Electronic Journal of Statistics
Vol. 11 (2017) 5018–5051
ISSN: 1935-7524
https://doi.org/10.1214/17-EJS1337SI

Sparsity information and regularization in the horseshoe and other shrinkage priors

Juho Piironen and Aki Vehtari

Helsinki Institute for Information Technology, HIIT
Department of Computer Science, Aalto University
e-mail: juho.piironen@aalto.fi; aki.vehtari@aalto.fi

Abstract: The horseshoe prior has proven to be a noteworthy alternative for sparse Bayesian estimation, but has previously suffered from two problems. First, there has been no systematic way of specifying a prior for the global shrinkage hyperparameter based on the prior information about the degree of sparsity in the parameter vector. Second, the horseshoe prior has the undesired property that there is no possibility of specifying separately information about sparsity and the amount of regularization for the largest coefficients, which can be problematic with weakly identified parameters, such as the logistic regression coefficients in the case of data separation. This paper proposes solutions to both of these problems. We introduce a concept of effective number of nonzero parameters, show an intuitive way of formulating the prior for the global hyperparameter based on the sparsity assumptions, and argue that the previous default choices are dubious based on their tendency to favor solutions with more unshrunk parameters than we typically expect a priori. Moreover, we introduce a generalization to the horseshoe prior, called the regularized horseshoe, that allows us to specify a minimum level of regularization to the largest values. We show that the new prior can be considered as the continuous counterpart of the spike-and-slab prior with a finite slab width, whereas the original horseshoe resembles the spike-and-slab with an infinitely wide slab. Numerical experiments on synthetic and real world data illustrate the benefit of both of these theoretical advances.

MSC 2010 subject classifications: Primary 62F15.
Keywords and phrases: Bayesian inference, sparse estimation, shrinkage priors, horseshoe prior.

Received June 2017.

1. Introduction

This paper deals with sparse Bayesian estimation and is an extension to our earlier work (Piironen and Vehtari, 2017a). We consider statistical models with a large number of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$ but so that it is reasonable to assume that only some of them are far from zero. A typical example – and also the case we will mostly focus in this paper – is a regression or classification problem with a large number of predictor variables out of which we expect only a few to be relevant and therefore have a regression coefficient distinguishable from zero.

A vast number of different estimators, both Bayesian and non-Bayesian, have been proposed for these problems. In the non-Bayesian literature the sparse problems are typically handled by Lasso (Tibshirani, 1996) or one of its generalizations (for an overview, see e.g., Hastie, Tibshirani and Wainwright, 2015). We focus on the probabilistic approach and carry out full Bayesian inference on the problem.

Two prior choices dominate the Bayesian literature: two component discrete mixture priors known as the spike-and-slab (Mitchell and Beauchamp, 1988; George and McCulloch, 1993), and a variety of continuous shrinkage priors (see e.g., Polson and Scott, 2011, and references therein). The spike-and-slab prior is intuitively appealing as when the spike is taken to be a delta-spike in the origin, it is equivalent to Bayesian model averaging (BMA) (Hoeting et al., 1999) over the variable combinations, and often has good performance in practice. The disadvantages are that the results can be sensitive to prior choices (slab width and prior inclusion probability) and that the posterior inference can be computationally demanding with a large number of variables, due to the huge model space. The inference could be sped up by analytical approximations using either expectation propagation (EP) (Hernández-Lobato, Hernández-Lobato and Suárez, 2010, 2015) or variational inference (VI) (Titsias and Lázaro-Gredilla, 2011), but this comes at the cost of a substantial increase in the amount of analytical work needed to derive the equations separately for each model and a more complex implementation.

The continuous shrinkage priors on the other hand are easy to implement, provide convenient computation using generic sampling tools such as Stan (Stan Development Team, 2017), and can yield as good or better results. A particularly interesting example is the horseshoe prior (Carvalho, Polson and Scott, 2009, 2010)

$$\begin{aligned} \theta_j \,|\, \lambda_j, \tau &\sim \mathrm{N}\big(0, \tau^2 \lambda_j^2\big)\,, \\ \lambda_j &\sim \mathrm{C}^+(0,1)\,, \quad j = 1, \ldots, D, \end{aligned} \tag{1.1}$$

which has shown comparable performance to the spike-and-slab prior in a variety of examples where a sparsifying prior on the model parameters θ_j is desirable (Carvalho, Polson and Scott, 2009, 2010; Polson and Scott, 2011). The horseshoe is one of the so called global-local shrinkage priors, meaning that there is a global hyperparameter τ that shrinks all the parameters towards zero, while the heavy-tailed half-Cauchy priors for the local hyperparameters λ_j allow some θ_j to escape the shrinkage (see Sec. 2.1 for more thorough discussion).

Despite its good performance in many problems, the horseshoe prior has previously suffered from two shortcomings. First, there has been no consensus on how to carry out inference for the global hyperparameter τ which determines the overall sparsity in the parameter vector $\boldsymbol{\theta}$ and therefore has a large impact on the results. We prefer full Bayesian inference (see Sec. 3.1) but the existing methodology has been lacking a systematic way of placing a prior for τ based on the information about the sparsity. Second, the horseshoe prior has the undesired property that the parameters far from zero are not regularized at all

(see Sec. 2.3). While this is often considered as a key strength of the prior, it can be harmful especially when the parameters are only weakly identified by the data, for instance in the case of a flat likelihood due to separable data in logistic regression.

We propose a solution to both of these problems. We introduce a concept of effective number of nonzero parameters m_{eff} (Sec. 3.3), derive analytically its relationship between the global shrinkage parameter τ, and show an easy and intuitive way of formulating the prior for τ based on the prior information about the sparsity of $\boldsymbol{\theta}$. Based on these theoretical considerations, we argue that the previously proposed default priors are dubious based on the prior they impose on m_{eff}, and that they yield good results only when τ (and therefore m_{eff}) is strongly identified by the data. Moreover, we introduce a generalization of the horseshoe prior, called the regularized horseshoe, that operates otherwise similarly as the original horseshoe but allows specifying the regularization to the coefficients that are far from zero (see Sec. 2.3). We show that the regularized horseshoe can be considered as the continuous counterpart of the spike-and-slab prior with a finite slab width, whereas the original horseshoe resembles the spike-and-slab with an infinitely wide slab. The benefit of both of these theoretical advances will be illustrated with examples on synthetic and real world data (Sec. 4).

As a final remark, although we focus on the horseshoe in our discussion, we want to emphasize that both of these ideas could also be applied to other shrinkage priors, and several promising alternatives to the horseshoe have been proposed during the recent years (Bhattacharya et al., 2015; Zhang, Reich and Bondell, 2016; Ghosh, Li and Mitra, 2017).

2. Horseshoe prior and its extension

This section discusses the horseshoe and its connection to the spike-and-slab prior (Section 2.2). We also present an extension (Section 2.3) that both helps understanding the theoretical properties of the original horseshoe and – as will be demonstrated in Section 4 – robustifies the prior and improves its practical performance.

