Penalising the complexity of extensions of the Gaussian distribution

Penalizzazione della complessità relativa alle estensioni della distribuzione normale

Diego Battagliese and Brunero Liseo

Abstract The Gaussian distribution has ever been the most popular and usable device in the field of statistics. Even in the context of penalised complexity (PC) priors, the normal density has a particular meaning, especially because we can consider it as a base model which could be extended both in terms of tail thickness and skewness. We derive the numerical PC prior for the shape parameter of the skew-normal density and the analytical PC prior for the degrees of freedom of the t-distribution. We also perform an approximation of the Kullback-Leibler divergence (KLD) in the the skew-normal model.

Abstract *La distribuzione normale ha sempre ricoperto un ruolo fondamentale in statistica. Anche nel caso delle PC prior essa riveste un ruolo importante, giacché può essere estesa sia per via di una componente di curtosi sia per una di asimmetria. Qui deriviamo la PC prior numerica per il parametro di forma di una normale asimmetrica e l'espressione analitica della PC prior per i gradi di libertà di una t di Student. Inoltre, proponiamo un'approssimazione della KLD quando l'estensione è in termini di asimmetria.*

Key words: PC priors, Skew-normal distribution, Student t-distribution, Kullback-Leibler divergence.

Diego Battagliese
Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161, Rome, e-mail: diego.battagliese@uniroma1.it

Brunero Liseo
Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161, Rome, e-mail: brunero.liseo@uniroma1.it

1 Introduction

In many practical statistical works, datasets reveal departures from symmetry, hence something more flexible than the normal model is needed. The skew-normal distribution [1] 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

$$f(x;\mu,\sigma,\lambda) = \frac{2}{\sigma}\,\phi\Big(\frac{x-\mu}{\sigma}\Big)\,\Phi\Big(\lambda\frac{x-\mu}{\sigma}\Big), \quad x\in\mathbb{R}, \quad \lambda\in(-\infty,+\infty), \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard Gaussian pdf and CDF respectively . Also the t-distribution extends the Gaussian in terms of robustness. Penalised complexity priors have been proposed in [4] and are based on the KLD between the simpler and the complex models. For a review of the principles behind the construction of a Penalised Complexity prior, see [4].

2 PC prior for the shape parameter in the skew-normal model

We can look at the skew-normal model as a flexible version of the normal distribution, where the latter represents the base model. In fact, for a particular value of λ, i.e. $\lambda = 0$, the density in (1) boils down to the normal density as $\Phi(0) = 1/2$. An important feature of the PC prior for λ is the invariance with respect to the location and scale parameters.

Proposition 1 (Invariance wrt location-scale) *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.*

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

can be written as

$$\int_{\mathscr{T}} 2\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}t^2\Big)\Phi\Big(\lambda t\Big)\log\Big\{2\Phi\Big(\lambda t\Big)\Big\}dt, \tag{3}$$

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

In other words, the resulting PC prior for λ does not depend on μ and σ. Suppose $X \sim SN(0,1,\lambda)$, the distance in terms of δ is

$$d(\delta) = \sqrt{2\mathrm{KLD}(\delta)} = \sqrt{2\int_{\mathscr{X}} 2\;\phi(x)\;\Phi\{\lambda(\delta)x\}\log[2\;\Phi\{\lambda(\delta)x\}]\;dx}, \tag{4}$$

where $\delta = \delta(\lambda) = \frac{\lambda}{\sqrt{1+\lambda^2}}$, $\delta \in (-1,1)$. The distance function in (4) is symmetric around 0, as well as the KLD. The minimum is at 0, where $d(0) = 0$, while the maximum is attained at the boundary values. The distance is exponentially distributed. We must be careful in making the change of variable to get the prior for δ, because we have to handle each monotone curve separately. In particular, the function $d(\delta)$ is monotone on $(-1,0)$ and on $(0,1)$. Then, the pdf for δ is

