Submitted to Statistical Science

Penalising model component complexity: A principled, practical approach to constructing priors

Daniel Simpson*, **Håvard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye**

University of Bath, NTNU, University of Tromsø The Arctic University

Abstract. This document contains supplementary material for Simpson *et al.* (2016).

1. THE "DISTANCE" TO A SINGULAR MODEL

As PC priors are defined relative to a base model, which is essentially a distinguished point in the parameter space, we occasionally run into the difficulty that this point (denoted $\xi = 0$) is fundamentally different from the points $\xi > 0$. In particular, the base model is occasionally singular to the other distributions and we need to define a useful notion of distance from a singular point in the parameter space. We are saved in the context of this paper by noting that the singular model $\pi(x|\xi = 0)$ is the end point of a curve in model space $t \to \pi(x|\xi = t)$, $t \geq 0$.

Consider the ϵ-distance $d_\epsilon(t) = \sqrt{2\text{KLD}\left(\pi(x|\xi = t) \| \pi(x|\xi = \epsilon)\right)}$, which is finite for every $\epsilon > 0$. If $d_\epsilon(t) = \mathcal{O}(\epsilon^{-p})$, then we can define the renormalised distance $\tilde{d}_\epsilon(t) = \epsilon^{p/2} d_\epsilon(t)$. Using this renormalisation, $\tilde{d}(t) = \lim_{t \downarrow 0} \tilde{d}_\epsilon(t)$ is finite and parameterisation invariant and we can use it to define PC priors for singular models.

2. THE STUDENT-T CASE

In this section we will study the Student-t case focusing solely on the degrees of freedom (d.o.f.) parameter $\nu = 1/\xi$, keeping the precision fixed. This is an important non-trivial case, since the Student-t distribution is often used to robustify the Gaussian distribution. Inference based on the Gaussian distribution is well known to be vulnerable to model deviation. This can result in a significant degradation of estimation performance (Lange, Little and Taylor, 1989; Pinheiro,

Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, United Kingdom (e-mail: d.simpson@bath.ac.uk) Department of Mathematical Sciences, NTNU, Norway (e-mail: hrue@math.ntnu.no; thigm85@gmail.com; andrea.riebler@math.ntnu.no) Department of Mathematics and Statistics, University of Tromsø The Arctic University (e-mail: sigrunn.sorbye@uit.no)

*Corresponding author

Liu and Wu, 2001; Masreliez and Martin, 1977), and using the Student-t distribution can account for deviations caused by heavier tails, see for example Jacquier, Polson and Rossi (2004), Chib, Nardari and Shephard (2002) and Juárez and Steel (2010) for applications in the econometric literature.

The base model for the Student-t distribution is the Gaussian, which occurs when $\nu = \infty$. To maintain the interpretability of the precision parameter of the distribution when $\nu < \infty$, we will use a standardised version of the Student-t with unit precision for all $\nu > 2$. This follows the advice from Cox and Reid (1987), promoting that a parameter with a more orthogonal interpretation will ease the (later) joint prior specification of the precision and the d.o.f.. Our interpretation of the Occam's razor principle implies that the mode of $\pi_d(d)$ must be at $d = 0$, corresponding to the Gaussian distribution. It turns out that any proper prior for ν with finite expectation violates this principle and promotes overfitting, as $\pi_d(0) = 0$.

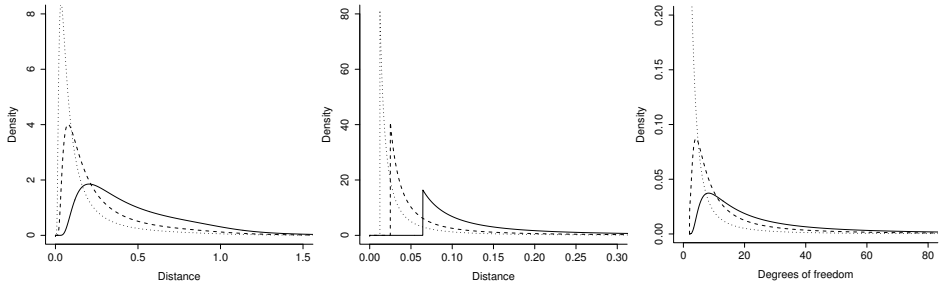
THEOREM 5. *Let $\pi_\nu(\nu)$ be an absolutely continuous prior for $\nu > 2$ where $E(\nu) < \infty$, then $\pi_d(0) = 0$ and the prior overfits in the sense of the practical version of Informal Definition 1.*

The proof is given in Section 4. 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 $d = 0$.