2.1. Horseshoe prior for linear regression

Consider the single output linear Gaussian regression model with several input variables, given by

$$y_i = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{x}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\big(0, \sigma^2\big), \quad i = 1, \ldots, n\,, \tag{2.1}$$

where $\mathbf{x}$ is the D-dimensional vector of inputs, $\boldsymbol{\beta}$ contains the corresponding weights and σ^2 is the noise variance. The horseshoe prior is set for the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)$

$$\begin{aligned} \beta_j \,|\, \lambda_j, \tau &\sim \mathrm{N}\big(0, \tau^2 \lambda_j^2\big), \\ \lambda_j &\sim \mathrm{C}^+(0, 1)\,, \quad j = 1, \ldots, D. \end{aligned} \tag{2.2}$$

If an intercept term β_0 is included in model (2.1), we give it a relatively flat prior, because there is usually no reason to shrink it towards zero. As discussed in the introduction, the horseshoe prior has been shown to possess several desirable theoretical properties and good performance in practice (Carvalho, Polson and Scott, 2009, 2010; Polson and Scott, 2011; Datta and Ghosh, 2013; van der Pas, Kleijn and van der Vaart, 2014). The intuition is the following: the global parameter τ pulls all the weights globally towards zero, while the thick half-Cauchy tails for the local scales λ_j allow some of the weights to escape the shrinkage. Different levels of sparsity can be accommodated by changing the value of τ: with large τ all the variables have very diffuse priors with very little shrinkage, but letting $\tau \to 0$ will shrink all the weights β_j to zero.

The above can be formulated more formally as follows. Let $\mathbf{X}$ denote the n-by-D matrix of observed inputs and $\mathbf{y}$ the observed targets. The conditional posterior for the coefficients $\boldsymbol{\beta}$ given the hyperparameters and data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ can be written as

$$\begin{aligned} p(\boldsymbol{\beta} \,|\, \boldsymbol{\Lambda}, \tau, \sigma^2, \mathcal{D}) &= \mathrm{N}\big(\boldsymbol{\beta} \,|\, \bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\big), \\ \bar{\boldsymbol{\beta}} &= \tau^2 \boldsymbol{\Lambda} \left(\tau^2 \boldsymbol{\Lambda} + \sigma^2 (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1} \hat{\boldsymbol{\beta}}, \\ \boldsymbol{\Sigma} &= (\tau^{-2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}, \end{aligned}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\big(\lambda_1^2, \ldots, \lambda_D^2\big)$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ is the maximum likelihood solution (assuming the inverse exists). If the predictors are uncorrelated with zero mean and variances $\mathrm{Var}(x_j) = s_j^2$, then $\mathbf{X}^\mathsf{T}\mathbf{X} \approx n\,\mathrm{diag}\big(s_1^2, \ldots, s_D^2\big)$, and we can approximate

$$\bar{\beta}_j = (1 - \kappa_j)\hat{\beta}_j, \tag{2.3}$$

where

$$\kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2 s_j^2 \lambda_j^2} \tag{2.4}$$

is the *shrinkage factor* for coefficient β_j. The shrinkage factor describes how much coefficient β_j is shrunk towards zero from the maximum likelihood solution ($\kappa_j = 1$ meaning complete shrinkage and $\kappa_j = 0$ no shrinkage). From (2.3) and (2.4) it is easy to verify that $\bar{\boldsymbol{\beta}} \to 0$ as $\tau \to 0$, and $\bar{\boldsymbol{\beta}} \to \hat{\boldsymbol{\beta}}$ as $\tau \to \infty$.

The result (2.4) holds for any prior that can be written as a scale mixture of Gaussians like (2.2), regardless of the prior for λ_j. The horseshoe employs independent half-Cauchy priors for all λ_j, and for this choice one can show that, for fixed τ and σ, the shrinkage factor (2.4) follows the prior

$$p(\kappa_j \,|\, \tau, \sigma) = \frac{1}{\pi} \frac{a_j}{(a_j^2 - 1)\kappa_j + 1} \frac{1}{\sqrt{\kappa_j}\sqrt{1 - \kappa_j}}, \tag{2.5}$$

where $a_j = \tau\sigma^{-1}\sqrt{n}\, s_j$. When $a_j = 1$, this reduces to $\mathrm{Beta}\big(\frac{1}{2}, \frac{1}{2}\big)$ which looks like a horseshoe, see Figure 1. Thus, a priori, we expect to see both relevant

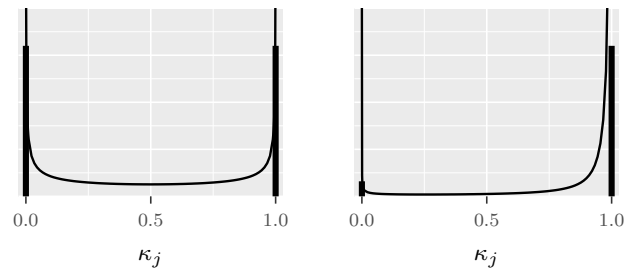

FIG 1. *The continuous curves show the densities for the shrinkage factor* (2.4) *for the horseshoe prior* (2.2) *when* $a_j = \tau\sigma^{-1}\sqrt{n}\, s_j = 1$ *(left) and when* $a_j = 0.1$ *(right). The bars denote the corresponding point mass function for the spike-and-slab prior* (2.7) *with infinite slab width* $c \to \infty$*, when* $\pi/(1-\pi) = 1$ *(left) and* $\pi/(1-\pi) = 0.1$ *(right). To aid visualization, the bars illustrating the point masses are scaled and show only the relative probability masses.*

($\kappa_j = 0$, no shrinkage) and irrelevant ($\kappa_j = 1$, complete shrinkage) variables. By changing the value of τ, the prior for κ_j places more mass either close to 0 or 1. For instance, choosing τ so that $a_j = 0.1$ favors complete shrinkage ($\kappa = 1$) and thus we expect more coefficients to be close to zero a priori. Notice though that for a fixed τ, the sparsity assumptions will be dependent on the input dimension D, and to get around this issue, we need to consider the values of all the shrinkage factors κ_j together. Using this idea, Section 3 discusses an intuitive way of designing a prior distribution for τ based on the assumptions about the number of nonzero components in $\boldsymbol{\beta}$.

Notice also that those variables which vary on larger scale s_j are treated as more relevant a priori, which is the reason why we usually scale all the variables to have unit variance $s_j^2 = 1$, unless the original scales really carry information about the relevances. Another way would be to use the original scales but adjust scales for the local parameters accordingly $\lambda_j \sim \mathrm{C}^+\big(0, s_j^{-2}\big)$.

