Penalised Complexity priors for copula estimation

Diego Battagliese[1], Clara Grazian[2], Brunero Liseo[1], Cristiano Villa[3]

[1] MEMOTEF Department, Sapienza University of Rome, Rome, Italy
[2] University of New South Wales, Sydney, Australia
[3] SMSAS, University of Kent, Canterbury, United Kingdom

E-mail for correspondence: `diego.battagliese@uniroma1.it`

Abstract: We consider a multivariate model with independent marginals as a benchmark for a generic multivariate model where the marginals are not independent. The Penalised Complexity (PC) prior takes natural place in such a context, as we can include in the simpler model an extra-component taking into account for dependence. In this paper, the additional component is represented by the parameter of the Gaussian copula density function. We show that the PC prior for a generic copula parameter can be derived regardless of the parameters of the marginal densities. Then, we propose a hierarchical PC prior for the Gaussian copula model.

Keywords: PC prior; Gaussian copula; Intrinsic prior; Hierarchical PC prior.

1 Introduction

In many statistical models it is natural to have a nested structure. Consider a model of a given complexity, one way to obtain a richer and more flexible model is to include an extra-component so that the simpler model would be nested in the more complex one. We may think, for instance, of a situation where we want to model the joint distribution of several random variables through a copula function. In the case of dependence among variables, the joint density can be expressed as the product of the marginal distributions times a copula function, on the contrary, the joint density boils down to the only product of the marginals when the latter are independent. We derive the PC prior for the correlation parameter of the Gaussian copula by exploiting the following result on the Kullback-Leibler divergence (KLD).

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Notice that the KLD is used to measure the distance between the two models. For a review of the principles behind the construction of a Penalised Complexity prior, see Simpson *et al.* (2017).

2 Method

We consider as a base model a certain multivariate density where the marginal densities are independent. Then, we could render this model more flexible by allowing a copula function to account for dependence, on the basis of the Sklar's representation. The flexible model is

$$M_1 = \{f_{\mathbf{X};\phi}(\mathbf{x};\phi), \mathbf{x} \in \mathbb{R}^k, \phi \in \mathbb{R}^q\}, \tag{1}$$

where, according to the Sklar's theorem, the joint density can be written

$$f_{\mathbf{X};\phi}(\mathbf{x};\phi) = \prod_{j=1}^{k} f_j(x_j;\underline{\theta}_j) c_\psi(F_1(x_1;\underline{\theta}_1), \ldots, F_k(x_k;\underline{\theta}_k);\psi), \tag{2}$$

and $\phi = \{\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_k, \psi\}$.
Furthermore, let

$$M_0 : \qquad \mathbf{X} \sim f_0(\mathbf{x};\varphi) = f_{\mathbf{X};\varphi}(\mathbf{x};\varphi) = \prod_{j=1}^{k} f_j(x_j;\underline{\theta}_j) \tag{3}$$

be the base model, where f_0 is 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. Here, $\varphi = \{\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_k, \psi = \psi_0\}$. Then, the theorem below follows

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

$$\mathrm{KLD}(f_{\mathbf{X}} \| f_{\mathbf{Y}}) = \int_{[0,1]^k} c(u_1, \ldots, u_k;\psi) \log c(u_1, \ldots, u_k;\psi) du_1 \ldots du_k, \tag{4}$$

where $c(u_1, \ldots, u_k)$ *represents the copula function associated with the density of* $\mathbf{X}$ *and* $U_j \sim \mathrm{Unif}(0,1),\ j = 1, \ldots, k.$

The theorem above states that the distance between a generic multivariate density and the one with independent marginals can be expressed as the distance between the copula density function and the independence copula. This result allows us to derive the PC prior for the copula parameter regardless of the parameters of the marginals. Notice also that it applies to any copula function and any dimension. Nevertheless, apart from the case of equicorrelation, for multidimensional elliptical copulas we need to define a multivariate PC prior. Suppose now to have only two marginal distributions, on the basis of Theorem 1 we can write

$$\mathrm{KLD}(f_{\mathbf{X};\phi}\|f_{\mathbf{X};\varphi}) = \int_{\mathcal{U}}\int_{\mathcal{V}} c(u,v;\rho)\log c(u,v;\rho)dudv, \tag{5}$$

where c is the density function of a bivariate Gaussian copula with parameter ρ. Then, $\mathrm{KLD}(\rho) = -\frac{1}{2}\log(1-\rho^2)$, and the prior is easily obtained

$$\pi^{PC}(\rho) = \frac{\theta}{2}\exp\left(-\theta\sqrt{-\log(1-\rho^2)}\right)\frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}. \tag{6}$$

