Annali del Dipartimento di Metodi e Modelli
per l'Economia il Territorio e la Finanza
1/2020, pp. 33-46, DOI: 10.13133/2611-6634/509

E-ISSN 2611-6634
ISSN 2385-0825

SAPIENZA
UNIVERSITÀ EDITRICE

Diego Battagliese*

Extensions of the Univariate PC Prior

Abstract. In the present paper, we describe and explore new methods for constructing joint penalised complexity prior distributions, PC priors hereafter, on the additive model components that build up a more flexible model starting from a base model which would not include those components. Although, an extension to the multivariate case has been already proposed, it is difficult to handle, especially in high-dimensional problems. So, we need something more manageable, particularly for computational purposes. We propose two different constructions of the multivariate PC prior, one based on the conditional PC prior distributions via the Hammersley-Clifford theorem, the other one based on the marginal PC prior distributions by means of a copula approach.

Keywords: Multivariate PC prior, Hammersley-Clifford theorem, Copula modeling, Gaussian copula.

1. Introduction

Penalised complexity priors have been proposed by Simpson, Rue, Riebler, Martins and Sørbye (2017). As pointed out by Simpson et al. (2017), in the univariate case 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 into the prior by selecting a value for the hyperparameter of the PC prior in order to make it as uninformative as possible. The distance between the two models is penalised by assigning to it an exponential distribution whose rate parameter constitutes the shrinking parameter that establishes the informativeness of the PC prior. Here, we are not claiming that 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. For a general review of objective priors see Consonni, Fouskakis, Liseo and Ntzoufras (2018).

Among the main advantages of these priors we can mention the invariance with respect 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 with respect to reparameterisation, but, in general, for certain values of the shrinkage parameter, the PC prior approaches the Jeffreys' one.

Another relevant feature is that, in several models, PC priors are invariant with respect to the other parameters lying both in the base and the complex models, 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 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.

Simpson et al. (2017) 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 not be 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 posterior distributions that give no evidence to the base model, even when it is the true one, and as a consequence, we would not be able to understand whether the evidence for the more flexible

* Memotef, Sapienza University of Rome, Via del Castro Laurenziano, 9 I-00161 Rome, Italy; e-mail: diego.battagliese@uniroma1.it

model come from the data or it is just induced by the prior distribution.
Here, we consider, as a base model, the one such that a particular value of the flexibility parameter ξ makes the component to disappear from 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 by construction and this allows their general use in testing problems.

2. The principled construction of the univariate PC prior

Simpson et al. (2017) defined four basic principles behind the construction of a PC prior for 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 prevents us to choose a different base model coming from our belief about it.
- **Measure of Complexity.** The increased complexity is measured by the Kullback-Leibler divergence

$$\text{KLD}(f\|g) = \int_{\mathcal{X}} f(x;\xi) \log\left(\frac{f(x;\xi)}{g(x)}\right) dx, \tag{1}$$

  where $g(x) = f(x;\ \xi = \xi_0)$, with ξ_0 being a particular value of ξ that simplifies the model. The Kullback-Leibler divergence is a measure of the information lost when g is used to approximate the density f. Therefore, it is not a metric and to obtain 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

$$\frac{\pi_d(d+\nu)}{\pi_d(d)} = r^{\nu}, \qquad d, \nu \geq 0 \tag{2}$$

  where $\pi_d(d) = \theta \exp(-\theta d)$ and $r = \exp(-\theta)$. The penalisation of an additional distance ν does not depend on the initial amount of distance d. Roughly speaking, the constant decay-rate means that we are equally penalising 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. Also in this case, nothing prevents 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 with this property is the exponential distribution, while in the discrete case, the geometric distribution share this property as well.
  We define the PC prior by means of a change of variable

$$\pi(\xi) = \pi_d(d(\xi)) \left| \frac{\partial d(\xi)}{\partial \xi} \right|. \tag{3}$$

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

$$\text{Prob}(Q(\xi) > W) = \alpha. \tag{4}$$

  This is the crucial point of the principled procedure, since some choices of the hyperparameter can make the prior unnecessarily or unintentionally very informative.
  Anyhow, the choice of θ is a user task and this renders the PC prior close to a weakly informative prior. Notice that $Q(\xi)$ is a generic transformation 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 the prior is.

We would like to remark that there is also another principle embedded in the construction above. In particular, we could obtain asymmetric versions of the PC prior by simply assigning different weights to sections of the parameter space where the distance function is monotone. Let us think for instance of a complex model being a standard skew-normal distribution, while the base model is represented by a standard normal. In this case the distance is not a one-to-one function, rather it is symmetric around zero. So, by assigning half an exponential to each part of the parameter space we would end up with a symmetric PC prior, otherwise we would have its asymmetric counterpart.
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 (Sørbye and Rue, 2016).