2.2. Spike-and-slab prior

The spike-and-slab (Mitchell and Beauchamp, 1988; George and McCulloch, 1993) is a popular shrinkage prior that is often considered as the "gold standard" for sparse Bayesian estimation. The prior is often written as a two-component mixture of Gaussians

$$\begin{aligned} \beta_j \,|\, \lambda_j, c, \varepsilon &\sim \lambda_j \,\mathrm{N}\big(0, c^2\big) + (1-\lambda_j)\,\mathrm{N}\big(0, \varepsilon^2\big), \\ \lambda_j &\sim \mathrm{Ber}(\pi), \qquad j = 1, \ldots, D, \end{aligned} \tag{2.6}$$

so that $\varepsilon \ll c$ and the indicator variable $\lambda_j \in \{0, 1\}$ denotes whether the coefficient β_j is close to zero (comes from the "spike", $\lambda_j = 0$) or nonzero (comes from the "slab", $\lambda_j = 1$). Often we set $\varepsilon = 0$, that is, the spike is taken to be a delta spike at the origin δ_0, although also $\varepsilon > 0$ could be used (George and McCulloch, 1993). The user then has to specify the values (or priors) for

the slab width c and the prior inclusion probability π, which encodes the prior information about the sparsity of the coefficient vector $\boldsymbol{\beta}$. Fixing c is probably the most common approach but by giving it a hyperprior, one can obtain a more heavy-tailed, such as Laplacian, slab (Johnstone and Silverman, 2004).

With the choice $\varepsilon = 0$, the prior (2.6) can be written analogous to (2.2) as

$$\begin{aligned} \beta_j \mid \lambda_j, c &\sim \mathrm{N}\big(0, c^2\lambda_j^2\big), \\ \lambda_j &\sim \mathrm{Ber}(\pi), \qquad j = 1, \ldots, D, \end{aligned} \tag{2.7}$$

so instead of giving continuous priors for λ_js as in the horseshoe, here only two values ($\lambda_j = 0, 1$) are allowed. Thus also the shrinkage factor κ_j only has mass at $\kappa_j = \frac{1}{1+n\sigma^{-2}s_j^2c^2}$ and at $\kappa_j = 1$, and the probabilities are π and $1-\pi$, respectively. Letting $c \to \infty$, all the mass is concentrated at the extremes $\kappa_j = 0$ and $\kappa_j = 1$, and the resemblance to the horseshoe becomes obvious, see Figure 1. Given the similarity of the shrinkage profiles between the horseshoe and spike-and-slab, it is not surprising that the two priors have shown comparable performance in a variety of experiments (Carvalho, Polson and Scott, 2009, 2010; Polson and Scott, 2011). The next section discusses an extension of the horseshoe that closely resembles the spike-and-slab prior with a finite slab width $c < \infty$.

2.3. Regularized horseshoe

As discussed in Section 2.1, the horseshoe prior favors solutions $\beta_j \approx 0$ and $\beta_j \approx \hat{\beta}_j$, and it can be shown that under certain conditions, $\bar{\beta}_j \to \hat{\beta}_j$ when $|\hat{\beta}_j| \to \infty$ (Carvalho, Polson and Scott, 2010). While this guarantees that the strong signals will not be overshrunk – and is often considered to be one of the key assets of the prior – this property can also be harmful, especially when the parameters are weakly identified. An example of such case is the flat likelihood arising in logistic regression with separable data. As the horseshoe has Cauchy tails, in these problems it suffers basically from the same problems as the Cauchy prior, namely that the posterior means for the regression coefficients may vanish (Ghosh, Li and Mitra, 2017). Therefore it would be very useful to be able to control the amount of shrinkage for the largest coefficients, which in spike-and-slab prior (Sec. 2.2) is achieved by controlling the slab width.

To guarantee that the prior always shrinks the coefficients at least by a small amount, we introduce the following *regularized horseshoe* prior

$$\begin{aligned} \beta_j \mid \lambda_j, \tau, c &\sim \mathrm{N}\Big(0, \tau^2\tilde{\lambda}_j^2\Big), \quad \tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{c^2 + \tau^2\lambda_j^2}, \\ \lambda_j &\sim \mathrm{C}^+(0,1)\,, \quad j = 1, \ldots, D, \end{aligned} \tag{2.8}$$

where $c > 0$ is a constant that we assume is given for now. The intuition behind this definition is the following. When $\tau^2\lambda_j^2 \ll c^2$, meaning the coefficient β_j is close to zero, then $\tilde{\lambda}_j^2 \to \lambda_j^2$ and the prior (2.8) approaches the original horseshoe. However, when $\tau^2\lambda_j^2 \gg c^2$, meaning the coefficient is far from zero,

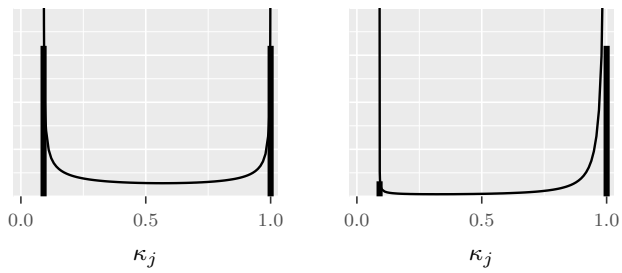

FIG 2. *Shrinkage profiles as in Figure 1 but now for the regularized horseshoe (2.8) and spike-and-slab (2.7) with a finite slab width $c = 1$. In comparison to Figure 1, for both priors the first mode is shifted from $\kappa_j = 0$ to $\kappa_j = 1/(1 + n\sigma^{-2}s_j^2c^2)$ (for the plots we have selected $n\sigma^{-2}s_j^2 = 10$).*

then $\tilde{\lambda}_j^2 \to c^2/\tau^2$ and the prior (2.8) approaches $\mathrm{N}\big(0, c^2\big)$. Thus the prior will shrink the small signals as the horseshoe but will also regularize even the largest coefficients as a Gaussian slab with variance c^2.

Another way to see this is to notice that the conditional prior for β_j can be factored as

$$p(\beta_j \mid \lambda_j, \tau, c) \propto \mathrm{N}\big(0, \tau^2\lambda_j^2\big)\,\mathrm{N}\big(0, c^2\big) \propto \mathrm{N}\Big(0, \tau^2\tilde{\lambda}_j^2\Big), \tag{2.9}$$

from which it is easy to see that depending on the relative magnitudes of $\tau^2\lambda_j^2$ and c^2, the prior operates (roughly) as the narrower one of the two factors. Therefore the role of $\mathrm{N}\big(0, c^2\big)$ is to "soft-truncate" the extreme tails of the horseshoe, thereby controlling the magnitude of the largest β_js. Letting $c \to \infty$, we recover the original horseshoe.

The shrinkage profile of the regularized horseshoe is illustrated in Figure 2 together with the spike-and-slab with the slab width c, which demonstrates the similarity of the two priors. Using $c < \infty$ has the advantage that it regularizes the parameters β_j when they are weakly identified, and allows us also to specify our prior information about the maximum effect β_j we expect to see. The benefit of the proposed approach is illustrated in Section 4.1.2.

