015	0.005	0.000	0.000	0.010
$\lambda = 3$	0.105	0.010	0.000	0.000	0.000	0.000	0.005
$\lambda = 5$	0.010	0.000	0.000	0.000	0.000	0.000	0.000
n=30							
$\lambda = 0$	0.980	0.915	0.875	0.790	0.715	0.000	0.935
$\lambda = 0.5$	0.755	0.400	0.300	0.145	0.100	0.000	0.400
$\lambda = 1$	0.165	0.035	0.010	0.000	0.000	0.000	0.045
$\lambda = 2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\lambda = 3$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\lambda = 5$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n=100							
$\lambda = 0$	1.000	0.940	0.935	0.870	0.845	0.420	0.955
$\lambda = 0.5$	0.135	0.030	0.040	0.015	0.010	0.000	0.050
$\lambda = 1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\lambda = 2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\lambda = 3$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\lambda = 5$	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 4.6: The frequency of times that $BF_{01} > 2$ over 200 samples.

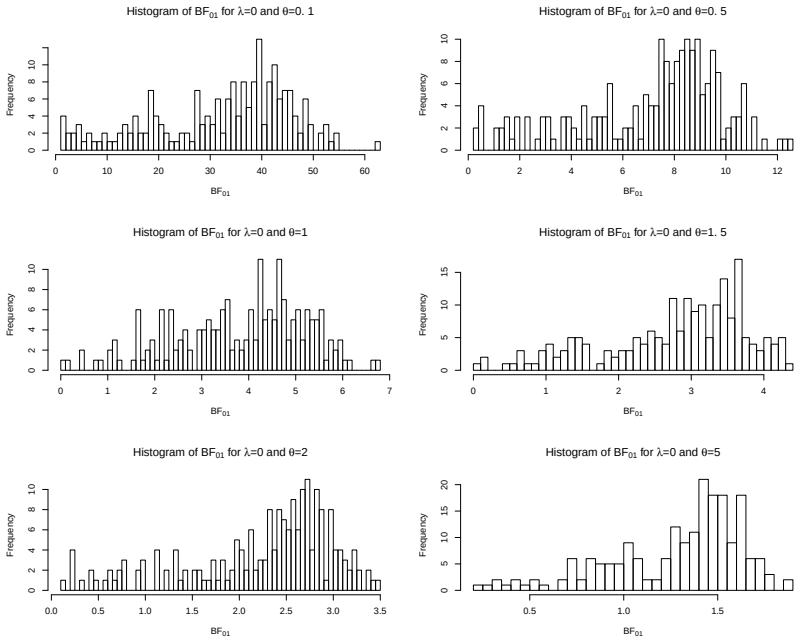

Figure 4.9: Bayes factor for $n = 30$, $\lambda = 0$ and diferent choices of θ.

in favour of the true model.
In figure ?? we can see again what happens with θ values very small or large. Now, the true asymmetry λ^* is 0.5. In the former case (small θ) the Bayes factor takes very large values, even if it should converge to zero, while in the latter case (large θ) we notice once again how the Bayes factor becomes closer and closer to 1.

4.5 The Frontier data set

As mentioned above, there are cases in which the MLE of λ is problematic. We are not referring to problems related to numerical maximization, rather to intrinsic properties of the likelihood function, that are not removable with a change of parameterization.
An example is provided by a challenging data set consisting of 50 simulated points from a SN$(0, 1, 5)$ (?). The particularity of such a data set is that the MLE of λ is infinite and therefore frequentist methods seems to suffer; however a solution has been proposed by ?. Figure ?? illustrate what is called the `frontier` data set (small circles), along with a nonparametric estimate of the density (continuous curve) and a parametric SN estimate (dashed curve). The parameter λ of the parametric SN curve is 8.14, so, a graphical approach would validate such a value for the shape paremeter. On the other hand, the MLE is infinite, corresponding to an half-normal distribution.
The MLE is not a satisfactory estimator, especially for small n and large λ. As claimed by ?, the probability to incur in an infinite MLE depends on the sign of the data rather than their actual value, and it is obvious that the same sign

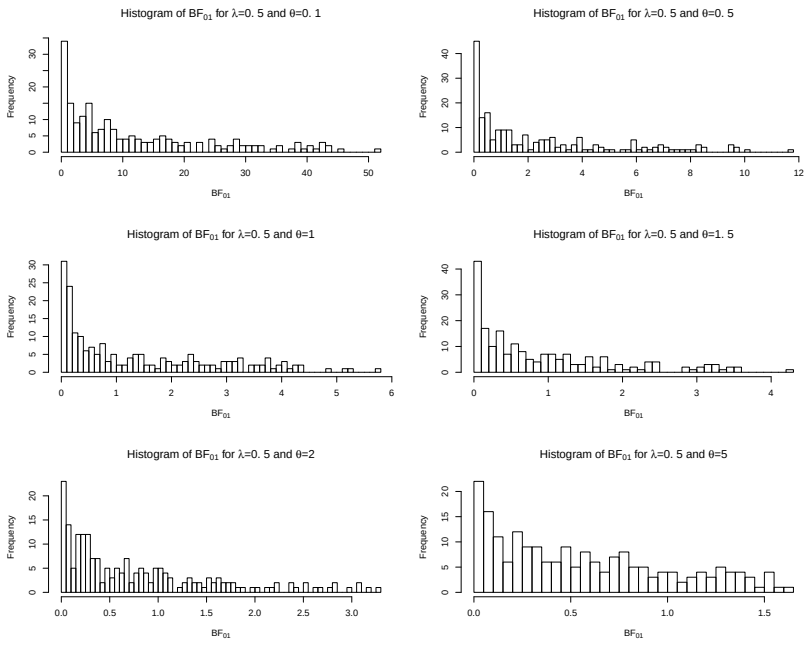

Figure 4.10: Bayes factor for $n = 30$, $\lambda = 0.5$ and diferent choices of θ.

is more likely to occur for strong asymmetry and a few data points. Alternative solutions to the MLE must be found and a natural way seems to be the employment of Bayesian strategies.
On the frequentist point of view, **?** proposed to tackle this problem by using a simple method. In the cases in which the MLE occurs at the frontier, they restart the maximization procedure and stop it when it reaches a loglikelihood value not significantly lower tha the maximum. This is the strategy used to estimate the parametric SN density in Figure **??**.
As a possible solution of the above anomaly we try to investigate the use of the PC prior for the shape parameter λ in the estimation problem. In addition, we are interested in observing if the rate parameter θ, for this particular data set, plays or not a key role, in the sense that it affects the estimation process.
We consider the posterior median as a point estimator. The Jeffreys' prior tends to be overconfident about λ (figure **??**), while the PC prior tends to provide good enough estimates in the range of θ in $(0.1, 2)$.
As expected from the estimation properties of the PC priors, it turns out that small values of θ provide very good results for the frontier data set, also given the strong asymmetry. Although the data set is large enough to make the prior influence vanish, it is still affected by very large values of θ, that make the shrinkage too heavy.
However, the PC prior do not overestimate the skewness more than the Jeffreys' prior, and this kind of behaviour is easy to interpret, as the PC prior preserves the base model, while the Jeffreys' prior spreads out at the boundaries.
Figures **??**, **??**, **??**, **??**, **??**, **??** and **??** show the simulated posterior distributions of λ for PC priors with varying θ as well as for the Jeffreys' prior. The posterior

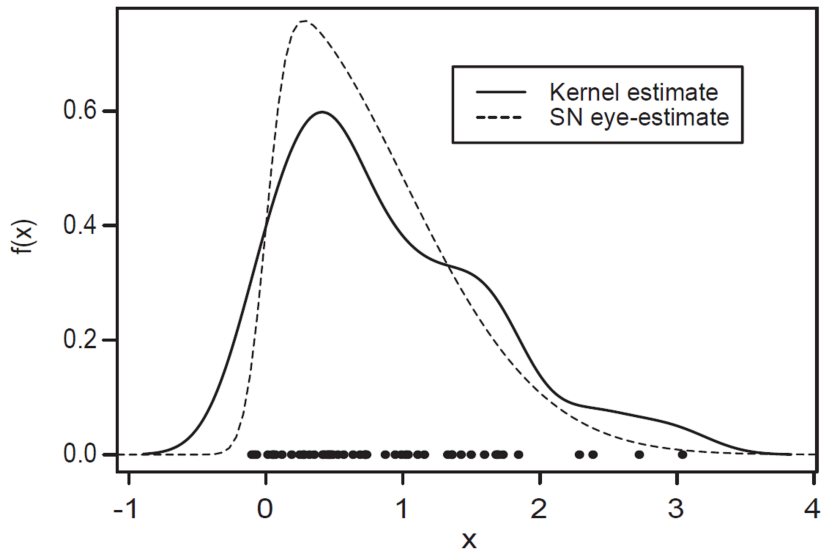

Figure 4.11: Simulated data points (small circles) leading to $\hat{\lambda} = \infty$ with non-parametric density estimate (continuous curve) and parametric SN density with shape parameter $\lambda = 8.14$ (dashed curve).

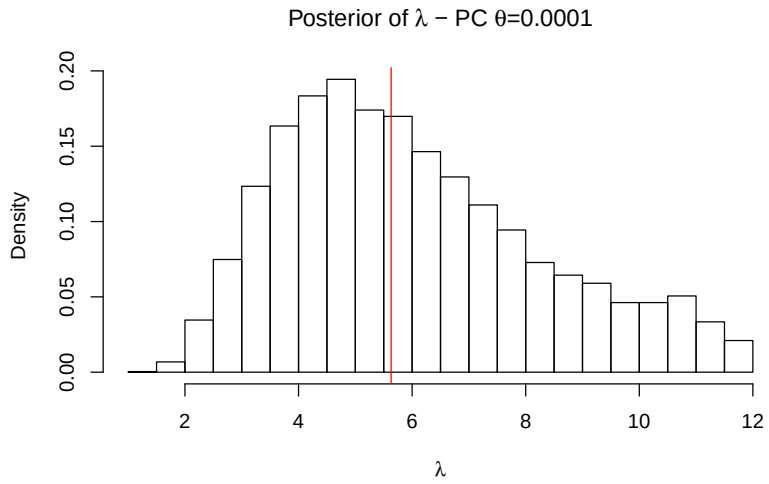

Figure 4.12: Simulated posterior of λ for the frontier data set using a PC prior with $\theta = 0.0001$. The mean is 6.015031 and the median 5.630749.

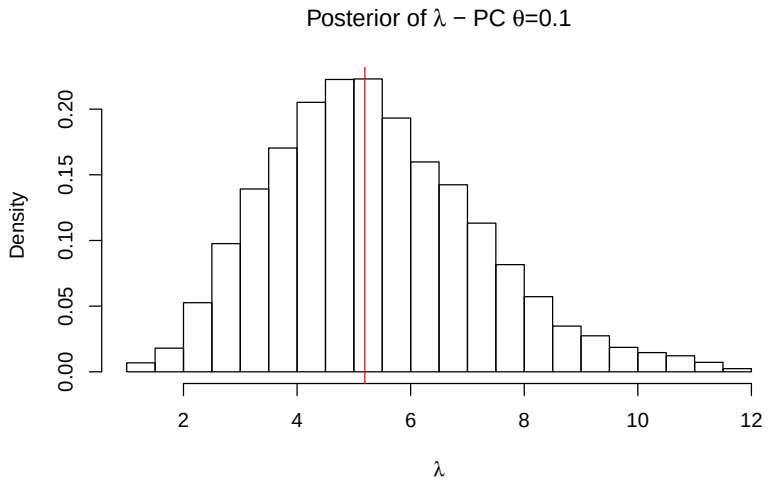

Figure 4.13: Simulated posterior of λ for the frontier data set using a PC prior with $\theta = 0.1$. The mean is 5.373173 and the median 5.193114.