Commonly used priors for ν include the exponential (Geweke, 2006) or the uniform (on a finite interval) distribution (Jacquier, Polson and Rossi, 2004), which, however, place zero density mass onto the base model which may cause overfitting in accordance with Theorem 5. Notable exceptions are the work of Fonseca, Ferreira and Migon (2008) who computed (various forms of) Jeffreys' priors in the case of linear regression models with Student-t errors, Juárez and Steel (2010) and Frühwirth-Schnatter and Pyne (2010) who use a proper prior with no integer moments, and Villa and Walker (2014) who provide an objective prior for discrete values of ν.

Consider the exponential prior for $\nu > 2$ with mean equal to 5, 10 and 20. Figure 1 (a) displays these priors converted to the distance scale $d = \sqrt{2\,\mathrm{KLD}}$. Similarly, Figure 1 (b) displays the corresponding priors resulting from uniform priors on $\nu = 2$ to 20, 50 and 100. As predicted by Theorem 5, the density at $d \to 0$ is zero for all the six priors. For the exponential priors in panel (a), the mode corresponds to $\nu = 7.0$, 17.0 and 37.0, respectively. This implies that Occam's razor does not apply, in the sense that the exponential prior *defines* that the posterior will shrink towards the respective mode rather than towards the Gaussian base model. The uniform prior behaves similarly as we increase the upper limit, we put more mass to large d.o.fs and the mode moves to the left. However, the finite support implies that the density in the distance scale is zero up to the point defined by the upper limit. If the *true* distribution was Gaussian then we would overfit the data using any of these priors.

The PC prior instead is defined to have the mode at $d = 0$. To choose the parameter λ for the exponential distribution for d, a notion of *scale* is required from the user. A simple choice is to provide (U, α) such that $\mathrm{Prob}(\nu < U) = \alpha$, giving $\lambda = -\log(\alpha)/d(U)$. Figure 1 (c) shows the corresponding priors for ν



(a) (b) (c)

Fig 1: Panel (a) shows the exponential prior for $\nu > 2$ with mean equal to 5 (solid), 10 (dashed) and 20 (dotted) transformed to the distance scale. Similarly, Panel (b) shows the uniform prior for ν from 2, to 20 (solid), 50 (dashed) and 100 (dotted) on the distance scale. Panel (c) shows the PC priors for ν with $U = 10$ and $\alpha = 0.2$ (solid), $\alpha = 0.5$ (dashed) and $\alpha = 0.8$ (dotted) on the d.o.f scale.

setting $U = 10$ and $\alpha = 0.2, 0.5$ and 0.8. Here, increasing α implies increasing the deviance from the Gaussian base model.

We conclude this section by noting that the PC prior for the degrees of freedom parameter cannot be computed analytically as the Kullback-Leibler divergence does not have a closed form. For $2 < \nu < 9$, we tabulated the prior using numerical integration. For $\nu > 9$, we computed the PC prior using the following asymptotic expansion, which has absolute error less than 1.6×10^{-10}.

$$\begin{aligned}\mathrm{KLD}(\nu) \quad = \quad & \frac{3}{4}\nu^{-2} + \frac{3}{2}\nu^{-3} + \frac{17}{8}\nu^{-4} + \frac{29}{10}\nu^{-5} + \frac{61}{12}\nu^{-6} + \frac{145}{14}\nu^{-7} + \frac{273}{16}\nu^{-8} + \frac{119}{6}\nu^{-9} + \\ & \frac{869}{20}\nu^{-10} + \frac{4\,121}{22}\nu^{-11} + \frac{6\,169}{24}\nu^{-12} - \frac{30\,035}{26}\nu^{-13} - \frac{21\,843}{28}\nu^{-14} + \\ & \frac{320\,779}{10}\nu^{-15} + \frac{995\,105}{32}\nu^{-16} - \frac{28\,689\,547}{34}\nu^{-17} - \frac{28\,558\,475}{36}\nu^{-18} + \\ & \frac{1\,110\,177\,193}{38}\nu^{-19} + \frac{1\,110\,701\,481}{40}\nu^{-20} + \mathcal{O}\left(\nu^{-21}\right)\end{aligned}$$

3. RESULTS FOR SECTION 7

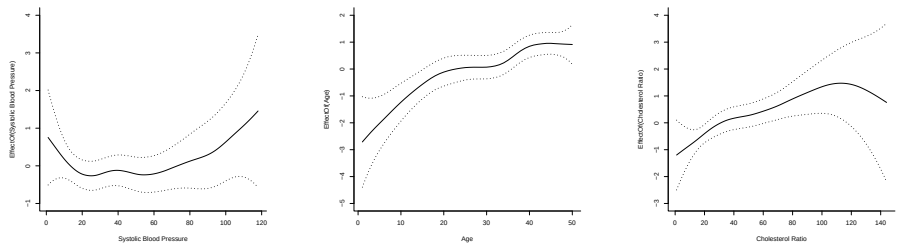
The resulting smooth effects from the case study in Section 7 are shown in Figure 2. They are similar to the results of Wood and Kohn (1998), with the differences mainly accounted for by our use of centred covariates.