It must be noted that the shrinkage profile in Figure 2 does not have *exactly* the same shape as the original horseshoe shifted and scaled from interval $(0, 1)$ to $(b, 1)$ with $b > 0$, although it is very close to this. There is slightly more mass near the left hand side mode, although the difference is too small to be visible in Figure 2. It is possible to retain the exact shape of the horseshoe by defining the modified local parameter $\tilde{\lambda}_j^2$ as

$$\tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{\frac{\sigma^2}{ns_j^2} + c^2 + \tau^2\lambda_j^2}. \tag{2.10}$$

The details of this result are spelled out in Appendix A. The reason why we define the prior as (2.8) is that unless n or the slab width c^2 is very small, the

term $\frac{\sigma^2}{ns_j^2}$ is typically small compared to c^2 and leaving it out has little influence in practice. On the other hand, formulation (2.8) is simpler and has the nice interpretation as a product of the original horseshoe and the Gaussian slab (2.9). Thus we report the result (2.10) only for completeness.

Our formulation requires choosing a value or prior for c. Unless substantial knowledge about the scale of the relevant coefficients exists, we generally recommend placing a prior for c instead of fixing it. Often a reasonable choice is

$$c^2 \sim \text{Inv-Gamma}(\alpha, \beta), \quad \alpha = \nu/2, \quad \beta = \nu s^2/2, \tag{2.11}$$

which translates to a Student-$t_\nu(0, s^2)$ slab for the coefficients far from zero and is typically a good default choice for a weakly informative prior. Another motivation for using inverse-Gamma is that it has a heavy right tail accompanied by a light left tail thereby preventing much mass from accumulating near zero. This is natural as we do not want to shrink those coefficients heavily towards zero that are already deemed to be far from zero. Still, we emphasize that our approach is not limited to this choice and also other hyperpriors are possible. It would also be possible to use variable specific slab widths c_i, but we do not explore this further in this paper and leave it for future investigation.

Finally, we would also like to point out that the tail-cutting idea of Equation (2.9) could be used more generally with other priors when relevant. For instance, we expect our idea to be useful with the horseshoe+ prior of Bhadra et al. (2017) which also has Cauchy tails and therefore suffers from the same problem as the horseshoe.

2.4. Hierarchical shrinkage

In addition to the problems with vanishing means, earlier we have also reported sampling issues with the original horseshoe even in simple regression problems (Piironen and Vehtari, 2015). Technically speaking, the problem arises due to posterior having an extreme funnel shape which is challenging for Markov chain Monte Carlo (MCMC) methods. The problem was revealed with the help of the divergence diagnostics of the NUTS algorithm (Hoffman and Gelman, 2014; Betancourt and Girolami, 2015; Betancourt, 2017a,b) when fitting the models in Stan.

The problem is related to the thick Cauchy tails of the prior, and to overcome the sampling issues, in our technical report we tentatively proposed replacing the half-Cauchy priors for the local parameters λ_j in (2.2) with half-t priors with small degrees of freedom, such as $\nu = 3$, and named this approach the "hierarchical shrinkage". With large enough ν, this seems to help with the sampling issues and remove the divergent transitions produced by NUTS, but the drawback is that the prior becomes less sparsifying. This is because when the tails of $p(\lambda_j)$ are made slimmer, we need to increase the value for τ to accommodate large signals, and therefore the prior is not able to shrink the small coefficients efficiently towards zero. Another limitation of this approach is that

in order to fight the problems arising from data separation in logistic regression (Sec. 2.3), we would also need to refrain from using half-Cauchy prior for the global parameter τ (which we might want to use, see Sec. 3) as this would also lead to Cauchy tails for $p(\beta_j)$.

The less sparsifying nature of the choice $\nu > 1$ will be demonstrated in practice in Section 4.1.2, where we also show that the regularized horseshoe (Sec. 2.3) clearly outperforms this approach. Thus we no longer recommend increasing the local degrees of freedom, but instead of using the regularized horseshoe.

3. The global shrinkage parameter

This section discusses the prior choice for the global hyperparameter τ. We begin with a short note on why we prefer full Bayesian inference for τ over point estimation, and then go on to discuss how we propose to set up the prior $p(\tau)$.

3.1. Full Bayes versus point estimation

In principle, one could use a plug-in estimate for τ, obtained either by cross-validation or maximum marginal likelihood (sometimes referred to as "empirical Bayes"). The maximum marginal likelihood estimate has the drawback that it is always in danger of collapsing to $\hat{\tau} = 0$ if the parameter vector happens to be very sparse. Moreover, rather than being computationally convenient, this approach might actually complicate matters as the marginal likelihood is not analytically available for non-Gaussian likelihoods. While cross-validation avoids the latter problem and possibly also the first one, it is computationally less efficient than the full Bayesian solution and fails to account for the posterior uncertainty. For these reasons we recommend full Bayesian inference for τ, and focus on how to specify the prior distribution.

3.2. Earlier approaches

Carvalho, Polson and Scott (2009) also recommend full Bayesian inference for τ, and following Gelman (2006), they propose prior

$$\tau \sim \mathrm{C}^+(0, 1), \tag{3.1}$$

whereas Polson and Scott (2011) recommend

$$\tau \,|\, \sigma \sim \mathrm{C}^+\big(0, \sigma^2\big). \tag{3.2}$$

If the target variable y is scaled to have marginal variance of one, unless the noise level σ is very small, both of these priors typically lead to quite similar posteriors. However, as we argue in Section 3.3, there is a theoretical justification

for letting τ scale with σ. The main motivation for using a half-Cauchy prior for τ is that it evaluates to a finite positive value at the origin, yielding a proper posterior and allowing even complete shrinkage $\tau \to 0$, while still having a thick tail which can accommodate a wide range of values. For these reasons, $\mathrm{C}^+\big(0, \eta^2\big)$ is a desirable choice when there are enough observations to let τ be identified by data. Still, we show that in several cases one can clearly benefit by choosing the scale η in a more careful manner than simply $\eta = 1$ or $\eta = \sigma$, because for most applications these choices place far to much mass for implausibly large values of τ. This point is discussed in Section 3.3. Moreover, the synthetic example in Section 4.1.1 shows that in some cases one could clearly benefit from even more informative prior.

van der Pas, Kleijn and van der Vaart (2014) study the optimal selection of τ in model

$$y_i \sim \beta_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\big(0, \sigma^2\big), \quad i = 1, \ldots, n. \tag{3.3}$$

They prove that in such a setting, the optimal value (up to a log factor) in terms of mean squared error and posterior contraction rates in comparison to the true $\boldsymbol{\beta}^*$ is

$$\tau^* = \frac{p^*}{n}, \tag{3.4}$$

where p^* denotes the number of nonzeros in the true coefficient vector $\boldsymbol{\beta}^*$ (assuming such exists). Their proofs assume that $n, p^* \to \infty$ and $p^* = o(n)$. Model (3.3) corresponds to setting $\mathbf{X} = \mathbf{I}$ and $D = n$ in the usual regression model (2.1). It is unclear whether and how this result could be extended to a more general $\mathbf{X}$, and how one should utilize this result when p^* is unknown (as it usually is in practice). In section 3.3, we formulate our method of constructing the prior $p(\tau)$ based on the prior information about p^*, and show that if p^* was known, our method would also give rise to result (3.4), but is more generally applicable.