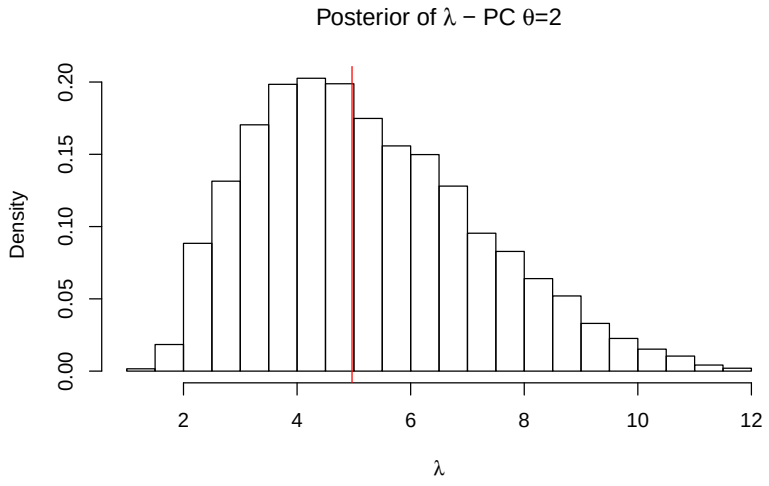

Figure 4.14: Simulated posterior of λ for the frontier data set using a PC prior with $\theta = 2$. The mean is 5.246213 and the median 4.967431.

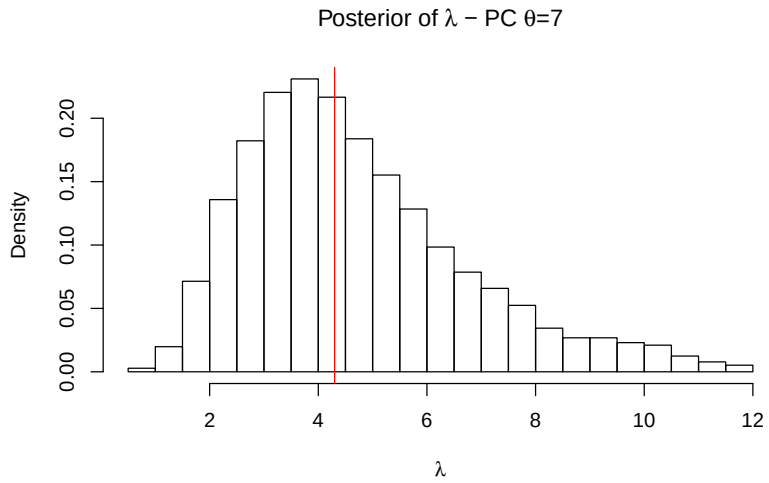

Figure 4.15: Simulated posterior of λ for the frontier data set using a PC prior with $\theta = 7$. The mean is 4.7024 and the median 4.300433.

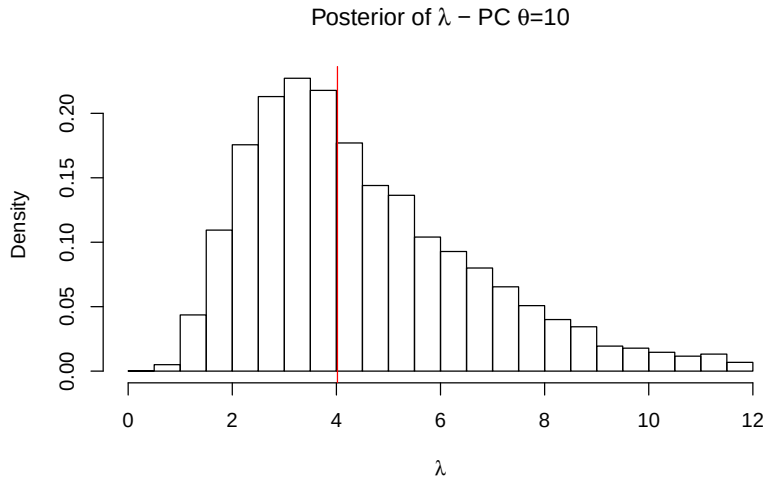

Figure 4.16: Simulated posterior of λ for the frontier data set using a PC prior with $\theta = 10$. The mean is 4.50517 and the median 4.020194.

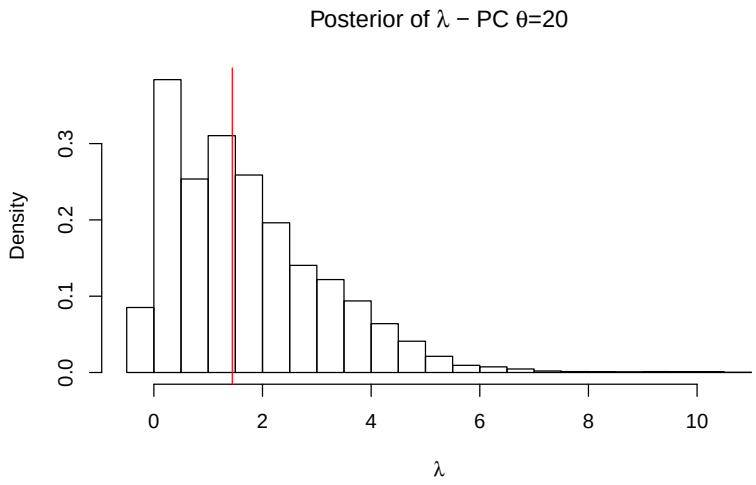

Figure 4.17: Simulated posterior of λ for the frontier data set using a PC prior with $\theta = 0.1$. The mean is 1.699234 and the median 1.443476.

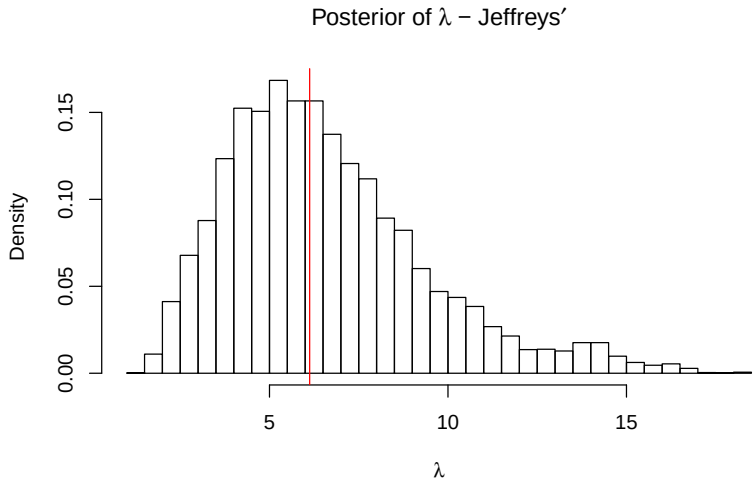

Figure 4.18: Simulated posterior of λ for the frontier data set using a Jeffereys' prior. The mean is 6.598168 and the median 6.125505.

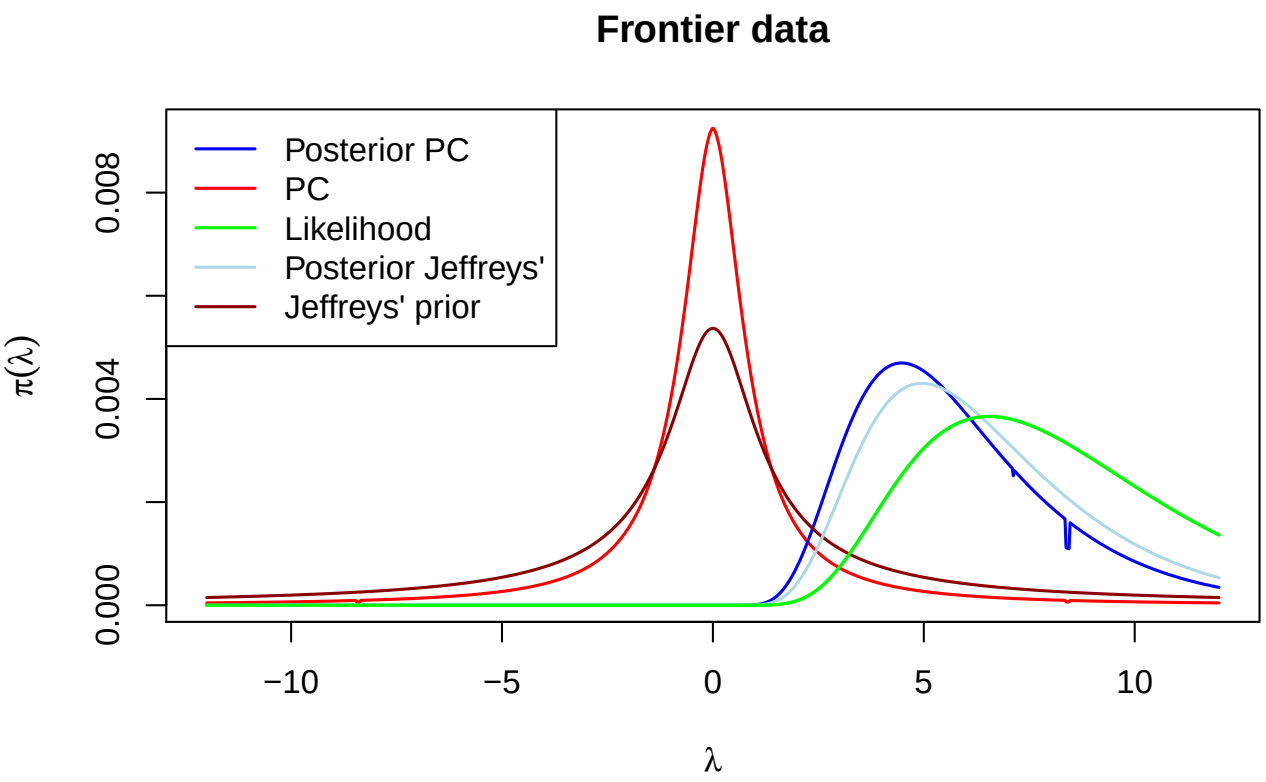

Figure 4.19: Numerical posteriors, priors and likelihood for the frontier data set. The PC prior has parameter $\theta = 0.1$.

is simulated by means of a Random Walk Metropolis-Hastings and the median is computed as a summary (red line).
The histograms are illustrative of the fact that, after a certain level of shrinkage, the more the shrinkage of the PC prior the less the accuracy in the estimation. Sure, it also depends on the very strong skewness enclosed into the frontier data set, although we already know how bad can be choosing a large value of θ. All the histograms are positive skewed, as we may see from the figures. However, the PC prior does a good job in the frontier data set as long as we keep small the value of θ.
In Figure ?? we roughly show the Jeffreys' and PC priors for the frontier data set. The curves are numerically computed since we do not know their density functions. In this case, the PC prior has a rate parameter $\theta = 0.1$. We can appreciate also in this graph the heavier right tale of the Jeffreys' prior distribution that is reflected in the posterior distribution as well, as we may also see from the histograms above.