4. PROOFS OF THEOREMS 1 AND 5

We give the proof of Theorem 5. The proof of Theorem 1 follows along the same lines. The KLD of approximating the unit precision Student-t with d.o.f. ν with a standard Gaussian is for large ν, $\mathrm{KLD} = \frac{3}{4}\nu^{-2} + \frac{3}{2}\nu^{-3} + \mathcal{O}(\nu^{-4})$. Since $\pi_\nu(\nu)$ has a finite first moment it is $o(\nu^{-2})$ as $\nu \to \infty$. Using the fact that $d = \sqrt{2\,\mathrm{KLD}}$, shows that $\pi_d(d) = o(1)$ as $d \to 0$.

5. PROOF OF THEOREM 2

Let $\boldsymbol{y}_n$ be n i.i.d. draws from $\pi(y|\zeta)$ and consider the hypothesis test $H_0 : \zeta = 0$ against the alternative $H_1 : \zeta \sim \pi(\zeta)$, where $\pi(0) \in (0, \infty)$. Let $m_i(\boldsymbol{y}_n)$ denote



(a) (b) (c)

Fig 2: Pointwise 95% credible intervals for the non-linear effects in the hierarchical model presented in Section 7. Panel (a) shows the effect of systolic blood pressure (BP), Panel (b) shows the effect of age at onset (Age), and Panel (c) shows the effect of low density lipoprotein cholesterol (CR).

the marginal likelihood under each model. As the domain of H_1 is open and contains only regular points for the model, the consistency of B_{01} under H_1 follows from Johnson and Rossell (2010). Assume that the model has regular asymptotic behaviour under H_0. Under H_0, Bayes' theorem implies that, for any $\zeta \geq 0$,

$$B_{01}(\boldsymbol{y}_n) = \frac{m_0(\boldsymbol{y}_n)}{m_1(\boldsymbol{y}_n)} = \frac{\pi(\boldsymbol{y}_n \mid \zeta = 0)}{\pi(\boldsymbol{y}_n \mid \zeta)} \frac{\pi(\zeta \mid \boldsymbol{y}_n)}{\pi(\zeta)} \xrightarrow{p} \frac{\pi(\boldsymbol{y}_n \mid \zeta = 0)}{\pi(\boldsymbol{y}_n \mid \zeta)} \frac{\exp(-v/2)}{\pi(0)} \sqrt{\frac{n}{8\pi v}},$$

where the second equality follows from Bochkina and Green (2014, Thm. 1), which states that $\pi(\zeta|\boldsymbol{y}_n)$ converges to a truncated normal distribution with mean parameter 0 and variance parameter $n^{-1/2}v$, and we have evaluated this asymptotic density at $\zeta = n^{-1/2}$. That $B_{01}(\boldsymbol{y}_n) = \mathcal{O}_p(n^{1/2})$ follows by noting that Bochkina and Green (2014, Assumption M1) implies that the first quotient is $\mathcal{O}_p(1)$.

When the model has irregular asymptotic behaviour, the result follows by replacing the truncated normal distribution by the appropriate Gamma distribution. It follows that $B_{01}(\boldsymbol{y}_n) = \mathcal{O}_p(n)$ in this case. The results of Bochkina and Green (2014) can also be used to extend this to the parameter invariant extensions of the non-local priors considered by Johnson and Rossell (2010). If $\pi(\zeta) = \mathcal{O}(\zeta^k)$, $k > -1$ as $\zeta \to 0$, then a simple extension of the above argument shows that $B_{01}(\boldsymbol{y}_n) = \mathcal{O}_p(n^{1/2+k})$ in the regular case and $B_{01}(\boldsymbol{y}_n) = \mathcal{O}_p(n^{1+k})$ in the irregular case.

6. PROOF OF THEOREM 3

The theorem follows by noting that as $\kappa \downarrow 0$, $\pi(\kappa) = \mathcal{O}\left(\pi_d\left(\kappa^{-1/2}\right)\kappa^{-3/2}\right)$ and as $\kappa \uparrow 1$, $\pi(\kappa) = \mathcal{O}\left(\pi_d\left(\sqrt{1-\kappa}\right)(1-\kappa)^{-1/2}\right)$.