3.3. Effective number of nonzero coefficients

Consider the prior distribution for the shrinkage factor of the jth regression coefficient for the linear Gaussian model, Eq. (2.5). The mean and variance can be shown to be

$$\mathrm{E}(\kappa_j \mid \tau, \sigma) = \frac{1}{1 + a_j}, \tag{3.5}$$

$$\mathrm{Var}(\kappa_j \mid \tau, \sigma) = \frac{a_j}{2(1 + a_j)^2}, \tag{3.6}$$

where $a_j = \tau\sigma^{-1}\sqrt{n}\, s_j$ as earlier. A given value for the global parameter τ can be understood intuitively via the prior distribution that it imposes on the

effective number of coefficients distinguishable from zero (or effective number of nonzero coefficients, for short), which we define as

$$m_{\text{eff}} = \sum_{j=1}^{D} (1 - \kappa_j). \tag{3.7}$$

When the shrinkage factors κ_j are close to 0 and 1 (as they typically are for the horseshoe prior), this quantity describes essentially how many active or unshrunk variables we have in the model. It serves therefore as a useful indicator of the effective model size.

Using results (3.5) and (3.6), the mean and variance of m_{eff} given τ and σ are given by

$$\mathrm{E}(m_{\text{eff}} \,|\, \tau, \sigma) = \sum_{j=1}^{D} \frac{a_j}{1 + a_j} \tag{3.8}$$

$$\mathrm{Var}(m_{\text{eff}} \,|\, \tau, \sigma) = \sum_{j=1}^{D} \frac{a_j}{2(1 + a_j)^2}, \tag{3.9}$$

Let us now assume that, in addition of having a zero mean, each variable also has a unit variance $s_j^2 = 1$. In this case the equations above simplify to

$$\mathrm{E}(m_{\text{eff}} \,|\, \tau, \sigma) = \frac{\tau\sigma^{-1}\sqrt{n}}{1 + \tau\sigma^{-1}\sqrt{n}} D, \tag{3.10}$$

$$\mathrm{Var}(m_{\text{eff}} \,|\, \tau, \sigma) = \frac{\tau\sigma^{-1}\sqrt{n}}{2(1 + \tau\sigma^{-1}\sqrt{n})^2} D. \tag{3.11}$$

The expression for the mean (3.10) is helpful. First, from this expression it is evident that to keep our prior information about m_{eff} consistent, τ must scale as $\sigma/\sqrt{n}$. Priors that fail to do so, such as (3.1), favor models of varying size depending on the noise level σ and the number of data points n. Second, if our prior guess for the number of relevant variables is p_0, it is reasonable to choose the prior so that most of the prior mass is located near the value

$$\tau_0 = \frac{p_0}{D - p_0} \frac{\sigma}{\sqrt{n}}, \tag{3.12}$$

which is obtained by solving equation $\mathrm{E}(m_{\text{eff}} \,|\, \tau, \sigma) = p_0$. Note that this is typically quite far from 1 or σ, which are used as scales for priors (3.1) and (3.2). For instance, if $D = 1000$ and $n = 200$, then prior guess $p_0 = 5$ gives about $\tau_0 = 3.6 \cdot 10^{-4}\sigma$.

To further develop the intuition about the connection between τ and m_{eff}, it is helpful to visualize the prior imposed on m_{eff} for different prior choices for τ. This is most conveniently done by drawing samples for m_{eff}; we first draw $\tau \sim p(\tau)$ and $\lambda_1, \ldots, \lambda_D \sim \mathrm{C}^+(0, 1)$, then compute the shrinkage factors $\kappa_1, \ldots, \kappa_D$ from (2.4), and finally m_{eff} from (3.7).

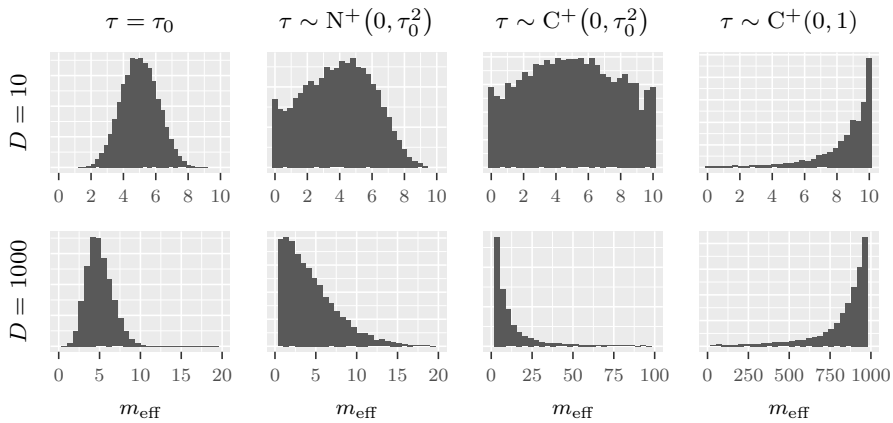

FIG 3. *Histograms of prior draws for m_{eff} (effective number of nonzero regression coefficients, Eq. (3.7)) imposed by different prior choices for τ, when the total number of input variables is $D = 10$ and $D = 1000$. τ_0 is computed from formula (3.12) assuming $n = 100$ observations with $\sigma = 1$ and $p_0 = 5$ as the prior guess for the number of relevant variables. Note the varying scales on the horizontal axes in the bottom row plots.*

Figure 3 shows histograms of prior draws for m_{eff} for some different prior choices for τ, with total number of variables $D = 10$ and $D = 1000$, assuming $n = 100$ observations with $\sigma = 1$. The first three priors utilize the value τ_0 which is computed from (3.12) using $p_0 = 5$ as our hypothetical prior guess for the number of relevant variables. Fixing $\tau = \tau_0$ results in a nearly symmetric distribution around p_0, while a half-normal prior with scale τ_0 yields a skewed distribution favoring solutions with $m_{\mathrm{eff}} < p_0$ but allowing larger values to also be accommodated. The half-Cauchy prior behaves similarly to the half-normal, but results in a distribution with a much thicker tail giving substantial mass also to values much larger than p_0 when D is large. Figure 3 also illustrates why prior $\tau \sim \mathrm{C}^+(0, 1)$ is often a dubious choice: it places far too much mass on large values of τ, consequently favoring solutions with most of the coefficients unshrunk. Thus when only a small number of the variables are relevant – as we typically assume – this prior results in sensible inference only when τ is strongly identified by data. Notice also that, if we changed the value of σ or n, the first three priors for τ would still impose the same prior for m_{eff}, but this is not true for $\tau \sim \mathrm{C}^+(0, 1)$.

This way, by studying the prior for m_{eff}, one can easily choose the prior for τ based on the information about the number of nonzero parameters. Because the prior information can vary substantially for different problems and the results depend on the information carried by the data, there is no globally optimal prior for τ that works for every single problem. Some recommendations, however, will be given in Section 5 based on these theoretical considerations and experiments presented in Section 4.