$$\pi(\delta) = \begin{cases} \sum_{i=1}^{2} \pi\{d_i(\delta)\} \left| \frac{\partial d_i(\delta)}{\partial \delta} \right| & \text{if } d(\delta) \in \Theta \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\Theta = (0,\infty)$ and we make use of Leibniz's Rule to numerically compute the derivative of the distance. The PC prior for δ is

$$\pi^{PC}(\delta|\theta) = \frac{\theta}{2} e^{-\theta\sqrt{2\mathrm{KLD}(\delta)}} \frac{|\mathrm{KLD}'(\delta)|}{\sqrt{2\mathrm{KLD}(\delta)}}, \tag{6}$$

where θ regulates the shrinkage of the prior mass towards the base model. The higher θ the more the shrinkage.

2.1 Approximation of the KLD

The prior above has not a closed form. To this aim we perform an approximation of the KLD based on the moments of the skew-normal distribution. The approximation works pretty good, but not so good on the tails, especially when the parameter θ is small as the probability mass spreads out at the boundaries. Here, we approximate the logarithm of the normal CDF by means of a quintic polynomial regression. The amazing fact is that the intercept α gets closer and closer to $-\log 2$ as we increase the degree of the polynomial regression, and this is crucial to have the $\mathrm{KLD}(\lambda = 0) = 0$. It is not convenient to consider more moments as the quintic approximation seems to work very well. Given $Y \sim SN(\lambda)$, the KLD can be written as

$$\mathbb{E}_Y[\log\{2\Phi(\lambda Y)\}] = \log 2 + \mathbb{E}_Y(\alpha + \beta\lambda Y + \xi\lambda^2 Y^2 + \gamma\lambda^3 Y^3 + \varepsilon\lambda^4 Y^4 + \eta\lambda^5 Y^5), \tag{7}$$

where α, β, ξ, γ, ε and η are the coefficients of the polynomial regression. So, the KLD can be approximated by the first five moments of the skew-normal distribution

$$\log 2 + \alpha + \beta\lambda\sqrt{\frac{2}{\pi}}\delta + \xi\lambda^2 + \gamma\lambda^3\sqrt{\frac{2}{\pi}}(3\delta - \delta^3) + 3\varepsilon\lambda^4 + \eta\lambda^5\sqrt{\frac{2}{\pi}}(15\delta - 10\delta^3 + 3\delta^5). \quad (8)$$

In this way, we would be able to derive an analytical PC prior for λ or δ.

2.2 Bayesian inference for the skew-normal model

We check out the frequentist properties of our PC prior and we compare it to the Jeffreys' prior in [3], in order to see if there could be a certain value of the parameter θ that can be interpreted as objective. We perform a simulation study for different values of the shape parameter, for various sample sizes and for several values of the shrinkage parameter, θ. For any combination we calculate the MSE of the posterior median, the coverage probabilities and the Bayes factor. The posterior median is a reasonable choice, especially for samples where the MLE is infinite, because this entails the non finiteness of the posterior mean, see [2]. The simulation study confirms that large values of θ are quite useless, in the sense that they produce more biased estimates, especially in samples where the true $\lambda \neq 0$. A large value of θ works well only when the true $\lambda = 0$. Anyhow, the gap with respect to a small value of θ vanishes for large sample sizes. In the current work we are interested to find a particular θ that can be interpreted as objective. The simulation study shows that for θ approximately equal to 0.5, the PC prior approaches the estimates produced by the Jeffreys' prior. So, if we had to choose a noninformative value for θ we would say approximately 0.5.