4.6 PC Prior for the Degrees of Freedom in the Student-t case

In this section we will analyse the Student-t distribution focusing solely on the degree of freedom parameter v. This is an important non-trivial case, since the Student-t distribution is often used to robustify the Gaussian distribution.
The base model for the Student-t distribution is the Gaussian, which occurs when $\nu = \infty$. The Occam's razor principle implies that the mode of $\pi_d(d)$ must be at $d = 0$, corresponding to the Gaussian distribution. Note that $\pi_d(d)$ stand for the prior on the distance scale. It turns out that any proper prior for ν with finite expectation violates this principle. The intuition is that if we want $\nu = \infty$ to be central in the prior, a finite expectation will bound the tail behaviour so that we cannot have the mode (or a non-zero density) at $d = 0$. This implies that Occam's razor does not apply, in the sense that the PC prior defines that the posterior will shrink towards the respective mode rather than towards the Gaussian base model. If the true distribution was Gaussian then we would overfit the data. Only for the rate parameter $\theta \to \infty$ of the exponential distribution that we assign to the distance scale, we would have the most of the mass at the base model, but in this case the expectation would not be finite.
? perform an approximation of the KLD(ν) as they did not find closed form for the KLD. For $2 < \nu < 9$, they tabulated the prior using numerical integration. For $\nu > 9$, they compute the PC prior using the following asymptotic expansion, which has absolute error less than 1.6×10^{-10}.

$$\begin{aligned}\text{KLD}(\nu) = &\frac{3}{4}\nu^{-2} + \frac{3}{2}\nu^{-3} + \frac{17}{8}\nu^{-4} + \frac{29}{10}\nu^{-5} + \frac{61}{12}\nu^{-6} + \frac{145}{14}\nu^{-7} + \frac{273}{16}\nu^{-8} + \\ &\frac{119}{6}\nu^{-9} + \frac{869}{20}\nu^{-10} + \frac{4121}{22}\nu^{-11} + \frac{6169}{24}\nu^{-12} - \frac{30035}{26}\nu^{-13} - \frac{21843}{28}\nu^{-14} + \\ &\frac{320779}{10}\nu^{-15} + \frac{995105}{32}\nu^{-16} - \frac{28689547}{34}\nu^{-17} - \frac{28558475}{36}\nu^{-18} + \\ &\frac{1110177193}{38}\nu^{-19} + \frac{1110701481}{04}\nu^{-20} + \mathcal{O}(\nu^{-21}). \quad (4.48)\end{aligned}$$

We can exploit Theorem ?? to derive an analytical expression of the KLD and consequently obtain the PC prior for ν. In addition, we show once again the invariance of the KLD with respect to the location-scale structure.
Thus, suppose to have a flexible model represented by a Student-t distribution, where the base model is the Normal distribution. In this case we can deal solely with the degrees of freedom, since the resulting PC prior is invariant to location and scale parameters. The reason of such an invariance can be easily explained by the following change of variable in the Kullback-Leibler divergence

$$\begin{aligned}\int_{\mathcal{X}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}} \\ \cdot \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}\right) dx \\ -\int_{\mathcal{X}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}} \log\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right) dx, \quad (4.49)\end{aligned}$$

where $Z = \frac{X-\mu}{\sigma}$, so that we have

$$\begin{aligned}\int_{\mathcal{Z}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}} \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right) dz \\ -\int_{\mathcal{Z}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}} \log\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\right) dz. \quad (4.50)\end{aligned}$$

As we can see, the Kullback-Leibler divergence does not depend on the location and scale parameters and as a consequence the PC prior does not depend too.
Given the previous result we can derive the PC prior for ν just considering the standard versions of the complex and base models, i.e. where the location and scale parameters are 0 and 1 respectively.
Then, the KLD is

$$\mathrm{KLD}(\nu) = \int_{\mathcal{X}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \log\left(\frac{\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}}\right) dx, \quad (4.51)$$

and it is easy to notice that we can split the integral as follows

$$\begin{aligned}\mathrm{KLD}(\nu) = \int_{\mathcal{X}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right) dx \\ -\int_{\mathcal{X}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \log\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\right) dx. \quad (4.52)\end{aligned}$$

Now, we can easily distinguish what the two integrals represent. The first integral is equal to the minus entropy of a Student-t distribution, while the second one can be written as the second moment of a Student-t distribution plus the logarithm of the normalizing constant of a standard normal distribution. In practice, we are considering equation (??) of Theorem ??.
Therefore

$$\mathrm{KLD}(\nu) = -\frac{\nu+1}{2}\left[\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right] - \log\left[\sqrt{\nu}\, B\left(\frac{\nu}{2}, \frac{1}{2}\right)\right] + \frac{1}{2}\frac{\nu}{\nu-2} - \log\left(\frac{1}{\sqrt{2\pi}}\right), \quad (4.53)$$

where Ψ is the digamma function and B is the beta function.
The resulting prior is defined only for $\nu > 2$ since the second moment of the Student-t distribution exists only for more than two degrees of freedom.
Then, the double of the KLD is

$$\mathrm{A}(\nu) = -(\nu+1)\left[\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right] - 2\log\left[\sqrt{\nu}\, B\left(\frac{\nu}{2}, \frac{1}{2}\right)\right] + \frac{\nu}{\nu-2} - 2\log\left(\frac{1}{\sqrt{2\pi}}\right), \quad (4.54)$$

and the distance

$$d(\nu) = \sqrt{\mathrm{A}(\nu)}. \quad (4.55)$$

The PC prior for the degrees of freedom is obtained by multiplying the exponential prior assigned to the distance scale by the absolute value of the derivative of such a distance

$$\pi(\nu) = \theta \exp(-\theta d(\nu)) \left|\frac{\partial d(\nu)}{\partial \nu}\right|, \quad (4.56)$$

where

$$\frac{\partial d(\nu)}{\partial \nu} = \frac{\frac{1}{4}\left(-\frac{2}{\nu} - \frac{4}{(\nu-2)^2} + (\nu+1)\Psi^{(1)}\left(\frac{\nu}{2}\right) - (\nu+1)\Psi^{(1)}\left(\frac{\nu+1}{2}\right)\right)}{d(\nu)}, \quad (4.57)$$

where, in turn, the numerator is the derivative of the Kullback-Leibler divergence and $\Psi^{(1)}$ is the trigamma function. The PC prior for ν is easily obtained

$$\pi(\nu) = \theta \exp(-\theta\sqrt{\mathrm{A}(\nu)}) \frac{\left|\frac{1}{4}\left(-\frac{2}{\nu} - \frac{4}{(\nu-2)^2} + (\nu+1)\Psi^{(1)}\left(\frac{\nu}{2}\right) - (\nu+1)\Psi^{(1)}\left(\frac{\nu+1}{2}\right)\right)\right|}{\sqrt{\mathrm{A}(\nu)}}. \quad (4.58)$$

So, exploiting the alternative representation of the KLD we are able to derive the analytical expression of the PC prior for ν.
Figure ?? shows the PC prior obtained from the approximation of the KLD in (??) (black curve) and our PC prior derived from the alternative representation of the KLD (blue curve).

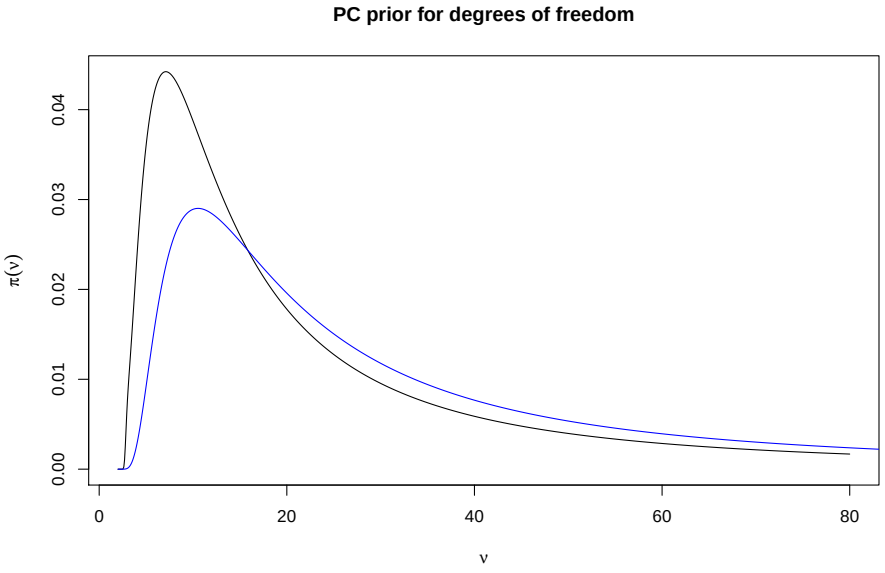

Figure 4.20: Simpson's approximation of $\pi^{PC}(\nu)$(black) vs our prior (blue), where the rate parameter $\theta = 10$

Chapter 5

Extending the Univariate PC Prior

? propose also an extension of the univariate PC prior to the multivariate setting $\underline{\xi}$, with base model $\underline{\xi} = \underline{0}$. The multivariate extension proposed by **?** preserves all the features of the univariate case. Given that many multivariate parameters spaces are not $\mathbb{R}^n$, we will let $\mathcal{M}$ be a subset of a smooth n-dimensional manifold.
First of all, assume that $d(\underline{\xi})$ has a non-vanishing Jacobian. For each $r \geq 0$, the level sets $\underline{\gamma} \in S_r = \{\underline{\xi} \in \mathcal{M} : d(\underline{\xi}) = r\}$ are a system of disjoint embedded submanifolds of $\mathcal{M}$. Roughly speaking, the submanifolds S_r are the leaves of a foliation. To better understand, consider, for instance, that in the bivariate case the KLD would be a sort of cone; so the cone is cut in many slices where each slice has a uniform distribution. Hence the natural lifting of the PC prior concept onto $\mathcal{M}$ is the prior that is exponentially distributed in $d(\underline{\xi})$ and uniformly distributed on the leaves $S_{d(\underline{\xi})}$. Then, we should find a mapping $\varphi(\cdot)$ such that $(d(\underline{\xi}), \varphi(\underline{\xi})) = \underline{g}(\underline{\xi})$. This mapping allow us to get a local representation for the multivariate PC prior as

$$\pi(\underline{\xi}) = \frac{\lambda}{|S_{d(\underline{\xi})}|} \exp(-\lambda d(\underline{\xi}))|\det(\mathbf{J}(\underline{\xi}))|, \tag{5.1}$$

where $J_{ij} = \frac{\partial g_i}{\partial \xi_j}$ is the Jacobian of $\underline{g}$. Computational geometry tools can be used to evaluate (**??**) approximately in low dimensions.
Anyhow, when the level sets are simplexes or spheres, exact expressions for the PC prior can be found. These situations occur when $d(\underline{\xi})$ is a linear or a quadratic function, i.e.