2.1 Multivariate case

Simpson et al. (2017) also proposed 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 Simpson et al. (2017) 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 allows us to get a local representation for the multivariate PC prior as

$$\pi(\underline{\xi}) = \frac{\lambda}{|S_{d(\underline{\xi})}|} \exp(-\lambda d(\underline{\xi}))|\det(\mathbf{J}(\underline{\xi}))|, \tag{5}$$

where $J_{ij} = \frac{\partial g_i}{\partial \xi_j}$ is the Jacobian of $\underline{g}$. Computational geometry tools can be used to approximately evaluate (5) 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.

$$d(\underline{\xi}) = h(\underline{b}^T \underline{\xi}), \quad \underline{b} > 0, \quad \underline{\xi} \in \mathbb{R}^n_+ \tag{6}$$

or

$$d(\underline{\xi}) = h(\frac{1}{2}\underline{\xi}^T \mathbf{H} \underline{\xi}), \quad \mathbf{H} > 0, \quad \underline{\xi} \in \mathbb{R}^n, \tag{7}$$

for some function $h(\cdot)$ satisfying $h(0) = 0$. 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

$$\pi(\underline{\xi}) = \lambda \exp(-\lambda d(\underline{\xi})) \frac{(n-1)!}{r(\underline{\xi})^{n-1}} |h'(r(\underline{\xi}))|, \quad r(\underline{\xi}) = h^{-1}(d(\underline{\xi})), \tag{8}$$

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

$$\pi(\underline{\xi}) = \lambda \exp(-\lambda d(\underline{\xi})) \frac{\Gamma\left(\frac{n}{2}+1\right)}{n\pi^{\frac{n}{2}} r(\underline{\xi})^{n-2}} \left| h'\left(\frac{1}{2} r(\underline{\xi})^2\right) \right|, \quad r(\underline{\xi}) = \sqrt{2h^{-1}(d(\underline{\xi}))}. \tag{9}$$

Anyhow, the general multivariate case for the PC prior is hard. In our opinion, there are some issues related to (5). 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 (5) is very difficult to compute for high dimensional models, apart from the cases where the distance function

is linear or quadratic, so something easier to use 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.

3. 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 1 (Positivity condition)**.** *A distribution with density $f(x_1, \dots, x_p)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the positivity condition if $f(x_1, \dots, x_p) > 0$ for all $x_1, \dots, x_p$ 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 1 (Hammersley-Clifford)**.** *Let $(X_1, \dots, X_p)$ have joint density $f(x_1, \dots, x_p)$ which satisfies the positivity condition. Then for all choices of $(\varepsilon_1, \dots, \varepsilon_p) \in supp(f)$*

$$f(x_1, \dots, x_p) \propto \prod_{j=1}^{p} \frac{f_{X_j|X_{-j}}(x_j|x_1, \dots, x_{j-1}, \varepsilon_{j+1}, \dots, \varepsilon_p)}{f_{X_j|X_{-j}}(\varepsilon_j|x_1, \dots, x_{j-1}, \varepsilon_{j+1}, \dots, \varepsilon_p)}. \tag{10}$$

For the proof see Besag (1974).
Note that the Hammersley-Clifford theorem does not guarantee the existence of a joint probability distribution for every choice of the conditionals (Johansen and Evers, 2007). In Bayesian modeling such problems mostly arise when using improper prior distributions.
Our proposal is to derive the one-dimensional conditional PC priors and then check whether they are compatible to create a joint PC prior.

3.1 Joint PC Prior for the bivariate Uniform model

As a first example consider a random vector having as a base model a 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

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

Here we confine ourselves to the case $a, b \leq 1$ in order to avoid problems with the positive definiteness of KLD in non-regular models.
Before applying the Hammersley-Clifford theorem we need to compute the full conditional PC priors. Let us compute the conditional KLDs

$$\text{KLD}(a|b) = \int_0^a \int_0^b \frac{1}{ab} \log\left(\frac{1}{ab}\right) dxdy = -\log(ab), \tag{11}$$

where b is not a random variable, rather it is just a parameter with values between 0 and 1. For the sake of clarity, notice that with abuse of notation we denote the conditional KLD in (11) like a conditional distribution, as it will be

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

turned into the prior for a, while b will represent just a parameter. Obviously, the same result holds when computing the conditional KLD of b given the parameter a

$$\text{KLD}(b|a) = \int_0^a \int_0^b \frac{1}{ab} \log\left(\frac{1}{ab}\right) dxdy = -\log(ab). \tag{12}$$