7. PROOF OF THEOREM 4

For convenience, we will prove Theorem 4 for general priors with $\pi_d(0) \in (0, \infty)$. Consider a prior $\pi(\sigma)$ on the standard deviation with $\pi(0) = 1$ and define the re-scaled prior as $\pi^\lambda(\sigma) = \lambda\pi(\lambda\sigma)$. The following theorem shows that, for

sufficiently large $\lambda = \lambda(p) \uparrow \infty$, the marginal prior for $\boldsymbol{\beta}$ has mass on δ–sparse vectors.

THEOREM 6. *Let $\pi(\sigma)$ be a non-increasing prior on σ such that $\pi(0) = 1$ and let $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}^2)$, where $D_{ii} \sim \pi^\lambda(\sigma)$. Set $\delta_p = p^{-1}$. Sufficient condition for the prior on the δ_p–dimension to be centred at the true sparsity s are that $\lambda \geq \mathcal{O}\left(\frac{p}{\log(p)}\left[1 - \frac{s}{p}\right]\right)$ and $\pi^\lambda(p^{-1}) \leq \mathcal{O}\left(\frac{p}{\log(p)}\left[1 - \frac{s}{p}\right]\right)$, where s is the true sparsity of the target vector.*

PROOF. The result follows by noting that

$$\begin{aligned}
1 - \alpha &= 2\int_0^\delta \int_0^\infty (2\pi\sigma^2)^{-1/2} \exp\left(\frac{\beta^2}{2\sigma^2}\right) \pi^\lambda(\sigma)\, d\sigma\, d\beta \\
&\gtrsim \lambda\pi(\lambda\delta) \int_0^\delta \int_0^1 \sigma^{-1} \exp\left(\frac{\beta^2}{2\sigma^2}\right) d\lambda\, d\beta \\
&\gtrsim \lambda\pi(\lambda\delta) \int_0^\delta \log\left(1 + \frac{4}{\beta^2}\right) e^{-\frac{\beta^2}{2}}\, d\beta \gtrsim \lambda\pi(\lambda\delta) \log\left(1 + \frac{4}{\delta^2}\right) \operatorname{erf}(2^{-1/2}\delta) \\
&\gtrsim \lambda\pi(\lambda\delta)\delta \log(\delta)^{-1},
\end{aligned}$$

where the first inequality comes from the definition of $\pi^\lambda(\sigma)$ and the second follows from standard bounds on the exponential integral. Noting that

$$\int_1^\infty (2\pi\sigma^2)^{-1/2} \exp\left(\frac{\beta^2}{2\sigma^2}\right) \pi^\lambda(\sigma)\, d\sigma \lesssim 1$$

and $\pi^\lambda(\sigma) \lesssim 1$, a similar calculations yield $1 - \alpha \lesssim \lambda\delta \log(\delta^{-1})$. It follows that $\alpha = p^{-1}s$ when $\lambda\pi\left(\frac{\lambda}{p}\right) \lesssim \frac{p}{\log(p)}\left(1 - \frac{s}{p}\right) \lesssim \lambda$, which implies the result. □

Theorem 4 follows from the assumption that $s \leq \mathcal{O}\left(\frac{p}{\log(p)}\right)$ and using the Taylor expansion on the Lambert-W function to get the upper bound.

REFERENCES

BOCHKINA, N. A. and GREEN, P. J. (2014). The Bernstein-von Mises theorem and nonregular models. *The Annals of Statistics* **42** 1850–1878.

CHIB, S., NARDARI, F. and SHEPHARD, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics* **108** 281–316.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B* **49** 1–39.

FONSECA, T. C. O., FERREIRA, M. A. R. and MIGON, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika* **95** 325–333.

FRÜHWIRTH-SCHNATTER, S. and PYNE, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11** 317–336.

GEWEKE, J. (2006). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics* **8** S19–S40.

JACQUIER, E., POLSON, N. G. and ROSSI, P. E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics* **122** 185–212.

JOHNSON, V. E. and 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.

JUÁREZ, M. A. and STEEL, M. F. J. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics* **28** 52–66.

Lange, K., Little, R. and Taylor, J. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84** 881–896.

Masreliez, C. and Martin, R. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *IEEE Transactions on Automatic Control* **22** 361–371.

Pinheiro, J., Liu, C. and Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* **10** 249–276.

Villa, C. and Walker, S. G. (2014). Objective Prior for the Number of Degrees of Freedom of a t Distribution. *Bayesian Analysis* **9** 197–220.

Wood, S. and Kohn, R. (1998). A Bayesian Approach to Robust Binary Nonparametric Regression. *Journal of the American Statistical Association* **93** 203–213.