$$d(\underline{\xi}) = h(\underline{b}^T \underline{\xi}), \quad \underline{b} > 0, \quad \underline{\xi} \in \mathbb{R}^n_+ \tag{5.2}$$

or

$$d(\underline{\xi}) = h(\frac{1}{2}\underline{\xi}^T \mathbf{H} \underline{\xi}), \quad \mathbf{H} > 0, \quad \underline{\xi} \in \mathbb{R}^n, \tag{5.3}$$

for some function $h(\cdot)$ satisfying $h(0) = 0$, typically $h(a) = \sqrt{2a}$. The linear case is useful for deriving the PC prior for general correlation matrices. Think for instance of a multivariate Gaussian copula where the marginals have different

pair correlations, so we can change the parameterisation and get the distance $d(\underline{\xi})$ as a linear function.
In the linear case with $\underline{b} = \underline{1}$, the PC prior is

$$\pi(\underline{\xi}) = \lambda \exp(-\lambda d(\underline{\xi})) \frac{(n-1)!}{r(\underline{\xi})^{n-1}} |h'(r(\underline{\xi}))|, \quad r(\underline{\xi}) = h^{-1}(d(\underline{\xi})), \tag{5.4}$$

while in the quadratic case with $\mathbf{H} = \mathbf{I}$, the PC prior is

$$\pi(\underline{\xi}) = \lambda \exp(-\lambda d(\underline{\xi})) \frac{\Gamma\left(\frac{n}{2}+1\right)}{n\pi^{\frac{n}{2}} r(\underline{\xi})^{n-2}} \left| h'\left(\frac{1}{2} r(\underline{\xi})^2\right) \right|, \quad r(\underline{\xi}) = \sqrt{2h^{-1}(d(\underline{\xi}))}. \tag{5.5}$$

Anyhow, the general multivariate case for the PC prior is hard. In our opinion, there are some issues related to (**??**). First of all, we want to compute a prior for many parameters but we have only one distance and the distribution we assign to such a distance is a univariate exponential density function. Then, we should know the level sets $S_r = \{\underline{\xi} \in \mathcal{M} : d(\underline{\xi}) = r\}$ and their geometry, in order to define the local mapping $\varphi(\cdot)$ that allows us to build the Jacobian matrix. In our opinion this latter is the most difficult part, especially from a computational point of view. Finally, formula (**??**) is very difficult to compute for high dimensional models, apart from the cases where the distance function is linear or quadratic, so something more easy is needed, especially for computational purposes.
In the next section we explore the construction of the multivariate PC prior via the Hammersley-Clifford theorem and in the last section we propose to use a copula to connect the univariate PC priors. Anyhow, for orthogonal parameters the multivariate PC prior is simply the product of univariate PC prior distributions.

5.1 The Hammersley-Clifford Theorem

An interesting property of the full conditionals, which the Gibbs sampler is based on, is that they fully specify the joint distribution, as Hammersley and Clifford proved in 1970[1]. Note that the set of marginal distributions does not have this property.

Definition 2 (Positivity condition)**.** A distribution with density $f(x_1, \ldots, x_p)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the positivity condition if $f(x_1, \ldots, x_p) > 0$ for all $x_1, \ldots, x_p$ with $f_{X_i}(x_i) > 0$.

The positivity condition thus implies that the support of the joint density f is the Cartesian product of the support of the marginals f_{X_i}.

Theorem 3 (Hammersley-Clifford)**.** *Let $(X_1, \ldots, X_p)$ satisfy the positivity condition and have joint density $f(x_1, \ldots, x_p)$. Then for all $(\xi_1, \ldots, \xi_p) \in supp(f)$*

$$f(x_1, \ldots, x_p) \propto \prod_{j=1}^{p} \frac{f_{X_j|X_{-j}}(x_j|x_1, \ldots, x_{j-1}, \xi_{j+1}, \ldots, \xi_p)}{f_{X_j|X_{-j}}(\xi_j|x_1, \ldots, x_{j-1}, \xi_{j+1}, \ldots, \xi_p)} \tag{5.6}$$

[1] Hammersley and Clifford actually never published this result, as they could not extend the theorem to the case of non-positivity.

Proof. We have

$$f(x_1, \ldots, x_{p-1}, x_p) = f_{X_p|X_{-p}(x_p|x_1,\ldots,x_{p-1})f(x_1,\ldots,x_{p-1})} \tag{5.7}$$

and by complete analogy

$$f(x_1, \ldots, x_{p-1}, \xi_p) = f_{X_p|X_{-p}(\xi_p|x_1,\ldots,x_{p-1})f(x_1,\ldots,x_{p-1})} \tag{5.8}$$

thus

$$\begin{aligned} f(x_1, \ldots, x_p) &\overset{(??)}{=} \underbrace{f(x_1, \ldots, x_{p-1})}_{\overset{(??)}{=} f(x_1,\ldots,x_{p-1},\xi_p)/f_{X_p|X_{-p}}(\xi_p|x_1,\ldots,x_{p-1})} f_{X_p|X_{-p}}(x_p|x_1, \ldots, x_{p-1}) \\ &= f(x_1, \ldots, x_{p-1}, \xi_p) \frac{f_{X_p|X_{-p}}(x_p|x_1, \ldots, x_{p-1})}{f_{X_p|X_{-p}}(\xi_p|x_1, \ldots, x_{p-1})} \\ &= \ldots \\ &= f(\xi_1, \ldots, \xi_p) \frac{f_{X_1|X_{-1}}(x_1|\xi_2, \ldots, \xi_p)}{f_{X_1|X_{-1}}(\xi_1|\xi_2, \ldots, \xi_p)} \cdots \frac{f_{X_p|X_{-p}}(x_p|x_1, \ldots, x_{p-1})}{f_{X_p|X_{-p}}(\xi_p|x_1, \ldots, x_{p-1})}. \end{aligned}$$

The positivity condition guarantees that the conditional densities are non-zero. □

Note that the Hammersley-Clifford theorem does not guarantee the existence of a joint probability distribution for every choice of the conditionals, as the following example shows. In Bayesian modeling such problems mostly arise when using improper prior distributions.

Example 1. Consider the following model

$$\begin{aligned} X_1|X_2 &\sim \text{Exp}(\lambda X_2) \\ X_2|X_1 &\sim \text{Exp}(\lambda X_1) \end{aligned}$$

for which it would be easy to design a Gibbs sampler. Trying to apply the Hammersley-Clifford theorem, we obtain

$$f(x_1, x_2) \propto \frac{f_{X_1|X_2}(x_1|\xi_2) \cdot f_{X_2|X_1}(x_2|x_1)}{f_{X_1|X_2}(\xi_1|\xi_2) \cdot f_{X_2|X_1}(\xi_2|x_1)} = \frac{\lambda\xi_2 e^{-\lambda x_1\xi_2} \cdot \lambda x_1 e^{-\lambda x_1 x_2}}{\lambda\xi_2 e^{-\lambda\xi_1\xi_2} \cdot \lambda x_1 e^{-\lambda x_1\xi_2}} \propto e^{-\lambda x_1 x_2}.$$

The integral $\int\int \exp(-\lambda x_1 x_2) dx_1 dx_2$ however is not finite, thus there is no two-dimensional probability distribution with $f(x_1, x_2)$ as its density.

5.1.1 Joint PC Prior for the Uniform Model

As another example consider a random vector having as base model an Uniform distribution on the unit square

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{Unif}\left[(0,1) \times (0,1)\right],$$

while the more flexible model has random edges a and b less than or equal to 1

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{Unif}\left[(0,a) \times (0,b)\right].$$

If we considered values of a and b greater than 1 we would have problems with the positive definiteness of the KLD, since we would have a nonregular model. Before applying the Hammersley-Clifford theorem we need to compute the full conditional PC priors. Let us compute the conditional KLDs

$$\text{KLD}(a|b) = \int_0^a \int_0^b \frac{1}{ab} \log\left(\frac{1}{ab}\right) dxdy = -\log(ab), \tag{5.9}$$

where b is not a random variable, rather it is just a parameter with values between 0 and 1. Obviously, the same result holds when computing the conditional KLD of b given the parameter a

$$\text{KLD}(b|a) = \int_0^a \int_0^b \frac{1}{ab} \log\left(\frac{1}{ab}\right) dxdy = -\log(ab). \tag{5.10}$$

Hence, the distances are the same

$$d(a|b) = \sqrt{-2\log(ab)} \tag{5.11}$$

$$d(b|a) = \sqrt{-2\log(ab)} \tag{5.12}$$

Now, it is easy to compute the conditional PC priors. Let us do it just once, since the same holds for the other full conditional

$$\pi(a|b) = \mu e^{-\mu\sqrt{-2\log(ab)}} \left| -\frac{\frac{2}{a}}{2\sqrt{-2\log(ab)}} \right|, \tag{5.13}$$

where μ is the rate parameter of the exponential distribution assigned to the distance scale.
Then

$$\pi(a|b) = \frac{\mu}{a} e^{-\mu\sqrt{-2\log(ab)}} \frac{1}{\sqrt{-2\log(ab)}}. \tag{5.14}$$

The prior for b given a with rate parameter θ is

$$\pi(b|a) = \frac{\theta}{b} e^{-\theta\sqrt{-2\log(ab)}} \frac{1}{\sqrt{-2\log(ab)}}. \tag{5.15}$$

Suppose to set $a^* = b^* = \frac{1}{2}$, then according to the Hammersley-Clifford theorem the joint density can be written as

$$\pi(a,b) \propto \frac{\pi(a|b^*) \cdot \pi(b|a)}{\pi(a^*|b^*) \cdot \pi(b^*|a)} \tag{5.16}$$

$$= \frac{\frac{\mu}{a} e^{-\mu\sqrt{-2\log\left(\frac{a}{2}\right)}} \frac{1}{\sqrt{-2\log\left(\frac{a}{2}\right)}} \cdot \frac{\theta}{b} e^{-\theta\sqrt{-2\log(ab)}} \frac{1}{\sqrt{-2\log(ab)}}}{2\mu e^{-\mu\sqrt{-2\log\left(\frac{1}{4}\right)}} \frac{1}{\sqrt{-2\log\left(\frac{1}{4}\right)}} \cdot 2\theta e^{-\theta\sqrt{-2\log\left(\frac{a}{2}\right)}} \frac{1}{\sqrt{-2\log\left(\frac{a}{2}\right)}}}$$

$$\propto \frac{1}{ab} e^{-(\mu-\theta)\sqrt{-2\log\left(\frac{a}{2}\right)}} e^{-\theta\sqrt{-2\log(ab)}} \frac{1}{\sqrt{-2\log(ab)}}. \tag{5.17}$$

The choice of $\mu = \theta = \lambda$ will provide a symmetric prior

$$\pi(a,b) = \frac{1}{ab}\exp\left(-\lambda\sqrt{-2\log(ab)}\right)\frac{1}{\sqrt{-2\log(ab)}}. \tag{5.18}$$

Let us study the behaviour of the prior at the limits

$$\text{if} \quad a,b \to 1 \begin{cases} \frac{1}{ab} \to 1 \\ \exp\left(-\lambda\sqrt{-2\log(ab)}\right) \to 1 \\ \frac{1}{\sqrt{-2\log(ab)}} \to \infty \end{cases}, \tag{5.19}$$

on the other hand

$$\text{if} \quad a,b \to 0 \begin{cases} \frac{1}{ab} \to \infty \\ \exp\left(-\lambda\sqrt{-2\log(ab)}\right) \to 0 \\ \frac{1}{\sqrt{-2\log(ab)}} \to 0 \end{cases}. \tag{5.20}$$

The Hammersley-Clifford construction creates two questions. The first one is related to the positiveness of the resulting joint distribution, while the second one is related to the integrability of the joint density itself.
Let us integrate the density (**??**)

$$\int_0^1\int_0^1 \frac{\exp\left(-\lambda\sqrt{-2\log(ab)}\right)}{ab\sqrt{-2\log(ab)}}dadb = \frac{1}{\lambda^3}, \tag{5.21}$$

so, the proper joint PC prior should be written as

$$\pi(a,b) = \frac{\lambda^3}{ab}\exp\left(-\lambda\sqrt{-2\log(ab)}\right)\frac{1}{\sqrt{-2\log(ab)}}. \tag{5.22}$$

Finally, both the conditions are satisfied. This means that the aforementioned construction makes sense for the Uniform model.