Hence, the distances are the same

$$d(a|b) = \sqrt{-2\log(ab)} \tag{13}$$

$$d(b|a) = \sqrt{-2\log(ab)} \tag{14}$$

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

$$\pi^{PC}(a|b) = \mu e^{-\mu\sqrt{-2\log(ab)}} \left| -\frac{\frac{2}{a}}{2\sqrt{-2\log(ab)}} \right|, \tag{15}$$

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

$$\pi^{PC}(a|b) = \frac{\mu}{a} e^{-\mu\sqrt{-2\log(ab)}} \frac{1}{\sqrt{-2\log(ab)}}. \tag{16}$$

The prior for b given a with rate parameter θ is

$$\pi^{PC}(b|a) = \frac{\theta}{b} e^{-\theta\sqrt{-2\log(ab)}} \frac{1}{\sqrt{-2\log(ab)}}. \tag{17}$$

Suppose to set $(\varepsilon_1, \varepsilon_2) = (a^*, b^*) = \left(\frac{1}{2}, \frac{1}{2}\right)$, then according to the Hammersley-Clifford theorem the joint density can be written as

$$\pi^{PC}(a,b) \propto \frac{\pi^{PC}(a|b^*) \cdot \pi^{PC}(b|a)}{\pi^{PC}(a^*|b^*) \cdot \pi^{PC}(b^*|a)} \tag{18}$$

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

$$\propto \frac{1}{ab} e^{-(\mu-\theta)\sqrt{-2\log\left(\frac{a}{2}\right)}} e^{-\theta\sqrt{-2\log(ab)}} \frac{1}{\sqrt{-2\log(ab)}}. \tag{19}$$

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

$$\pi^{PC}(a,b) \propto \frac{1}{ab} \exp\left(-\lambda\sqrt{-2\log(ab)}\right) \frac{1}{\sqrt{-2\log(ab)}}. \tag{20}$$

Let us study the behaviour of the prior at the boundaries of the parameter space

$$\text{if} \quad a,b \to 1 \begin{cases} \frac{1}{ab} \to 1 \\ \exp\left(-\lambda\sqrt{-2\log(ab)}\right) \to 1 \\ \frac{1}{\sqrt{-2\log(ab)}} \to \infty \end{cases}; \tag{21}$$

on the other hand

$$\text{if} \quad a,b \to 0 \begin{cases} \frac{1}{ab} \to \infty \\ \exp\left(-\lambda\sqrt{-2\log(ab)}\right) \to 0 \\ \frac{1}{\sqrt{-2\log(ab)}} \to 0 \end{cases}. \tag{22}$$

The Hammersley-Clifford construction poses two questions. The first one is related to the positiveness of the resulting joint distribution, while the second one is related to its integrability.
Let us integrate the density in (20)

$$\int_0^1 \int_0^1 \frac{\exp\left(-\lambda\sqrt{-2\log(ab)}\right)}{ab\sqrt{-2\log(ab)}} dadb = \frac{1}{\lambda^3}, \tag{23}$$

so, the proper joint PC prior should be written as

$$\pi^{PC}(a,b) = \frac{\lambda^3}{ab}\exp\left(-\lambda\sqrt{-2\log(ab)}\right)\frac{1}{\sqrt{-2\log(ab)}}. \tag{24}$$

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

3.2 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

$$f_0(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}\left(x^2+y^2\right)}, \tag{25}$$

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

$$f_1(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]}. \tag{26}$$

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

$$\mathrm{KLD}(f_1\|f_0) = \iint_{\mathbb{R}^2}\frac{1}{2\pi}e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]}\log\left(e^{-\frac{1}{2}\left[(x-\mu_x)^2-x^2+(y-\mu_y)^2-y^2\right]}\right)dxdy, \tag{27}$$

or equivalently

$$\mathrm{KLD}(\mu_x,\mu_y) = \iint_{\mathbb{R}^2}\frac{1}{2\pi}e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]}\left(-\frac{1}{2}\left[\mu_x^2-2x\mu_x+\mu_y^2-2y\mu_y\right]\right)dxdy. \tag{28}$$

Notice that after integrating out x and y, the KLD above is just a function of the remaining parameters. In practice, the arguments of the KLDs in (27) and (28) give rise to the same meaning. Now, we can write

$$\begin{aligned}\mathrm{KLD}(f_1\|f_0) &= -\frac{1}{2}\mu_x^2 - \frac{1}{2}\mu_y^2 + \mu_x\mathbb{E}^{\sim f_1}[X] + \mu_y\mathbb{E}^{\sim f_1}[Y]\\ &= -\frac{1}{2}\mu_x^2 - \frac{1}{2}\mu_y^2 + \mu_x^2 + \mu_y^2\\ &= \frac{\mu_x^2+\mu_y^2}{2}.\end{aligned} \tag{29}$$