We conclude by pointing out a connection between our reference value (3.12) and the oracle result (3.4) for the simplified model (3.3). As pointed out in the

last section, model (3.3) corresponds to setting $\mathbf{X} = \mathbf{I}$ (which implies $n = D$ and $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$) in the usual regression model (2.1). Using this fact and repeating the steps needed to arrive at (3.12), we get

$$\tau_0 = \frac{p_0}{D - p_0}\sigma. \tag{3.13}$$

Suppose now that we select $p_0 = p^*$, that is, our prior guess is oracle. Using the same assumptions as van der Pas, Kleijn and van der Vaart (2014), namely that $n, p^* \to \infty$ and $p^* = o(n)$, and additionally that $\sigma = 1$, we get $\tau_0 \to p^*/D = \tau^*$. This result is natural, as it means it is optimal to choose τ so that the imposed prior for the effective number of nonzero coefficients m_{eff} is centered at the true number of nonzeros p^*. This further motivates why m_{eff} is a useful quantity.

3.4. Regularized horseshoe and other shrinkage priors

As discussed in Section 2.3, when we set $c < \infty$, the shrinkage profile of the regularized horseshoe (2.8) becomes approximately equivalent to that of the horseshoe shifted from interval $(0, 1)$ to $(b_j, 1)$, where $b_j = \frac{1}{1+n\sigma^{-2}s_j^2c^2}$. Thus the shrinkage factor under the regularized horseshoe satisfies approximately $\tilde{\kappa}_j = (1 - b_j)\kappa_j + b_j$, where κ_j denotes the shrinkage factor for the original horseshoe. From this we get $1 - \tilde{\kappa}_j = (1 - b_j)(1 - \kappa_j)$. Assuming further that all the variables have a unit variance $s_j^2 = 1$ and thus $b_j = b = \frac{1}{1+n\sigma^{-2}c^2}$, the effective model complexity under the regularized horseshoe satisfies

$$\tilde{m}_{\text{eff}} = (1 - b)m_{\text{eff}},$$

where m_{eff} is the effective number of nonzeros for the original horseshoe. Thus with a given τ, the effective complexity for the regularized horseshoe is always less than for the pure horseshoe, because those coefficients that are far from zero are still affected by the slab. Therefore we can naturally use result (3.12) also for the regularized horseshoe with p_0 as our prior guess for the number of coefficients far from zero, but remembering that those coefficients will experience the regularization by the slab.

The concept of effective number of nonzeros could also be used with shrinkage priors other than the (regularized) horseshoe, as long as the prior can be written as a scale mixture of Gaussians like (2.2). Many such alternatives have been proposed, including the double-exponential or Laplace (Park and Casella, 2008), Dirichlet-Laplace (Bhattacharya et al., 2015), R-square induced Dirichlet decomposition (Zhang, Reich and Bondell, 2016), and horseshoe+ (Bhadra et al., 2017). For instance, the Dirichlet-Laplace prior has a Dirichlet concentration hyperparameter a that strongly affects the sparsity properties of the prior, and based on the experiments of Bhattacharya et al. (2015) has a substantial effect on the results. It would therefore be interesting the investigate a prior information calibrated selection of a using our framework.

Depending on the prior, corresponding analytical results like (3.10) and (3.11) may or may not be available, but as long as one is able to sample both from

$p(\tau)$ and $p(\lambda_j)$, it is always easy to draw samples from the prior distribution for m_{eff}, and therefore investigate the effect of hyperprior or hyperparameter choice on the effective model complexity. It must be noted though, that for those prior for which the shrinkage factors κ are not near 0 or 1, the values of m_{eff} can be more difficult to interpret.

3.5. Non-Gaussian observation models

When the observation model is non-Gaussian, the exact analysis from Section 3.3 is analytically intractable. We can, however, perform the analysis using a Gaussian approximation to the likelihood. Using the second order Taylor expansion for the log likelihood, the approximate posterior for the regression coefficients given the hyperparameters becomes

$$\begin{aligned} p(\boldsymbol{\beta} \,|\, \boldsymbol{\Lambda}, \tau, \phi, \mathcal{D}) &\approx \mathrm{N}\big(\boldsymbol{\beta} \,|\, \bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\big), \\ \bar{\boldsymbol{\beta}} &= \tau^2 \boldsymbol{\Lambda} \left(\tau^2 \boldsymbol{\Lambda} + (\mathbf{X}^\mathsf{T} \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1}\right)^{-1} \hat{\boldsymbol{\beta}}, \\ \boldsymbol{\Sigma} &= (\tau^{-2} \boldsymbol{\Lambda}^{-1} + \mathbf{X}^\mathsf{T} \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1}, \end{aligned}$$

where $\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_n)$, $\tilde{\boldsymbol{\Sigma}} = \mathrm{diag}(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2)$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T} \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\mathsf{T} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{z}}$ (assuming the first inverse exists). Here ϕ denotes the possible dispersion parameter and $(\tilde{z}_i, \tilde{\sigma}_i^2)$ the location and variance for the ith Gaussian pseudo-observation. These are obtained from the first and second order derivatives of the log-likelihood terms $L_i(f_i, \phi)$ with respect to the linear predictor $f_i = \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i$ at the posterior mode $\bar{f}_i = \bar{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x}_i$ (Gelman et al., 2013, ch. 16.2)

$$\tilde{z}_i = \bar{f}_i - \frac{L_i'(\bar{f}_i, \phi)}{L_i''(\bar{f}_i, \phi)}, \quad \tilde{\sigma}_i^2 = -\frac{1}{L_i''(\bar{f}_i, \phi)}.$$

The fact that some of the observations are more informative than others – meaning $\tilde{\sigma}_i^2$ is not constant – makes further simplification somewhat difficult.

To proceed, we make the rough assumption that we can replace each $\tilde{\sigma}_i^2$ by a single variance term $\tilde{\sigma}^2$. Assuming further that the covariates are uncorrelated with zero mean and variances $\mathrm{Var}(x_j) = s_j^2$ (as in Sec. 3.3), the posterior mean for the jth coefficient satisfies $\bar{\beta}_j = (1 - \kappa_j)\hat{\beta}_j$ with shrinkage factor given by

$$\kappa_j = \frac{1}{1 + n\tilde{\sigma}^{-2}\tau^2 s_j^2 \lambda_j^2}. \tag{3.14}$$

The discussion in Section 3.3 therefore also approximately holds for the non-Gaussian observation model, except that σ^2 is replaced by $\tilde{\sigma}^2$. Still, this leaves us with the question, which value to choose for $\tilde{\sigma}^2$ to exploit this result in practice?

For the generalized linear models with y having a distribution in the exponential family with natural parameter θ and dispersion ϕ, the log likelihood for a single observation has the form (McCullagh and Nelder, 1989)

$$L = \frac{y\theta - B(\theta)}{A(\phi)} - C(y, \phi),$$

Table 1

The pseudo variances for the most commonly used generalized linear models to be used as approximate plug-in values for σ^2 in equations of Section 3.3. In practice when necessary, we usually replace μ by the sample mean $\bar{y}$. In the Gamma distribution, α denotes the shape parameter so that $\mathrm{Var}(y) = \mu^2/\alpha$. In the inverse Gaussian, λ is the shape parameter so that $\mathrm{Var}(y) = \mu^3/\lambda$. See McCullagh and Nelder (1989).