5.1.2 PC Prior for the mean vector of the Bivariate Normal distribution

As a second example, let us consider the PC prior for the vector of the means in the bivariate Gaussian distribution, where the covariance matrix is assumed to be the identity matrix.
Therefore, suppose to have the base model

$$f_0(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}, \tag{5.23}$$

and the more flexible model given by the introduction of the mean parameters μ_x and μ_y

$$f_1(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]}. \tag{5.24}$$

Then, the Kullback-Leibler divergence between the two models is

$$\mathrm{KLD}(f_1\|f_0) = \iint_{\mathbb{R}^2} \frac{1}{2\pi} e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]} \times \log\left(e^{-\frac{1}{2}\left[(x-\mu_x)^2-x^2+(y-\mu_y)^2-y^2\right]}\right) dxdy, \quad (5.25)$$

or equivalently

$$\mathrm{KLD}(\mu_x,\mu_y) = \iint_{\mathbb{R}^2} \frac{1}{2\pi} e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]} \times -\frac{1}{2}\left[\mu_x^2 - 2x\mu_x + \mu_y^2 - 2y\mu_y\right] dxdy. \quad (5.26)$$

Now, we can write

$$\begin{aligned}\mathrm{KLD}(f_1\|f_0) &= -\frac{1}{2}\mu_x^2 - \frac{1}{2}\mu_y^2 + \mu_x\mathbb{E}^{\sim f_1}[X] + \mu_y\mathbb{E}^{\sim f_1}[Y] \\ &= -\frac{1}{2}\mu_x^2 - \frac{1}{2}\mu_y^2 + \mu_x^2 + \mu_y^2 \\ &= \frac{\mu_x^2+\mu_y^2}{2}. \end{aligned} \quad (5.27)$$

Notice that the KLD has the same expression both for μ_x given μ_y and for μ_y given μ_x. In fact, when we consider either μ_x or μ_y as a parameter, the KLD is just a function of the other random variable.
It follows that the distance is

$$d(\mu_x,\mu_y) = \sqrt{\mu_x^2+\mu_y^2}. \quad (5.28)$$

Let us compute now the conditional PC prior for μ_x given μ_y

$$\begin{aligned}\pi^{PC}(\mu_x|\mu_y) &= \frac{\lambda_x}{2}\exp\left(-\lambda_x\sqrt{\mu_x^2+\mu_y^2}\right)\frac{1}{2\sqrt{\mu_x^2+\mu_y^2}}2|\mu_x| \\ &= \frac{\lambda_x}{2}\exp\left(-\lambda_x\sqrt{\mu_x^2+\mu_y^2}\right)\frac{|\mu_x|}{\sqrt{\mu_x^2+\mu_y^2}}. \end{aligned} \quad (5.29)$$

Equivalently, the PC prior for μ_y given μ_x is

$$\pi^{PC}(\mu_y|\mu_x) = \frac{\lambda_y}{2}\exp\left(-\lambda_y\sqrt{\mu_x^2+\mu_y^2}\right)\frac{|\mu_y|}{\sqrt{\mu_x^2+\mu_y^2}}. \quad (5.30)$$

Finally, we have all the ingredients to apply the Hammersley-Clifford theorem

$$\begin{aligned}\pi^{PC}(\mu_x,\mu_y) &\propto \frac{\pi^{PC}(\mu_x|\tilde{\mu_y})\cdot\pi^{PC}(\mu_y|\mu_x)}{\pi^{PC}(\tilde{\mu_x}|\tilde{\mu_y})\cdot\pi^{PC}(\tilde{\mu_y}|\mu_x)} \\ &= \frac{\frac{\lambda_x}{2}e^{-\lambda_x\sqrt{\mu_x^2+\tilde{\mu_y}^2}}\frac{|\mu_x|}{\sqrt{\mu_x^2+\tilde{\mu_y}^2}}\cdot\frac{\lambda_y}{2}e^{-\lambda_y\sqrt{\mu_x^2+\mu_y^2}}\frac{|\mu_y|}{\sqrt{\mu_x^2+\mu_y^2}}}{\frac{\lambda_x}{2}e^{-\lambda_x\sqrt{\tilde{\mu_x}^2+\tilde{\mu_y}^2}}\frac{|\tilde{\mu_x}|}{\sqrt{\tilde{\mu_x}^2+\tilde{\mu_y}^2}}\cdot\frac{\lambda_y}{2}e^{-\lambda_y\sqrt{\mu_x^2+\tilde{\mu_y}^2}}\frac{|\tilde{\mu_y}|}{\sqrt{\mu_x^2+\tilde{\mu_y}^2}}} \\ &\propto \frac{|\mu_x\mu_y|}{\sqrt{\mu_x^2+\mu_y^2}}e^{-\lambda_x\sqrt{\mu_x^2+\tilde{\mu_y}^2}+\lambda_y\sqrt{\mu_x^2+\tilde{\mu_y}^2}-\lambda_y\sqrt{\mu_x^2+\mu_y^2}}. \end{aligned} \quad (5.31)$$

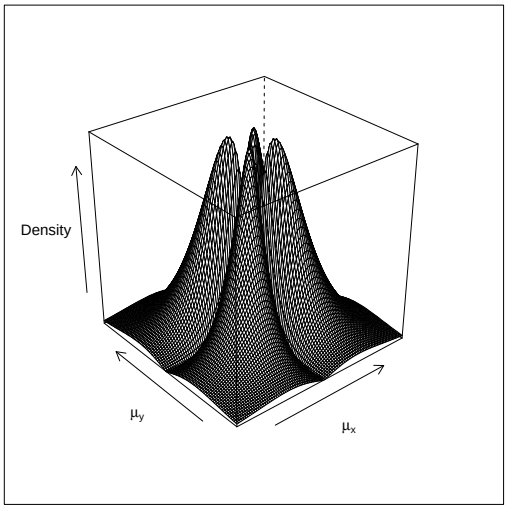

Figure 5.1: Joint PC prior for the mean vector of a bivariate Gaussian density, where $\lambda = 0.1$.

Now, let $\lambda_x = \lambda_y = \lambda$, then

$$\pi^{PC}(\mu_x, \mu_y) \propto \frac{|\mu_x \mu_y|}{\sqrt{\mu_x^2 + \mu_y^2}} e^{-\lambda\sqrt{\mu_x^2+\mu_y^2}}. \tag{5.32}$$

Figure ?? shows the joint PC prior obtained via the Hammersley-Clifford theorem, where the rate parameter λ is set equal to 0.1. We may notice that the Hammersley-Clifford construction makes the joint PC prior similar to a non-local prior (?).
Alternatively, we can express the parameterization in terms of the polar coordinates

$$\begin{cases} \mu_x = r\cos\varphi \\ \mu_y = r\sin\varphi \\ \mu_x^2 + \mu_y^2 = r^2 \end{cases} \tag{5.33}$$

so that our prior can be written as

$$\begin{aligned} \pi^{PC}(r, \varphi) &\propto \frac{r^2|\sin\varphi\cos\varphi|}{r}\exp(-\lambda r) \\ &= r\exp(-\lambda r)|\sin\varphi\cos\varphi|, \end{aligned} \tag{5.34}$$

where $r\exp(-\lambda r) \propto \text{Gamma}(r|\nu = 2, \lambda)$. Figure ?? shows the same PC prior with the parameterisation in terms of the polar coordinates, where $\varphi \in (0, 2\pi)$ and $r \in (0, 10)$; even in this case the rate parameter $\lambda = 0.1$.
Finally, let's make some clarification on the Hammersley-Clifford construction. We can choose any value of $\tilde{\mu_x}$ and $\tilde{\mu_y}$ belonging to the support of the joint density. They do not affect the kernel of the distribution as they are just constants; the joint density will result up to a proportionality constant. Anyhow, we recommend to do not choose values of $\tilde{\mu_x}$ and $\tilde{\mu_y}$ corresponding to the base

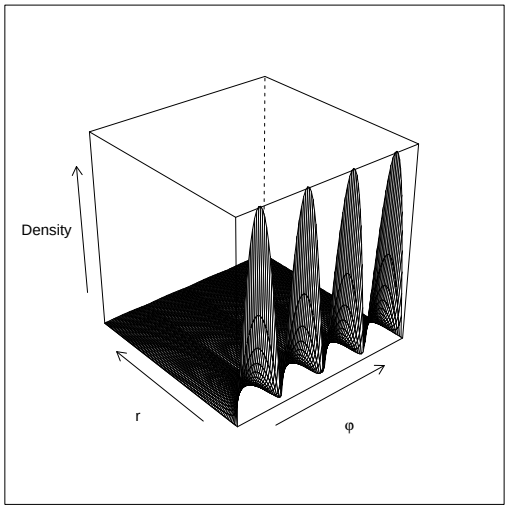

Figure 5.2: Joint PC prior in terms of the polar coordinates, φ and r, for the mean vector of a bivariate normal density, where $\lambda = 0.1$.

model, as one could incur in issues related to the existence of the joint density. In addition, it is worth to clarify that the non-locality of the PC prior constructed via the Hammersley-Clifford theorem is not a consequence of $\tilde{\mu_x}$ and $\tilde{\mu_y}$, rather a consequence of the conditional densities.

5.1.3 PC Prior in the Bivariate Skew-Normal Model

The scalar skew-normal density can be extended to the d-dimensional case by considering the following density function

$$f_d(\underline{x}; \Omega, \underline{\alpha}) = 2\phi_d(\underline{x}; \Omega)\Phi(\underline{\alpha}^{\mathrm{T}}\underline{x}), \qquad \underline{x} \in \mathbb{R}^d, \tag{5.35}$$

where Ω is a positive-definite $d \times d$ correlation matrix, $\phi_d(\underline{x}; \Sigma)$ is the density function of a $\mathrm{N}_d(0, \Sigma)$ variate and α is a d-dimensional vector parameter.
Our aim is to derive a joint density function for the elements contained in the vector parameter $\underline{\alpha}$, i.e. $\alpha_1, \alpha_2, \ldots, \alpha_d$. Notice that we refer to a variable Z with density (??) as a *normalized multivariate skew-normal* variate. In the cases where we deal with a bivariate skew-normal density, the parameters of interest are α_1 and α_2.
To make the multivariate skew-normal more concrete we have a closer look at the bivariate skew-normal distribution.
From equation (??), we define the bivariate skew-normal as

$$f(x_1, x_2; \alpha_1, \alpha_2, \omega) = 2\phi_2(x_1, x_2; \omega)\Phi(\alpha_1 x_1 + \alpha_2 x_2), \tag{5.36}$$

where ω is the off-diagonal element of Ω.
The components of $\underline{\alpha}$ are

$$\alpha_1 = \frac{\delta_1 - \delta_2\omega}{\sqrt{(1-\omega^2)(1-\omega^2-\delta_1^2-\delta_2^2+2\delta_1\delta_2\omega)}}, \tag{5.37}$$

$$\alpha_2 = \frac{\delta_2 - \delta_1\omega}{\sqrt{(1-\omega^2)(1-\omega^2-\delta_1^2-\delta_2^2+2\delta_1\delta_2\omega)}}. \tag{5.38}$$

If we know the marginal distributions $SN(\alpha(\delta_1))$, $SN(\alpha(\delta_2))$ and ω it is possible to calculate the density of the bivariate skew-normal distribution. Note that $\alpha(\delta_j) = \frac{\delta_j}{\sqrt{1-\delta_j^2}} \neq \alpha_j$.
The following proposition provides the marginals of a bivariate skew-normal distribution.