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

$$d(\mu_x,\mu_y) = \sqrt{\mu_x^2+\mu_y^2}. \tag{30}$$

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

$$\begin{aligned}\pi^{PC}(\mu_x|\mu_y) &= \frac{\lambda_x}{2}\exp\left(-\lambda_x\sqrt{\mu_x^2+\mu_y^2}\right)\frac{1}{2\sqrt{\mu_x^2+\mu_y^2}}2|\mu_x|\\ &= \frac{\lambda_x}{2}\exp\left(-\lambda_x\sqrt{\mu_x^2+\mu_y^2}\right)\frac{|\mu_x|}{\sqrt{\mu_x^2+\mu_y^2}}.\end{aligned} \tag{31}$$

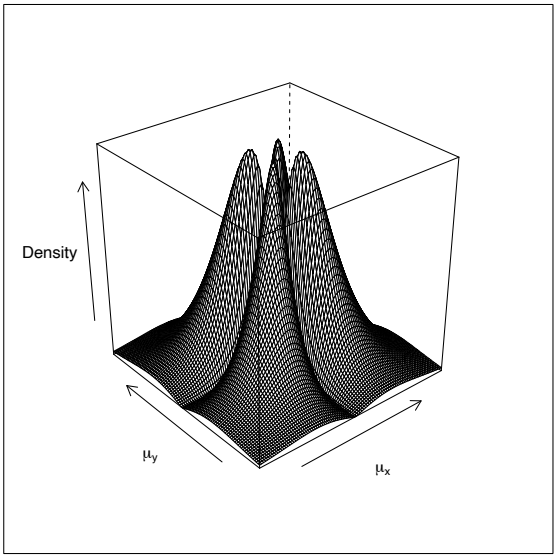

Figure 1. Joint PC prior for the mean vector of a bivariate Gaussian density, where $\lambda = 0.1$.

Equivalently, the PC prior for μ_y given μ_x is

$$\pi^{PC}(\mu_y|\mu_x) = \frac{\lambda_y}{2}\exp\left(-\lambda_y\sqrt{\mu_x^2+\mu_y^2}\right)\frac{|\mu_y|}{\sqrt{\mu_x^2+\mu_y^2}}. \tag{32}$$

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

$$\begin{aligned}\pi^{PC}(\mu_x,\mu_y) &\propto \frac{\pi^{PC}(\mu_x|\tilde{\mu}_y)\cdot\pi^{PC}(\mu_y|\mu_x)}{\pi^{PC}(\tilde{\mu}_x|\tilde{\mu}_y)\cdot\pi^{PC}(\tilde{\mu}_y|\mu_x)}\\ &= \frac{\frac{\lambda_x}{2}e^{-\lambda_x\sqrt{\mu_x^2+\tilde{\mu}_y^2}}\frac{|\mu_x|}{\sqrt{\mu_x^2+\tilde{\mu}_y^2}}\cdot\frac{\lambda_y}{2}e^{-\lambda_y\sqrt{\mu_x^2+\mu_y^2}}\frac{|\mu_y|}{\sqrt{\mu_x^2+\mu_y^2}}}{\frac{\lambda_x}{2}e^{-\lambda_x\sqrt{\tilde{\mu}_x^2+\tilde{\mu}_y^2}}\frac{|\tilde{\mu}_x|}{\sqrt{\tilde{\mu}_x^2+\tilde{\mu}_y^2}}\cdot\frac{\lambda_y}{2}e^{-\lambda_y\sqrt{\mu_x^2+\tilde{\mu}_y^2}}\frac{|\tilde{\mu}_y|}{\sqrt{\mu_x^2+\tilde{\mu}_y^2}}}\\ &\propto \frac{|\mu_x\mu_y|}{\sqrt{\mu_x^2+\mu_y^2}}e^{-\lambda_x\sqrt{\mu_x^2+\tilde{\mu}_y^2}+\lambda_y\sqrt{\mu_x^2+\tilde{\mu}_y^2}-\lambda_y\sqrt{\mu_x^2+\mu_y^2}}.\end{aligned} \tag{33}$$

Now, let us assume, for simplicity, $\lambda_x = \lambda_y = \lambda$, then

$$\pi^{PC}(\mu_x,\mu_y) \propto \frac{|\mu_x\mu_y|}{\sqrt{\mu_x^2+\mu_y^2}}e^{-\lambda\sqrt{\mu_x^2+\mu_y^2}}. \tag{34}$$

Figure 1 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 (Johnson and Rossell, 2010).