Model	Link	$\partial\mu/\partial f$	$V(\mu)$	$A(\phi)$	Pseudo variance $\tilde{\sigma}^2$
Gaussian	Identity	1	1	σ^2	σ^2
Binomial	Logit	$\mu(1-\mu)$	$\mu(1-\mu)$	1	$\mu^{-1}(1-\mu)^{-1}$
Poisson	Log	μ	μ	1	μ^{-1}
Gamma	Inverse	$-\mu^2$	μ^2	α^{-1}	$\mu^{-2}\alpha^{-1}$
Inverse Gaussian	Inverse squared	$-\mu^3/2$	μ^3	λ^{-1}	$4\mu^{-3}\lambda^{-1}$

for some specific functions $A(\cdot)$, $B(\cdot)$ and $C(\cdot)$. The pseudo variance for a given observation y is then (see Appendix B)

$$\tilde{\sigma}^2 = -\frac{1}{L''} = A(\phi)\left[\frac{1}{V(\mu)}\left(\frac{\partial\mu}{\partial f}\right)^2 - (y-\mu)\frac{\partial}{\partial f}\left(\frac{1}{V(\mu)}\frac{\partial\mu}{\partial f}\right)\right]^{-1}, \tag{3.15}$$

where $V(\mu) = B''(\theta)$ is the variance function, $\mu = \mathrm{E}(y) = B'(\theta)$ the expected value. The simplified expressions for the most commonly used generalized linear models with their canonical links are listed in Table 1.

We observe that $\tilde{\sigma}^2$ is a product between $A(\phi)$ and a term that in general depends on f, μ, and y, although for canonical links the dependence from y vanishes because the derivative of $(\frac{\partial\mu}{\partial f})/V(\mu)$ with respect to f is zero (McCullagh and Nelder, 1989, ch. 2). Thus it makes sense that for those non-Gaussian models that have a dispersion parameter ϕ (like Gamma and inverse Gaussian models), the pseudo variance and therefore also τ should scale with $A(\phi)$. For the non-constant multiplier of $A(\phi)$ in (3.15) we can use, for example, value obtained by setting μ equal to the sample mean of $\bar{y}$. This approach, although crude, seems to work reasonably well in practice. For instance, in binary classification, if we have the same number of observations from both classes, then $\mu = 0.5$, yielding $\tilde{\sigma}^2 = 4$ which was observed to give good results in our earlier study (Piironen and Vehtari, 2017a).

3.6. More complex models

Although we limit our discussion to generalized linear models, different authors have employed horseshoe prior in various other models. For instance, Faulkner and Minin (2017) consider trend filtering for modelling time series and use horseshoe as a sparsifying prior on the kth-order forward differences to express prior assumptions about the number of rapid changes in the underlying signal. As another example, Ghosh and Doshi-Velez (2017) use horseshoe prior over the weights in Bayesian neural networks to effectively turn off some of the nodes in the network.

When the model gets more complicated, one cannot simply use the reference result (3.12) to guide the hyperprior choice because it is derived assuming the linear model (2.1). Unless similar results can easily be derived for the model of interest, we recommend a pragmatic approach of drawing from the prior for different values of τ and studying the effect on the sparsity (how many coefficients fall below a certain threshold) to get an idea of a reasonable range of values. Although this strategy may seem crude, we argue that it should still be better than not thinking about the prior at all. Moreover, our results (Sec. 4.2) suggest that with a weakly informative prior such as $\tau \sim \mathrm{C}^+\big(0, \tau_0^2\big)$, having even a rough ballpark figure of the correct magnitude for τ_0 can already be a clear improvement compared to the simple $\tau \sim \mathrm{C}^+(0, 1)$.

4. Experiments

This section illustrates the benefit of the theoretical advances on synthetic and real world data. All considered models were fitted using Stan[1] (codes in the supplementary material) with the default settings unless otherwise stated, running 4 chains, 2000 samples each, first halves discarded as warmup.

4.1. Synthetic data

4.1.1. Toy example

We first illustrate the impact of the hyperprior choice $p(\tau)$ with a toy example similar to the one discussed by van der Pas, Kleijn and van der Vaart (2014). Consider model (3.3), where each y_i is generated by adding Gaussian noise with $\sigma^2 = 1$ to the corresponding signal β_i. We generated 100 data realizations with $n = 400$ and the true $\boldsymbol{\beta}_*$ having $p_* = 20$ nonzero entries equal to $A = 1, 2, \ldots, 10$ with the rest of the entries being zeros. We then computed the mean squared error (MSE) between the estimated posterior mean $\bar{\boldsymbol{\beta}}$ and the true $\boldsymbol{\beta}_*$ assuming the pure horseshoe prior with hyperpriors $\tau \sim \mathrm{C}^+(0, 1)$, $\tau \sim \mathrm{C}^+\big(0, \tau_0^2\big)$ and $\tau = \tau_0$, where τ_0 is calculated from Equation (3.13) with the oracle prior guess $p_0 = p_*$. Although the last choice reflects stronger prior information than we would typically expect to have in practice, the purpose of this setup is simply to demonstrate that one can substantially improve the inferences using our framework provided substantial prior knowledge exists. Notice though, that even when we set $\tau = \tau_0$, we do not treat τ as completely fixed, because it depends on σ which is treated as an unknown parameter with vague prior $p(\sigma^2) \propto \sigma^{-2}$.

Figure 4 shows the MSE for the three hyperpriors for different values of A. For each prior, the error is largest close to $A = \sqrt{2 \log 400} \approx 3.5$, which is called the "universal threshold" for this problem (Johnstone and Silverman, 2004; van der Pas, Kleijn and van der Vaart, 2014). Below this threshold the

[1]The experiments with the original horseshoe on the real world data are taken from Piironen and Vehtari (2017a) and were run using Stan version 2.12.0, whereas the experiments with the regularized horseshoe are run using newer version 2.15.1

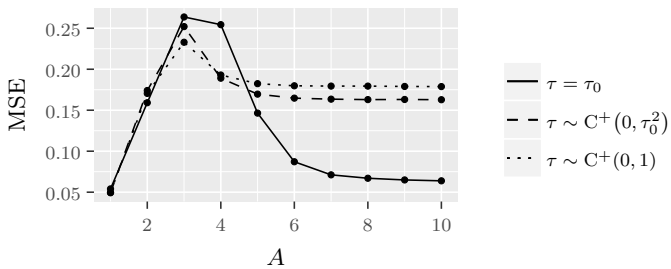

FIG 4. Toy example*: Mean squared error (MSE) between the estimated and the true coefficient vector of length $n = 400$ on average over 100 different data realizations. The true coefficient vector has $p_* = 20$ elements with a nonzero value equal to A and the rest are zeros.*

nonzero components in $\boldsymbol{\beta}$ are too small to be detected and are thus shrunk too heavily towards zero which introduces error. For $A = 4$ the informative $\tau = \tau_0$ actually yields the worst results due to this overshrinkage (see discussion below), but gives clearly superior results for larger A. The choice $\tau \sim \mathrm{C}^+\big(0, \tau_0^2\big)$ gives better results than $\tau \sim \mathrm{C}^+(0, 1)$ but is clearly inferior to $\tau = \tau_0$.