The latter prior is proper, clearly symmetric as it depends on ρ only through the square and the absolute value, and has any odd moment equal to zero. Simpson *et al.* (2017) proposed to use a probability statement on a tail event to select the parameter θ. This latter plays a key role as it regulates the shrinkage of the prior towards the base model, so a wrong choice of this parameter may be misleading, especially in Bayesian hypothesis testing.
From an objective point of view, we calculate the intrinsic prior for the rate parameter θ and then we specify the hyperparameter of such an intrinsic prior distribution by maximizing the variance of the hierarchical PC prior for ρ where the intrinsic prior is put on θ. The procedure to derive the intrinsic prior is borrowed from Pérez and Berger (2002) as it coincides with the expected-posterior prior.
We use $\pi^N(\theta) = \frac{1}{\theta}$ as an improper starting distribution, then the intrinsic prior is given by

$$\pi^I(\theta) = \int_{-1}^{1}\pi(\theta|\rho_\ell)f(\rho_\ell|H_0)d\rho_\ell, \tag{7}$$

where $f(\rho_\ell|\theta_0)$ is the PC prior in (6) calculated in θ_0, say the null hypothesis, and $\pi(\theta|\rho_\ell) = \frac{\pi^N(\theta)f(\rho_\ell|\theta)}{m^N(\rho_\ell)}$, where in turn ρ_ℓ represents the training sample. If there is no subset of ρ_ℓ for which $0 < m^N(\rho_\ell) < \infty$, then ρ_ℓ is called *minimal training sample*. Berger and Pericchi (1996) showed that often it will simply be a sample of size $\max(\dim(\theta))$. So, we need just an observation to convert the improper starting distribution into a proper prior. Therefore

$$\pi^I(\theta) = \frac{\theta_0}{(\theta+\theta_0)^2} \tag{8}$$

will be proper. We set the hyperparameter θ_0 in an objective manner. In particular, we numerically maximize the variance with respect to θ_0, i.e.

$$\max_{\theta_0}\int_{-1}^{1}\int_{0}^{\infty}\rho^2\pi^{PC}(\rho|\theta)\pi^I(\theta|\theta_0)d\theta d\rho. \tag{9}$$

The maximizer is $\theta_0 = 0.491525$ as it renders the prior as flat as possible.

3 Simulation study and real data

We check out the frequentist performance of our hierarchical PC prior via a simulation study. For each true 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. We use the Jeffreys' prior as competitor for inference, while for Bayesian hypothesis test we use the Arc-sine prior, since it is proper.
As one can expect, for $\rho = 0$, our hierarchical PC prior is superior to the Jeffreys' prior in terms of MSE; this is because of the little spike at the base model induced by the hierarchical approach. However, for intermediate correlations, there seems to be a bias-variance trade-off; the Jeffreys' prior looks less biased but less efficient, whilst the PC prior seems to be more biased but more efficient. To compare overall values of ρ^*, we also compute an overall MSE, and the latter is basically in favour of our prior.
We use Bayes factor to select among models. Theorem 1 allows us to write

$$B_{01} = \frac{c_\rho(u, v; \rho)|_{\rho=0}}{\int c_\rho(u, v; \rho)\pi^{PC}(\rho|\theta_0 = 0.491525)d\rho}. \tag{10}$$

We compute the frequency of times that $B_{01} \leq 0.5$. It turns out to be basically smaller for the PC prior compared to the Arc-sine prior when the true model is the base model, whilst it is larger when the true ρ deviates far away from the independence model, especially for small sample sizes.
Finally, we analyze the `danube` data set which contains ranks of base flow observations for two stations situated at Scharding (Austria) on the Inn river and at Nagymaros (Hungary) on the Danube. The data have been pre-processed to remove any time trend. Specifically, a linear time series model with 12 seasonal components is fitted. Then residuals are extracted. The correlation between time series is computed over the residuals, otherwise we would carry back correlation within the series. The results of the Bayesian test are in line with the ones of the frequentist test, providing strong evidence for $\rho \neq 0$.

References

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

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

Simpson, D., Rue, H., Riebler, A., Martins, T.G. and Sørbye, S.H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with Discussion). *Statistical Science*, **32**, 1 – 28.