2.3 Bayesian hypothesis testing

We use our PC prior for a Bayesian hypotheses test and we compare it to the Jeffreys' prior in [3], namely a $t(\lambda|\mu = 0, \sigma = \pi/2, \nu = 1/2)$. The proposition stated in Sect. 2 is very important as 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. The Bayes factor for testing

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

can be written as

$$\mathrm{BF}_{01}(x) = \frac{\prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)|_{\lambda=0}}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)\pi^{PC}(\lambda|\theta)d\lambda}. \quad (9)$$

We use a uniform importance distribution. Indeed 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 if the parameter θ is small and the asymmetry is close to 0. However, choices of small or large θ are suboptimal, in terms of convergence of the Bayes factor towards the true model. If we draw from a PC prior with a parameter θ too small, it is more likely to get extreme values of λ. Then, the marginal likelihood will be close to 0, as long as λ and x_i will have opposite signs. Suppose to draw values of λ from a PC prior with $\theta \to 0$, then for a generic x_i

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

On the other hand, the Bayes factor gives no evidence for the true model when the PC prior has a large θ. It doesn't matter what the true model is, and for $\theta \to \infty$ it will be exactly equal to 1. For $\theta \to \infty$ the PC prior becomes a Dirac centered at 0. Then

$$\text{BF}_{01}(x) = \frac{f(x|\lambda)|_{\lambda=0}}{\int_{-\infty}^{\infty} f(x|\lambda)\mathbb{I}_{\{\lambda=0\}} d\lambda} = 1, \tag{10}$$

where $\mathbb{I}_{\{\lambda=0\}}$ denotes the Dirac distribution and $f(x|\lambda)$ is the likelihood function. Simulations seem to favor a $\theta = 2$. The comparison with the Jeffreys' prior encourages the use of our prior.

3 PC prior for the degrees of freedom of the *t*-distribution

For the Gaussian base model, the Kullback-Leibler divergence can be resorted in terms of entropy and second moment of the more complex model. We use the following result

Theorem 1 (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*

$$\text{KLD}(g\|f) = -H(Y) + \frac{1}{2}\left\{\mathbb{E}(Y^2) + \log(2\pi)\right\}, \tag{11}$$

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

Proof.

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

□

We exploit the theorem above to derive the PC prior for the degrees of freedom, ν, of a t-distribution. The base model for the t-distribution is the Gaussian, which occurs when $\nu = \infty$. In [4] there is an approximation of the KLD, whilst Theorem 1 allows us to derive an analytical expression for the KLD and consequently for the PC prior. In addition, once again the prior is invariant with respect to the location-scale structure. Therefore

$$\mathrm{KLD}(\nu) = -\frac{\nu+1}{2}\left\{\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right\} - \log\left\{\sqrt{\nu}\, B\left(\frac{\nu}{2}, \frac{1}{2}\right)\right\} + \frac{1}{2}\frac{\nu}{\nu-2} - \log\left(\frac{1}{\sqrt{2\pi}}\right), \quad (12)$$

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

$$\pi(\nu) = \theta e^{-\theta\sqrt{\mathrm{A}(\nu)}} \frac{\left|\frac{1}{4}\left\{-\frac{2}{\nu} - \frac{4}{(\nu-2)^2} + (\nu+1)\Psi^{(1)}\left(\frac{\nu}{2}\right) - (\nu+1)\Psi^{(1)}\left(\frac{\nu+1}{2}\right)\right\}\right|}{\sqrt{\mathrm{A}(\nu)}}, \quad (13)$$

where $A(\nu) = 2\mathrm{KLD}(\nu)$ and $\Psi^{(1)}$ is the trigamma function.

References

1. Azzalini, A.: A class of distributions which includes the normal ones. Scand. J. Statist. **12**, 171–178 (1985)
2. Liseo, B.: La classe delle densità normali sghembe: aspetti inferenziali da un punto di vista bayesiano. Statistica **50**, 59–70 (1990)
3. Liseo, B., Loperfido, N.: A note on reference priors for the scalar skew-normal distribution. J. Stat. Plan. Infer. **136**, 373–389 (2006)
4. Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H.: Penalising model component complexity: A principled, practical approach to constructing priors (with Discussion). Stat. Sci. **32**, 1–28 (2017)