Lemma 1. Suppose $\mathbf{X} = (X_1, X_2)^T \sim SN_2(x_1, x_2; \alpha_1, \alpha_2, \omega)$. Then $X_i \sim SN(\bar{\alpha}_i)$, $i = 1, 2$, where

$$\bar{\alpha}_1 = \frac{\alpha_1 + \omega\alpha_2}{\sqrt{1+(1-\omega^2)\alpha_2^2}}, \tag{5.39}$$

$$\bar{\alpha}_2 = \frac{\alpha_2 + \omega\alpha_1}{\sqrt{1+(1-\omega^2)\alpha_1^2}}. \tag{5.40}$$

Note that ω can be derived from the covariance or the correlation between X_1 and X_2. The relations for the bivariate case are the following

$$Cov(X_1, X_2) = \omega - \frac{2}{\pi}\delta_1\delta_2, \tag{5.41}$$

$$\rho_{12} = \frac{\omega - \frac{2}{\pi}\delta_1\delta_2}{\sqrt{(1-\frac{2}{\pi}\delta_1^2)(1-\frac{2}{\pi}\delta_2^2)}}. \tag{5.42}$$

Obviously, the parameter ω cannot be arbitrary, in the sense that the correlation matrix must be positive definite. In the bivariate case this leads to the following condition on ω

$$\delta_1\delta_2 - \sqrt{(1-\delta_1^2)(1-\delta_2^2)} < \omega < \delta_1\delta_2 + \sqrt{(1-\delta_1^2)(1-\delta_2^2)}, \tag{5.43}$$

and this condition gives also the attainable correlation level

$$\frac{(1-\frac{2}{\pi})\delta_1\delta_2 - \sqrt{(1-\delta_1^2)(1-\delta_2^2)}}{\sqrt{(1-\frac{2}{\pi}\delta_1^2)(1-\frac{2}{\pi}\delta_2^2)}} < \rho < \frac{(1-\frac{2}{\pi})\delta_1\delta_2 + \sqrt{(1-\delta_1^2)(1-\delta_2^2)}}{\sqrt{(1-\frac{2}{\pi}\delta_1^2)(1-\frac{2}{\pi}\delta_2^2)}}. \tag{5.44}$$

If $\delta_1 \neq \delta_2$, there exist values for which not any arbitrary correlation can be attained, i.e. $\rho_{\max} < 1$ and $\rho_{\min} > -1$. This is a well known property of the linear correlation for non-elliptic multivariate distributions.
Figure ?? shows attainable correlations ρ depending on $\delta = \delta_1 = \delta_2$. Note that for $\delta_1 \neq \delta_2$ also the maximal correlation can be smaller than one.
Figure ?? shows the joint PC prior, obtained via the Hammersley-Clifford theorem, for the vector $\underline{\alpha}$ of the bivariate skew-normal model. Notice that the prior

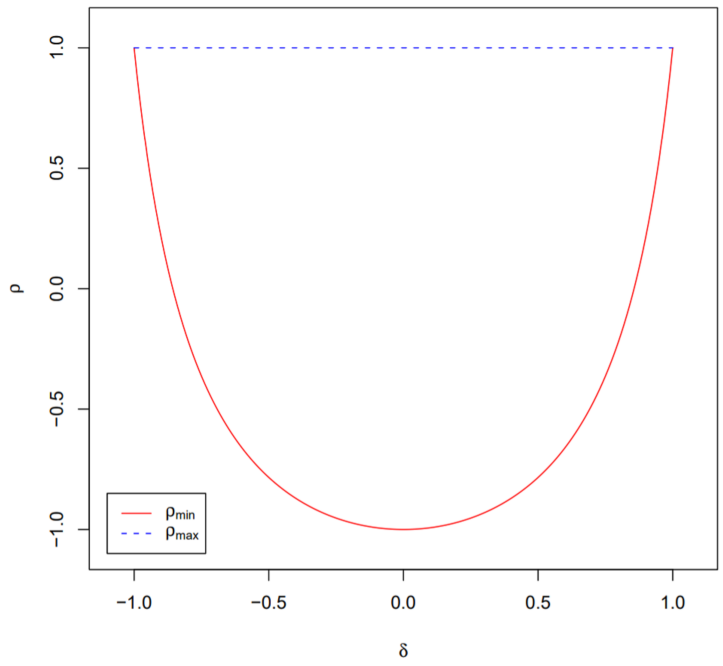

Figure 5.3: Attainable correlation ρ for $\delta = \delta_1 = \delta_2$.

must be numerically computed as the KLD has no closed form and it is given by

$$\text{KLD}(\alpha_1, \alpha_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2\phi_2(\mathbf{x})\Phi(\alpha_1 x_1 + \alpha_2 x_2) \log(2\Phi(\alpha_1 x_1 + \alpha_2 x_2))dx_1 dx_2. \tag{5.45}$$

In order to apply the Hammersley-Clifford theorem, we need to compute the univariate conditional PC priors and therefore we need to calculate the derivative of (**??**) both wrt α_1 and α_2, and to do that we use the Leibnitz's rule, i.e. we invert the integral and the derivative operators. In practice, we have

$$\frac{\partial \text{KLD}(\alpha_1, \alpha_2)}{\partial \alpha_1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2\phi_2(\mathbf{x})\phi(\alpha_1 x_1 + \alpha_2 x_2)x_1 \cdot (1 + \log(2\Phi(\alpha_1 x_1 + \alpha_2 x_2)))dx_1 dx_2, \tag{5.46}$$

and

$$\frac{\partial \text{KLD}(\alpha_1, \alpha_2)}{\partial \alpha_2} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2\phi_2(\mathbf{x})\phi(\alpha_1 x_1 + \alpha_2 x_2)x_2 \cdot (1 + \log(2\Phi(\alpha_1 x_1 + \alpha_2 x_2)))dx_1 dx_2. \tag{5.47}$$

Once again the integrals must be numerically computed.

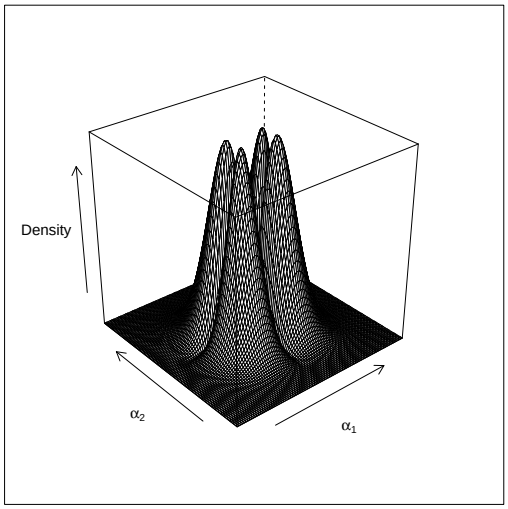

Figure 5.4: Joint PC prior for the vector $\underline{\alpha}$ of the bivariate skew-normal model, where the rate paramter $\lambda = 1$.

5.2 Copula based approach to construct a Multivariate PC Prior

In the previous section, we have explored the construction of the multivariate PC prior via the Hammersley-Clifford theorem. As we may see, apart from the Uniform model, the resulting multivariate PC prior exhibits non-locality in correspondence of the base model. So, whenever the base model is at the interior of the parameter space, the prior would show non-locality. Notice that this latter is a consequence of the non-locality of the conditional PC priors.
Therefore, the Hammersley-Clifford construction seems to lead to a sort of multivariate non-local prior. This could be useful in some situation, but in our case it is not adequate. In fact, we want the joint prior, penalising the distance from a base model, to have some mass at the base model itself in order to avoid overfitting (**?**).
In this section, we want to explore the construction of the multivariate PC prior more deeply, by looking at its embedded properties of the Kullback-Leibler divergence. Recall that in the bivariate case the KLD is a function of two parameters as it is obtained by considering as the base model the one where the two parameters themselves are absent.
Let's define now the KLD as a function of two generic parameters ξ_1, ξ_2

$$\mathrm{KLD}(\xi_1, \xi_2) = \mathrm{KLD}(\xi_1) + \mathrm{KLD}(\xi_1|\xi_2), \tag{5.48}$$

where $\mathrm{KLD}(\xi_1|\xi_2)$ is the conditional relative entropy. Equation (**??**) is validated by the following theorem.

Theorem 4 (Chain Rule for relative entropy)**.**

$$D(p(x,y)\|q(x,y)) = D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)). \tag{5.49}$$

Proof.

$$
\begin{aligned}
&D(p(x,y)\|q(x,y)) \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} && (5.50)\\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} && (5.51)\\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} && (5.52)\\
&= D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)). && (5.53)
\end{aligned}
$$

□

Another interesting property of the KLD is that the conditional relative entropy, $D(Y|X) \leq D(Y)$ (Information can't hurt), therefore $D(X,Y) \leq D(X) + D(Y)$, with the equality when X and Y are independent. Notice that, with abuse of notation, here D stands for the KLD. In addition, note that equation (**??**) can be also used to derive conditional PC priors based on the conditional relative entropy, as this latter is the difference between $\text{KLD}(\xi_1, \xi_2)$ and $\text{KLD}(\xi_1)$ or $\text{KLD}(\xi_2)$.
Whenever it happens that $D(X,Y) = D(X) + D(Y)$, the joint prior for the distance scales can be considered with independent components. So, in this case, the construction of the joint PC prior for a vector of parameters assumes orthogonality among the univariate distances and as a consequence among the marginal PC prior distributions, i.e. the joint PC prior is simply the product of the marginal PC priors. This is equivalent to assume a multivariate version of the exponential distribution (constituted by independent components) over the distance. Recall that, in the multivariate case, **?** still consider an univariate exponential distribution to penalise the multi-parameters distance.
Let now (X,Y) be a random vector with generic parameters ξ_1 and ξ_2, then for $D(\xi_1, \xi_2) = D(\xi_1) + D(\xi_2)$, the distance $d(\xi_1, \xi_2) = \sqrt{2D(\xi_1, \xi_2)}$ turns out to be the norm of the vector resulting from the linear combination of the basis vectors, in fact $d(\xi_1, \xi_2) = \sqrt{d(\xi_1, 0)^2 + d(0, \xi_2)^2}$. Notice also that $d(\xi_1, 0) = d(\xi_1)$, i.e. the conditional distance is equal to the marginal one, but not for instance if we consider a correlation structure in the joint density of (X,Y). Let's see an example.
Let us consider the base model

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{I} \right],$$

and the more complex model given by

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \mathbf{I} \right],$$

As we have already seen, the $\text{KLD}(\mu_x, \mu_y) = \frac{\mu_x^2 + \mu_y^2}{2}$. It is also the sum of the KLDs between univariate standard normals, i.e. $\text{KLD}(N(\mu_i, 1)\|N(0,1))$.