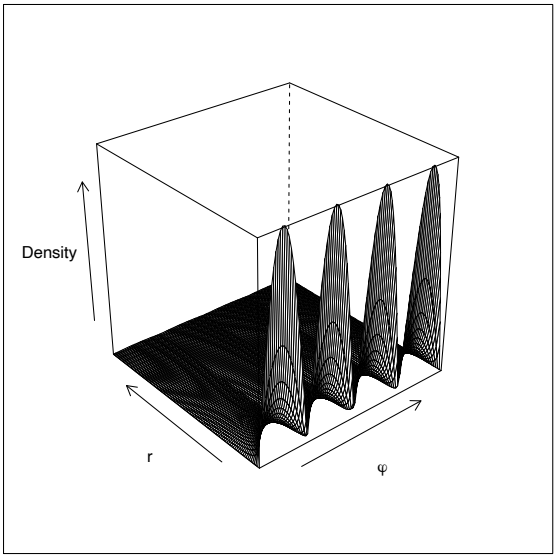

Figure 2. Joint PC prior in terms of the polar coordinates, φ and r, for the mean vector of a bivariate normal density, where $\lambda = 0.1$.

Alternatively, we can express the parameterization in terms of the polar coordinates

$$\begin{cases} \mu_x = r\cos\varphi \\ \mu_y = r\sin\varphi \\ \mu_x^2 + \mu_y^2 = r^2 \end{cases} \tag{35}$$

so that our prior can be written as

$$\begin{aligned} \pi^{PC}(r,\varphi) &\propto \frac{r^2|\sin\varphi\cos\varphi|}{r}\exp(-\lambda r) \\ &= r\exp(-\lambda r)|\sin\varphi\cos\varphi|, \end{aligned} \tag{36}$$

where $r\exp(-\lambda r) \propto \text{Gamma}(r|\nu = 2, \lambda)$. Figure 2 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 $\lambda = 0.1$.

3.3 PC Prior in the bivariate Skew-Normal model

As a third example, consider the bivariate skew-normal model. The scalar skew-normal density can be extended to the d-dimensional case by considering the following density function (Azzalini and Capitanio, 1999)

$$f_d(\underline{x};\Omega,\underline{\alpha}) = 2\phi_d(\underline{x};\Omega)\Phi(\underline{\alpha}^{\mathrm{T}}\underline{x}), \qquad \underline{x}\in\mathbb{R}^d, \tag{37}$$

where Ω is a positive-definite $d \times d$ correlation matrix, $\phi_d(\underline{x};\Sigma)$ is the density function of a $\text{N}_d(0,\Sigma)$ variate and $\underline{\alpha}$ is a d-dimensional vector parameter.
To make the multivariate skew-normal more concrete we have a closer look at the bivariate skew-normal distribution. From equation (37), we define the bivariate skew-normal as

$$f(x_1,x_2;\alpha_1,\alpha_2,\omega) = 2\phi_2(x_1,x_2;\omega)\Phi(\alpha_1 x_1 + \alpha_2 x_2), \tag{38}$$

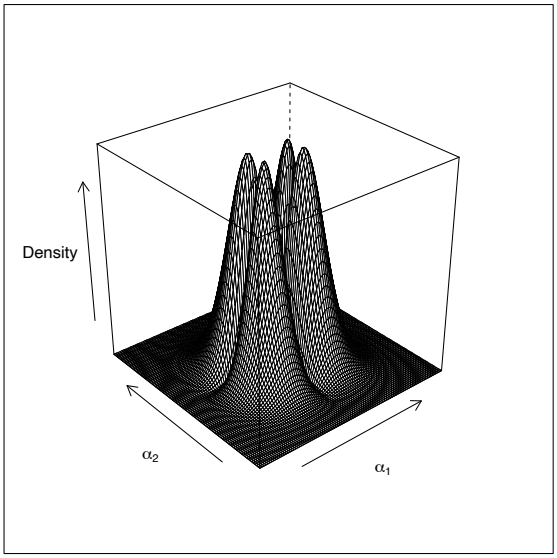

Figure 3. Joint PC prior for the vector $\underline{\alpha}$ of the bivariate skew-normal model, where the rate parameter $\lambda = 1$.

where ω is the off-diagonal element of Ω.
Figure 3 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 must be numerically computed as the KLD has no closed form and it is given by

$$\text{KLD}(\alpha_1, \alpha_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2\phi_2(\mathbf{x})\Phi(\alpha_1 x_1 + \alpha_2 x_2) \log(2\Phi(\alpha_1 x_1 + \alpha_2 x_2))dx_1 dx_2. \tag{39}$$