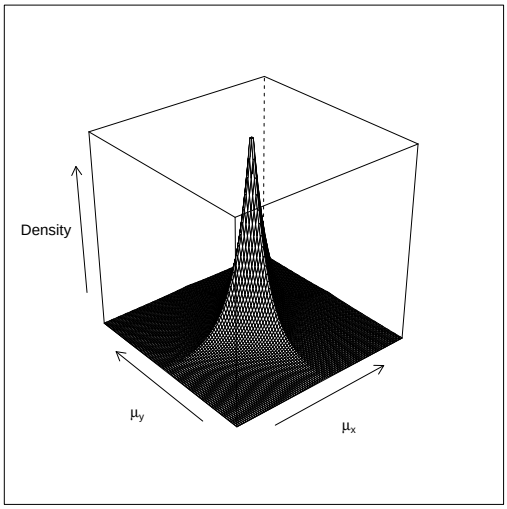

Figure 5.5: Bivariate PC prior for the means of a multivarite Gaussian distribution. Both penalisation rates are equal to one.

This corroborates the assumption of independence among the components of the random vector (μ_x, μ_y).
Therefore, the distance turns out to be orthogonal and as a consequence the joint PC prior over the mean vector is the product of the marginal PC priors, namely the double exponential priors.
In practice, we have $d(\mu_x, 0) = d(\mu_x) = \mu_x$ and $d(0, \mu_y) = d(\mu_y) = \mu_y$, and given $\mu_x \perp \mu_y$

$$\pi(d(\mu_x, \mu_y)) = \pi(d(\mu_x))\pi(d(\mu_y)) = \frac{\lambda_x}{2}\frac{\lambda_y}{2}e^{-\lambda_x\mu_x - \lambda_y\mu_y}, \tag{5.54}$$

since we give half an exponential to the positive part of each component; here μ_i is meant to be positive (given that the distance is positive) and λ_i is a rate parameter.
The corresponding joint PC prior for (μ_x, μ_y) is

$$\pi^{PC}(\mu_x, \mu_y) = \pi^{PC}(\mu_x)\pi^{PC}(\mu_y) = \frac{1}{4}\lambda_x\lambda_y e^{-\lambda_x|\mu_x| - \lambda_y|\mu_y|}, \tag{5.55}$$

where $\mu_x \in \mathbb{R}$, $\mu_y \in \mathbb{R}$. The prior is easily defined and we need only to elicit the rate parameters.
Figure **??** shows the bivariate PC prior over the random vector of the means, obtained assuming independent components. So, when the KLD is additive, the independence assumption is a natural consequence.
Unfortunately, this rarely happens, indeed usaully $\text{KLD}(\Xi_1, \Xi_2) < \text{KLD}(\Xi_1) + \text{KLD}(\Xi_2)$. This is a consequence of the fact that $\text{KLD}(\Xi_1|\Xi_2) \leq \text{KLD}(\Xi_1)$. Notice that the disequality is true only on average. Specifically, $\text{KLD}(\Xi_1|\Xi_2 = \xi_2)$ may be greater than or less than or equal to $\text{KLD}(\Xi_1)$, but on the average $\text{KLD}(\Xi_1|\Xi_2) = \sum_{\xi_2} p(\xi_2)\text{KLD}(\Xi_1|\Xi_2 = \xi_2) \leq \text{KLD}(\Xi_1)$. Here, with abuse of notation, we have denoted the random variables, formerly denoted by ξ_1 and

ξ_2, with their respective capital letters.
Consider for instance the base model given by

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma} \right],$$

where $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
Let's assume that in the more complex model has been added the vector of the means

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \mathbf{\Sigma} \right].$$

For the models above, the KLD turns out to be $\text{KLD}(\mu_x, \mu_y) = \frac{1}{2(1-\rho^2)}(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y)$. So, it is evident now that this KLD cannot equalize the sum of the KLDs over the marginal densities, namely $\text{KLD}(\mu_x) = \frac{\mu_x^2}{2}$ and $\text{KLD}(\mu_y) = \frac{\mu_y^2}{2}$. In this case it seems natural to add to the independence joint PC prior a further component that is able to take into account for dependence. In practice, we multiply the marginal PC prior distributions by a copula density function. Let's assume to use the Gaussian copula, both for practical purposes and its ability to handle high dimensions.
The Joint PC prior looks like

$$\pi^{PC}(\mu_x, \mu_y) = \frac{1}{4}\lambda_x\lambda_y e^{-\lambda_x|\mu_x| - \lambda_y|\mu_y|} \cdot c_\psi(F(\mu_x), G(\mu_y); \psi), \tag{5.56}$$

where F and G are the distribution functions of μ_x and μ_y respectively, ψ is the parameter of the Gaussian copula, while the PC prior density functions are in practice Laplace distributions.
Figure **??** shows the joint PC prior for the random vector of the means, obtained by the product of the marginal PC prior densities, with rate parameters equal to one, times a Gaussian copula density function with positive correlation parameter equal to 0.75, while Figure **??** shows the joint density where we consider a negative correlation parameter of the Gaussian copula equal to -0.75.
The multivariate PC prior obtained above is very easy to simulate from. Suppose that our goal is to estimate probability

$$H(\mu_x, \mu_y) = P(\mu_x \leq \mu_{x_0}, \mu_y \leq \mu_{y_0}), \tag{5.57}$$

where $\mu_x \sim f(\mu_x)$ and $\mu_y \sim g(\mu_y)$ and $H(\mu_x, \mu_y) = C_\psi(F(\mu_x), G(\mu_y); \psi)$ is the Gaussian copula defined in (**??**) with correlation parameter denoted by ψ. Then to simulate from the joint distribution it suffices to implement the following procedure with the first three steps coinciding with the simulation of a bivariate normal distribution, and the particular techniques of the fourth step depend on the specific marginal distributions $F(\mu_x)$ and $G(\mu_y)$. To sample from the Gaussian copula

- Generate independently two standard normal variables $z_1, z_2 \sim N(0, 1)$.
- Define correlated standard normal variables as $w_1 = z_1$ and $w_2 = \psi z_1 + \sqrt{1-\psi^2} z_2$.

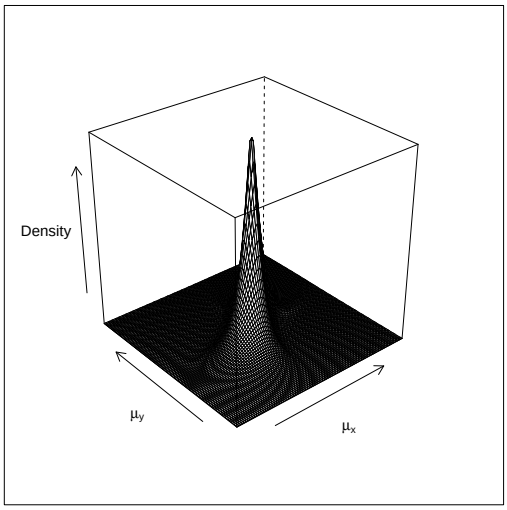

Figure 5.6: Bivariate PC prior for the means of a multivariate Gaussian distribution. The joint distribution is obtained through a Gaussian copula with correlation parameter equal to 0.75; both penalisation rates are equal to one.

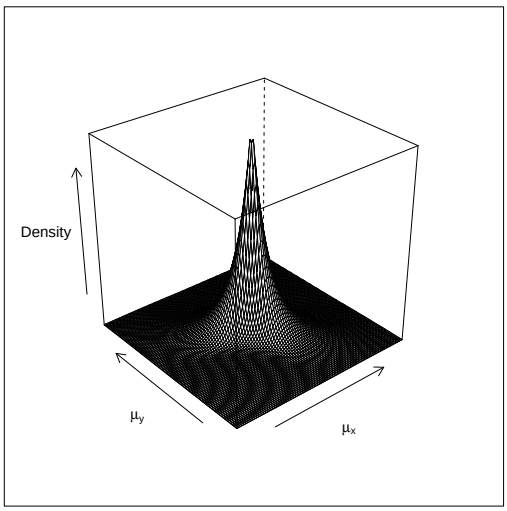

Figure 5.7: Bivariate PC prior for the means of a multivariate Gaussian distribution. The joint distribution is obtained through a Gaussian copula with correlation parameter equal to -0.75; both penalisation rates are equal to one.

- Set $u = \Phi(w_1)$, $v = \Phi(w_2)$.
- Set $\mu_x = F^{-1}(u)$, $\mu_y = G^{-1}(v)$. Exact implementation of this step depends on the distributions $F(\mu_x)$ and $G(\mu_y)$.

Repeat n times to obtain sample $(x_i, y_i), i = 1, \ldots, n \sim C_\psi(F(\mu_x), G(\mu_y); \psi)$. Notice that $H(\mu_x, \mu_y) = C_\psi(F(\mu_x), G(\mu_y); \psi)$, therefore $h(\mu_x, \mu_y) = f(\mu_x)g(\mu_y) \cdot c_\psi(F(\mu_x), G(\mu_y); \psi)$. In our opinion, this is a convenient way to construct multivariate PC prior distributions and, as we have just seen, it is very simple to simulate from them. In addition, the Gaussian copula density function is able to handle high dimensions. Finally, consider that we could take into account for skewness and kurtosis by simply replacing the Gaussian copula with a more flexible copula.
In the situation described above, one could think to use the discrepancy between $\text{KLD}(\mu_x, \mu_y)$ and $\text{KLD}(\mu_x) + \text{KLD}(\mu_y)$ to calibrate the association parameter of the copula that we want to use to construct the multivariate PC prior distribution.
Suppose to calculate the Kullback-Leibler divergence between $\text{KLD}(\mu_x, \mu_y)$ and $\text{KLD}(\mu_x) + \text{KLD}(\mu_y)$, in order to get a quantity that we may consider as the mutual information. So, we calculate

$$\begin{aligned}
&\text{KLD}(\text{KLD}(\mu_x, \mu_y) \| \text{KLD}(\mu_x) + \text{KLD}(\mu_y)) \\
&= \int\int \text{KLD}(\mu_x, \mu_y) \cdot \log \frac{\text{KLD}(\mu_x, \mu_y)}{\text{KLD}(\mu_x) + \text{KLD}(\mu_y)} d\mu_x d\mu_y && (5.58) \\
&= \int\int \frac{1}{2(1-\rho^2)}(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y) \cdot \log \frac{\frac{1}{2(1-\rho^2)}(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y)}{\frac{\mu_x^2+\mu_y^2}{2}} d\mu_x d\mu_y && (5.59) \\
&= \int\int \frac{1}{2(1-\rho^2)}(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y) \cdot \log \frac{(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y)}{(\mu_x^2 + \mu_y^2)(1-\rho^2)} d\mu_x d\mu_y. && (5.60)
\end{aligned}$$