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 (39) 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

$$\frac{\partial \text{KLD}(\alpha_1, \alpha_2)}{\partial \alpha_1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2\phi_2(\mathbf{x})\phi(\alpha_1 x_1 + \alpha_2 x_2)x_1 \cdot (1 + \log(2\Phi(\alpha_1 x_1 + \alpha_2 x_2)))dx_1 dx_2, \tag{40}$$

and

$$\frac{\partial \text{KLD}(\alpha_1, \alpha_2)}{\partial \alpha_2} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2\phi_2(\mathbf{x})\phi(\alpha_1 x_1 + \alpha_2 x_2)x_2 \cdot (1 + \log(2\Phi(\alpha_1 x_1 + \alpha_2 x_2)))dx_1 dx_2. \tag{41}$$

Once again the integrals must be numerically computed.

4. Copula based approach to construct the multivariate PC prior

In the previous section, we have explored the construction of the multivariate PC prior via the Hammersley-Clifford theorem. Apart from the Uniform model, the resulting multivariate PC priors exhibit 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 situations, but in our case it is not adequate. In fact, our goal is to construct a prior which penalizes

distance from the base model and which maintains some mass at the base model itself in order to avoid overfitting (Simpson et al., 2017).
In this section, we explore the construction of the multivariate PC prior 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. This latter is obtained by considering as the base model the one where the two parameters themselves are absent.
Let us define now the KLD as a function of two generic parameters ξ_1, ξ_2

$$\text{KLD}(\xi_1, \xi_2) = \text{KLD}(\xi_1) + \text{KLD}(\xi_1|\xi_2), \tag{42}$$

where $\text{KLD}(\xi_1|\xi_2)$ is the conditional relative entropy. Equation (42) is validated by the following theorem.

Theorem 2 (Chain Rule for relative entropy)**.**

$$\text{KLD}(p(x,y)\|q(x,y)) = \text{KLD}(p(x)\|q(x)) + \text{KLD}(p(y|x)\|q(y|x)). \tag{43}$$

Proof. Let us assume, for simplicity, that $p(\cdot,\cdot)$ and $q(\cdot,\cdot)$ are bivariate discrete distributions

$$\begin{aligned}
&\text{KLD}(p(x,y)\|q(x,y)) \\
&\quad= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} && (44)\\
&\quad= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} && (45)\\
&\quad= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} && (46)\\
&\quad= \text{KLD}(p(x)\|q(x)) + \text{KLD}(p(y|x)\|q(y|x)). && (47)
\end{aligned}$$

□

For the sake of clarity, notice that for joint probability mass functions $p(x,y)$ and $q(x,y)$, the conditional relative entropy $\text{KLD}(p(y|x)\|q(y|x))$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$ (Cover and Thomas, 2006). Another interesting property is that the conditional entropy $H(\xi_1|\xi_2) \leq H(\xi_1)$ (Conditioning reduces entropy), therefore $H(\xi_1, \xi_2) \leq H(\xi_1) + H(\xi_2)$, with the equality holding when ξ_1 and ξ_2 are independent. On the other hand, conditioning increases divergence. Notice that equation (42) can be also used to derive conditional PC priors based on the conditional relative entropy, as the 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 $\text{KLD}(\xi_1, \xi_2) = \text{KLD}(\xi_1) + \text{KLD}(\xi_2)$, the joint prior for the distance scales can be considered with independent components. 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. In the latter case, 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, i.e. the product of independent exponential densities, over the distance. Recall that, in the multivariate case, Simpson et al. (2017) still consider a 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 $\text{KLD}(\xi_1, \xi_2) = \text{KLD}(\xi_1) + \text{KLD}(\xi_2)$, the distance $d(\xi_1, \xi_2) = \sqrt{2\text{KLD}(\xi_1, \xi_2)}$ 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).
As an example, let us consider the base model of section 3.2

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

and the more complex model given by

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

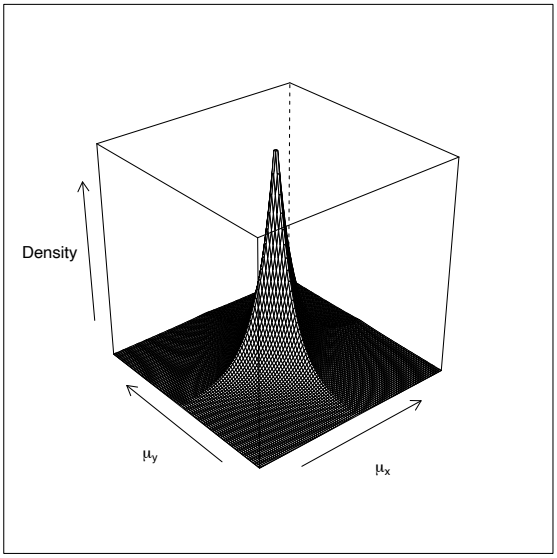

Figure 4. Bivariate PC prior for the means of a bivariate Gaussian distribution. Both penalisation rates are equal to one.

As we have already seen, the 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. KLD$(N(\mu_i, 1)\|N(0, 1))$. 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 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 $\mu_x \perp \mu_y$

$$\pi(d(\mu_x, \mu_y)) = \pi(d(\mu_x))\pi(d(\mu_y)) = \frac{\lambda_x}{2}\frac{\lambda_y}{2}e^{-\lambda_x\mu_x-\lambda_y\mu_y}, \tag{48}$$

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

$$\pi^{PC}(\mu_x, \mu_y) = \pi^{PC}(\mu_x)\pi^{PC}(\mu_y) = \frac{1}{4}\lambda_x\lambda_y e^{-\lambda_x|\mu_x|-\lambda_y|\mu_y|}, \tag{49}$$

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 4 shows the bivariate PC prior over the random vector of the means. The prior is obtained assuming independent components. So, when the KLD is additive, the independence assumption is a natural consequence.
Unfortunately, this rarely happens, indeed usually $H(\Xi_1, \Xi_2) < H(\Xi_1) + H(\Xi_2)$. This is a consequence of the fact that $H(\Xi_1|\Xi_2) \leq H(\Xi_1)$. Notice that the inequality is true only on average. Specifically, $H(\Xi_1|\Xi_2 = \xi_2)$ may be greater than or less than or equal to $H(\Xi_1)$, but on average $H(\Xi_1|\Xi_2) = \sum_{\xi_2} p(\xi_2)H(\Xi_1|\Xi_2 = \xi_2) \leq H(\Xi_1)$. Here, with abuse of notation, we have denoted the random variables, formerly denoted by ξ_1 and ξ_2, with their respective capital letters.
Now, 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 $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ is fixed and known.
Let us assume that in the more complex model the vector of the means has been added

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

For the models above, the KLD turns out to be $\text{KLD}(\mu_x, \mu_y) = \frac{1}{2(1-\rho^2)}(\mu_x^2 + \mu_y^2 - 2\rho\mu_x\mu_y)$. From the KLD expression we may notice how a ρ that is positive and close to one can mitigate the penalisation when $\mu_x = \mu_y$. Anyhow, this is not true in general for $\mu_x \neq \mu_y$. It is certainly true that negative values of ρ increase the KLD more than their positive counterparts and as a consequence boost the penalisation.
Moreover, it is evident that the KLD does not 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, according to the Sklar's theorem (Sklar, 1959), we multiply the marginal PC prior distributions by a copula density function. Let us assume to use the Gaussian copula, both for practical purposes and its ability to handle high dimensions. Notice that the example above is just an illustration of how the additive decomposition of the KLD can be broken.
Then, the joint PC prior looks like

$$\pi^{PC}(\mu_x, \mu_y) = \frac{1}{4}\lambda_x\lambda_y e^{-\lambda_x|\mu_x| - \lambda_y|\mu_y|} \cdot c_\psi(F(\mu_x), G(\mu_y); \psi), \tag{50}$$

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. However, we are aware that for this example the correlation parameter ψ of the copula function should depend on the correlation ρ of the data, otherwise we would have a joint PC prior where the copula component does not take into account for the correlation ρ. Equation (50) is meant to be just an illustration. Anyhow, the construction above need a further refinement in order to allow the parameter ψ to depend on ρ, therefore our aim for future developments is to derive a particular strategy in order to write $\psi(\rho)$. Furthermore, we would like to remark that the joint PC prior in (50) does not accomplish the obvious requirements of dependence that come from the KLD expression above, i.e. different penalisations according to the sign of ρ. For the sake of simplicity, we could even consider $\psi = \rho$, but in this case we would not consider the different decay of the prior depending on the the sign of ρ.
Figure 5 shows the joint PC prior for the random vector of the means. It is 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 6 shows the joint density where we consider a negative correlation parameter of the Gaussian copula equal to -0.75.