The double integral above represents a sort of evidence in favour of the independence assumption.
On the other hand, the following double integral

$$\begin{aligned}
&\text{KLD}(\text{KLD}(\mu_x) + \text{KLD}(\mu_y) \| \text{KLD}(\mu_x, \mu_y)) \\
&= \int\int \text{KLD}(\mu_x) + \text{KLD}(\mu_y) \cdot \log \frac{\text{KLD}(\mu_x) + \text{KLD}(\mu_y)}{\text{KLD}(\mu_x, \mu_y)} d\mu_x d\mu_y && (5.61) \\
&= \int\int \frac{\mu_x^2 + \mu_y^2}{2} \cdot \log \frac{\frac{\mu_x^2+\mu_y^2}{2}}{\frac{1}{2(1-\rho^2)}(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y)} d\mu_x d\mu_y && (5.62) \\
&= \int\int \frac{\mu_x^2 + \mu_y^2}{2} \cdot \log \frac{(\mu_x^2 + \mu_y^2)(1-\rho^2)}{(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y)} d\mu_x d\mu_y, && (5.63)
\end{aligned}$$

can be viewed as the evidence in favour of the dependence assumption and against the orthogonality hypothesis.
The double integrals above are not symmetric. In essence, this is not a deficiency of the KLD but a feature. The asymmetry can also be expressed in terms of

the sum

$$\begin{aligned}\mathrm{KLD}(\mathrm{KLD}(\mu_x,\mu_y)\|\mathrm{KLD}(\mu_x)+\mathrm{KLD}(\mu_y))\\ -\mathrm{KLD}(\mathrm{KLD}(\mu_x)+\mathrm{KLD}(\mu_y)\|\mathrm{KLD}(\mu_x,\mu_y)),\end{aligned}\tag{5.64}$$

but notice that it looks like the difference, since

$$\mathrm{KLD}(\mathrm{KLD}(\mu_x)+\mathrm{KLD}(\mu_y)\|\mathrm{KLD}(\mu_x,\mu_y))\tag{5.65}$$

is negative (recall that we are computing the KLD between KLDs and not between probability functions); nonetheless the asymmetry is alwalys positive. We define the asymmetric mutual information MI^*

$$\begin{aligned}MI^* = \int\int \mathrm{KLD}(\mu_x)+\mathrm{KLD}(\mu_y)\cdot\log\frac{\mathrm{KLD}(\mu_x)+\mathrm{KLD}(\mu_y)}{\mathrm{KLD}(\mu_x,\mu_y)}d\mu_x d\mu_y+\\ \int\int \mathrm{KLD}(\mu_x,\mu_y)\cdot\log\frac{\mathrm{KLD}(\mu_x,\mu_y)}{\mathrm{KLD}(\mu_x)+\mathrm{KLD}(\mu_y)}d\mu_x d\mu_y.\end{aligned}\tag{5.66}$$

The first addend is negative but it is always less than the second one, which is positive. Therefore $MI^* > 0$.
We need a criterion to infer the association parameter of the copula function and, in our opinion, this should be done for every model at hand, even though here we propose a common strategy.
First of all, since MI^* is always positive, we symmetrise it by letting it depend on the sign of the correlation within $\mathrm{KLD}(\mu_x,\mu_y)$.

$$MI^{**} = MI^*\cdot\mathrm{sgn}(\rho)\tag{5.67}$$

Then, the strategy consists of taking a function g, depending on the model at hand, that is able to map MI^{**} into the parameter of the copula function. Let D be the KLD, in order to simplify the notation.
For instance, for $\rho=-0.1$ we have that

$$\begin{aligned}MI^* &= D(D(\mu_x)+D(\mu_y)\|D(\mu_x,\mu_y))+D(D(\mu_x,\mu_y)\|D(\mu_x)+D(\mu_y))\\ &= -5.68156+11.1733\\ &= 5.49174,\end{aligned}\tag{5.68}$$

then $MI^{**}=-5.49174$.
In this particular case, we define $\psi=g(MI^{**})$ as follows

$$\psi=\frac{-5.49174/1000}{\sqrt{1+(-5.49174/1000)^2}},\tag{5.69}$$

where ψ is the parameter of the Gaussian copula density function and, in turn

$$g(MI^{**})=\frac{MI^{**}/1000}{\sqrt{1+(MI^{**}/1000)^2}}.\tag{5.70}$$

It turns out that $\psi=-0.005491657$.
Let's see what happens for a large and positive value of the correlation.
If $\rho=0.8$, then

$$\begin{aligned}MI^* &= D(D(\mu_x)+D(\mu_y)\|D(\mu_x,\mu_y))+D(D(\mu_x,\mu_y)\|D(\mu_x)+D(\mu_y))\\ &= -601.951+2898.53\\ &= 2296.579,\end{aligned}\tag{5.71}$$

and again $MI^{**} = 2296.579$.
By computing ψ as we have done above, we obtain $\psi = 0.9168528$.
What emerges from the last examples is that the joint PC prior derived above can also encapsulate the information inherent to the dependence among the variables which constitute the joint model (notice that we are not referring to the marginal PC prior distributions).
It is worth to mention that with this choice of g, the joint PC prior supports less dependence than ρ for small values of ρ, whilst it tends to support more dependence than ρ when ρ tends to one. For instance, by choosing $\rho = 0.5$, ψ will be approximately equal to 0.24, and, at a certain point, for increasing ρ, ψ will be greater than ρ.
Finally, suppose to calculate the KLD between the bivariate skew-normal and the bivariate normal distributions, i.e. $\text{KLD}(\alpha_1, \alpha_2)$ where α_1 and α_2 are shape parameters. In this case, $\text{KLD}(\alpha_1, \alpha_2) < \text{KLD}(\alpha_1, 0) + \text{KLD}(0, \alpha_2)$, so we need a criterion similar to the one described above to capture the dependence among the shape parameters. Notice that in this model $\text{KLD}(\alpha_1, 0) \neq \text{KLD}(\alpha_1)$, rather $\text{KLD}(\alpha_1, 0) = \text{KLD}(\bar{\alpha}_1)$, because we need to update $\bar{\alpha}_1 = \frac{\alpha_1 + \omega\alpha_2}{\sqrt{1+(1-\omega^2)\alpha_2^2}}$. Anyhow $\text{KLD}(\alpha_1, \alpha_2) < \text{KLD}(\bar{\alpha}_1) + \text{KLD}(\bar{\alpha}_2)$. In our opinion, it could be used the constraint in equation (**??**) in order to elicit the parameter ψ.
As another example, consider the KLD between the standard skew-t and the standard normal distributions, then the KLD will be a function of the shape parameter λ and the degrees of freedom ν, $\text{KLD}(\lambda, \nu)$. Even in this case $\text{KLD}(\lambda, \nu) < \text{KLD}(\lambda, \infty) + \text{KLD}(0, \nu) = \text{KLD}(\lambda) + \text{KLD}(\nu)$. For this model, the KLD (as a function of both the parameters) calculated at the base model with respect to one of the two parameters equates the KLD computed over the marginal densities, i.e. $\text{KLD}(\lambda)$ is the KLD between the standard skew-normal and normal distributions and it is equal to $\text{KLD}(\lambda, \infty)$, whilst $\text{KLD}(\nu)$ is the KLD between the standard Student-t and the standard normal distributions and it is equal to $\text{KLD}(0, \nu)$.
Finally, in the latter two examples, one could consider to multiply the marginal PC prior distributions by a Gaussian copula density function where a criterion to elicit the association parameter must be specified.

Bibliography

A. Azzalini. A class of distribution which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

A. Azzalini and A. Capitanio. Statistical applications of the mutivariate skew-normal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61:579–602, 1999.

T. Bacigal, V. Jagr, and R. Mesiar. Non-exchangeable random variables, archimax copulas and their fitting to real data. *Kybernetika*, 47(4):519–531, 2011.

C. L. Bayes and M. D. Branco. Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics*, 21:141–163, 2007.

J.O. Berger and J.M. Bernardo. On the development of reference priors. In *Bayesian Statistics*, volume 4, pages 35–60. Oxford University Press, London, 1992.

J.O. Berger and L.R. Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122, 1996.

J.O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *The Annals of Statistics*, 37:905–938, 2009.

J. M. Bernardo. Noninformative priors do not exist. *Journal of Statistical Planning and Inference*, 65:159–189, 1997.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics, Chichester, 1994.

A. Bhattacharya, D. Pati, N.S. Pillai, and D.B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110 (512):1479–1489, 2015.

M. Branco, M. Genton, and B. Liseo. Objective Bayesian analysis oof the skew-t distributions. *Scandinavian Journal of Statistics*, 40(1):63–85, 2013.

G. Casella and E. L. Lehmann. *Theory of Point Estimation*. Springer, 2nd edition, 1998.

A. P. Dawid. Invariant Prior Distributions. In *Encyclopedia of Statistical Sciences*, Wiley. New York, 2006.

M.M. De Queiroz, R.W.C. Silva, and R.H. Loschi. Shannon entropy and kullback–leibler divergence in multivariate log fundamental skew-normal and related distributions. *The Canadian Jounal of Statistics La revue canadienne de statistique*, 44:219–237, 2016.

F. Y. Edgeworth. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society, Serie B*, 71:499–512, 1908.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80: 27–38, 1993.

M. Franco-Villora, M. Ventrucci, and H. Rue. A unified view on Bayesian varying coefficient models. *Electronic Journal of Statistics*, 13(2):5334–5359, 2019.

C. Genest, M. Gendron, and M. Bourdeau-Brien. The advent of copulas in finance. *European Journal of Finance*, 15:609–618, 2009.

E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

W. Hoeffding. Massstabinvariante korrelationstheorie. *Schriften des mathematischen Seminars und des Instituts für Angewandte Mathematik der Universität Berlin*, 5:181–233, 1940.

H. Ishwaran and J.S Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Sciences and Cybernetics*, 4:227–241, 1968.

H. Jeffreys. *Theory of Probability*. Oxford University Press, New York, 1961.

V.E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:143–170, 2010.

W.H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(4):814–861, 1958.

D.V. Lindley. A statistical paradox. *Biometrika*, 44:187–192, 1957.

B. Liseo. La classe delle densità normali sghembe: aspetti inferenziali da un punto di vista bayesiano. *Statistica*, 50(1):71–82, 1990.

B. Liseo and N. Loperfido. A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference*, 136:373–389, 2006.

A.J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press, Princeton, NJ, 2nd edition, 2015.

A.J. Patton. Copula methods for forecasting multivariate time series. In *Handbook of economic forecasting, G. Elliott & A. Timmermann (Eds.)*, pages 899–960. Springer, New York, 2013.

J.M. Pérez and J.O. Berger. Expected-posterior prior distributions for model selection. *Biometrika*, 89:491–511, 2002.

C. P. Robert, N. Chopin, and J. Rousseau. Harold Jeffreys's theory of probability revisited. *Statistical Science*, 24:141–172, 2009.

C.P. Robert. A note on Jeffreys-Lindley paradox. *Statistica Sinica*, 3:601–608, 1993.

C.P. Robert and J. Rousseau. Some comments about James Watson's and Chris Holmes' 'Approximate models and robust decisions': Nonparametric Bayesian clay for robust decision bricks. *https://www.researchgate.net/publication/301878904*, 2016.

J. Rousseau and B. Szabo. Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics*, 45(2):833–865, 2017.

G. Salvadori, C. De Michele, N. T. Kottegoda, and R. Rosso. *Extremes in nature: An approach using copulas*, volume 56 of *Water science and technology library*. Springer, Berlin, 2007.

N. Sartori. Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *Journal of Statistical Planning and Inference*, 136:4259–4275, 2006.

D. Simpson, H. Rue, A. Riebler, T.G. Martins, and S.H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, 2017.

M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

S.H. Sørbie and H. Rue. Fractional gaussian noise: Prior specification and model comparison. *arXiv:1611.06399v1*, 2016.

S.H. Sørbie and H. Rue. Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6):923–935, 2017.

M. Ventrucci, D. Cocchi, G. Burgazzi, and A. Laini. Pc priors for residual correlation parameters in one-factor mixed models. *Statistical Methods and Application*, pages 1–21, 2019.

J. Watson and C. Holmes. Approximate models and robust decisions. *Statistical Science*, 31(4):465–489, 2016.