5. Conclusions

In this paper, we have described two different constructions of the multivariate PC prior, the former based on the Hammersley-Clifford theorem and the latter based on a copula representation.
Regarding the latter approach, the copula construction allows us to drop the assumption of penalising the multivariate distance by means of a one-dimensional exponential distribution. In addition, the resulting multivariate PC prior avoids the overfitting due to the prior mass absence at the base model. These two features are the main advantages of such an approach. Among the positive aspects of the aforementioned construction there is also the possibility to easily handle high-dimensional problems; it depends on the particular copula at hand, but for instance the Gaussian copula can be used for large dimensions. In fact, the Gaussian copula is very easy to simulate from, and this renders the copula approach more suitable from a computational point of view.
Furthermore, we can retrieve the orthogonality assumption by simply selecting the value of the copula parameter that makes the two components independent; for instance, in the Gaussian copula, this occurs for $\psi = 0$.
Copulas are really flexible tools that allow us to build a joint distribution starting from the marginals and including a function that accounts for dependence. Notice that the copula function can also take into account for skewness, kurtosis, left or right-tail dependence, among the others.
A more interesting extension would be to define a particular criterion in order to elicit the copula parameter. In our

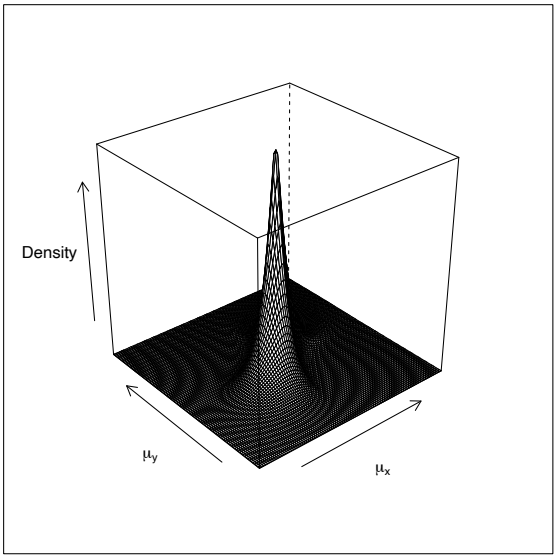

Figure 5. Bivariate PC prior for the means of a bivariate Gaussian distribution. The joint distribution is obtained through a Gaussian copula with correlation parameter equal to 0.75*; both penalisation rates are equal to one.*

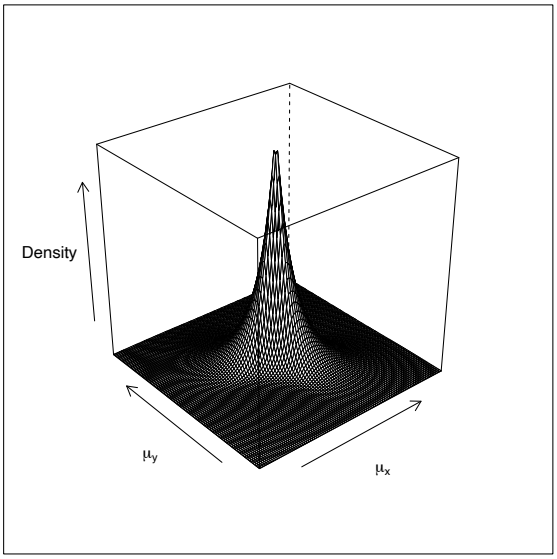

Figure 6. Bivariate PC prior for the means of a bivariate Gaussian distribution. The joint distribution is obtained through a Gaussian copula with correlation parameter equal to -0.75*; both penalisation rates are equal to one.*

opinion, this should be done for each model at hand by exploiting the information about the parameters of interest and by translating it into the association parameter of the copula. For instance, let us consider situations where one has some constraints over the dependence between the parameters of interest, and constraints are determined by the parameters of interest themselves. In this case, it would be natural to use a copula to connect the marginal distributions. Finally, unlike the Hammersley-Clifford construction, the copula based approach requires the knowledge of the marginal densities instead of the conditional ones.

References

AZZALINI A., CAPITANIO A. (1999), Statistical applications of the multivariate skew-normal distributions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 579-602.

BESAG J. (1974), Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2) 192-236.

CONSONNI G., FOUSKAKIS D., LISEO B., NTZOUFRAS I. (2018), Prior distributions for objective Bayesian analysis, *Bayesian Analysis*, 13(2) 627-679.

COVER T. M., THOMAS J. A. (2006), *Elements of information theory* (second edition). Wiley, New Jersey.

JOHANSEN A. M., EVERS L. (2007), *Monte Carlo Methods*, Lecture Notes. University of Bristol, Department of Mathematics.

JOHNSON V. E., ROSSELL D. (2010), 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.

SIMPSON D., RUE H., RIEBLER A., MARTINS T. G., SØRBYE S. H. (2017), Penalising model component complexity: a principled, practical approach to constructing priors (with Discussion), *Statistical Science*, 32(1) 1-28.

SKLAR M. (1959), Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, 8, 229-231.

SØRBYE S. H., RUE H. (2018), Fractional Gaussian noise: prior specification and model comparison, *Environmetrics*, 29(5-6) e2457. https://doi.org/10.1002/env.2457.