Figure 5 illustrates the data $\mathbf{y}$ and the estimated coefficients $\bar{\boldsymbol{\beta}}$ for one particular data realization when $A = 4$ and $A = 6$. In both cases the informative choice $\tau = \tau_0$ helps to shrink the zero components in $\boldsymbol{\beta}$ towards zero, but for $A = 4$ also overshrinks the nonzero components. The reason for the overshrinkage is that some observations y_i that correspond to zero signal $\beta_i = 0$ happen to have

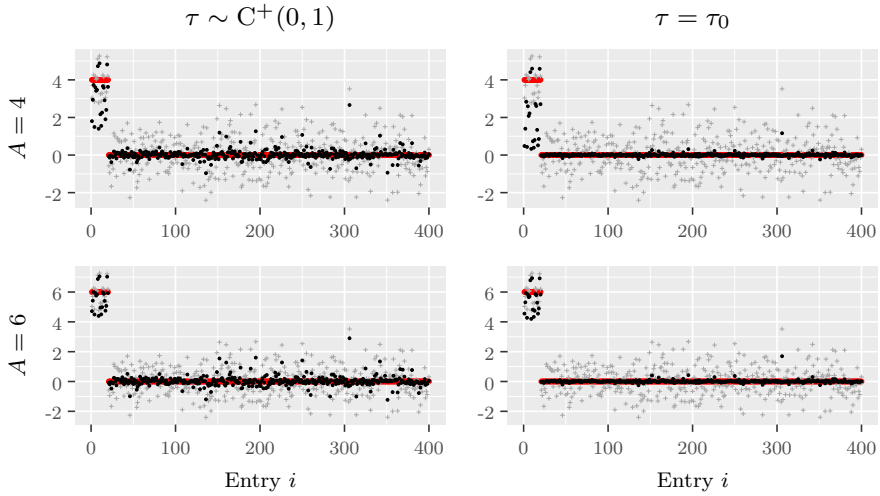

FIG 5. Toy example*: An example data realization $\mathbf{y} = (y_1, \ldots, y_n)$ (gray crosses), posterior mean $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_n)$ (black dots) and the true signal $\boldsymbol{\beta}_*$ (red lines) for $A = 4$ (top row) and $A = 6$ (bottom row). In both cases the oracle value for τ helps to shrink the zero components in $\boldsymbol{\beta}$ but also overshrinks the actual signals in the case $A = 4$.*

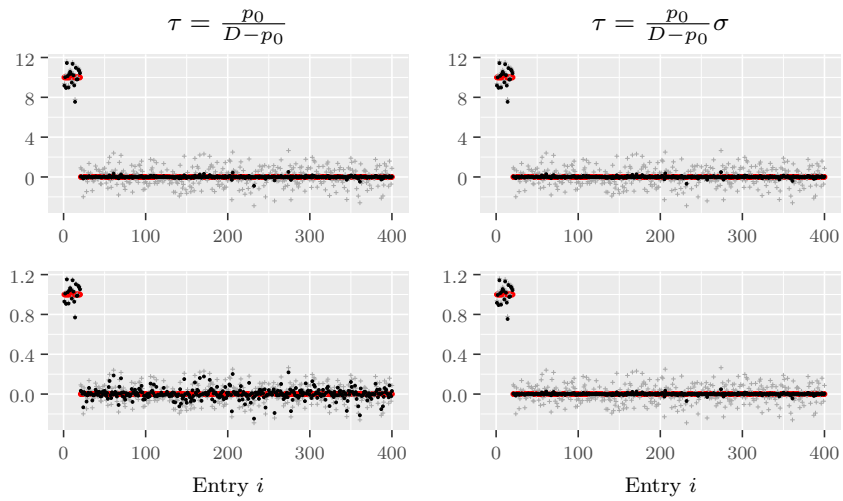

FIG 6. Toy example*: Top row: one data realization* $\mathbf{y}$ *(gray crosses), the posterior mean estimates* $\bar{\boldsymbol{\beta}}$ *(black dots) and the true signal* $\boldsymbol{\beta}_*$ *(red lines) when* $A = 10$ *and the global parameter is fixed either to* $\tau = \frac{p_0}{D-p_0}$ *or* $\tau = \frac{p_0}{D-p_0}\sigma$*. Bottom row: Otherwise the same but now the scale of the observations is changed by multiplying them by* 0.1*.*

similar magnitude to the observations coming from a nonzero signal $\beta_i = A$, and thus these irrelevant components "steal" from the limited budget for m_{eff}. For this particular value of A the overshrinkage of the actual signals happens to be worse in terms of MSE than undershrinkage of the zero components, and thus one would get better results by setting p_0 to be slightly above the true p_* (results not shown). For $A = 6$ the actual signals are large enough to be distinguished from zero, and the informative selection of τ yields substantially better estimate for $\boldsymbol{\beta}$.

Finally, Figure 6 illustrates the importance of scaling τ with the noise level σ. The top row shows one data realization and the posterior mean estimates $\bar{\boldsymbol{\beta}}$ when the global parameter is fixed either to $\tau = \frac{p_0}{D-p_0}$ or to $\tau = \frac{p_0}{D-p_0}\sigma$. For these data both yield essentially the same result, since the true noise variance σ^2 is one. However, when the observations are scaled by multiplying them by 0.1 (bottom row), the value for τ that does not scale with σ yields clearly worse results than in the first case, while the results for the latter value remain practically unchanged. What essentially happens is that when the observations are transformed to a smaller scale, then fixing $\tau = \frac{p_0}{n-p_0}$ increases the prior expectation for m_{eff} by the same factor (10 in this case) and thus it will favor solutions with many more coefficients far from zero.

This small experiment has relevance also regarding the more general model (2.1) because in the ideal case of uncorrelated predictors with unit variance $\mathbf{X}^\mathsf{T}\mathbf{X} = n\mathbf{I}$, we can think that we have a single observation of each β_j with variance σ^2/n. In practice these conditions are rarely met but the idea is still useful.

4.1.2. Classification with separable data

The purpose of this example is to illustrate the problem with the original horseshoe not controlling the magnitude of the largest coefficients and show how the regularized horseshoe can solve this issue. We generated $n = 30$ binary classification observations so that for the instances in the first class, the first two features were drawn from a Gaussian with mean 1 and scale 0.5, whereas in the other class the mean of the first two features was -1 (the data are visualized in Fig. 7). In addition, we generated 98 irrelevant features drawn from the standard Gaussian, so that the total input dimension was $D = 100$. We then fitted the standard logistic regression model with prior $\beta_0 \sim \mathrm{N}\big(0, 5^2\big)$ for the intercept and four different priors for the regression coefficients β_j: the original horseshoe, hierarchical shrinkage with $\nu = 3$, and regularized horseshoe with slab scales $c = 2$ and $c^2 \sim \text{Inv-Gamma}(2, 8)$, the latter of which corresponds to a Student-$t_4\big(0, 2^2\big)$ slab (see Sec. 2.3). In each case we used hyperprior $\tau \sim t_3^+\big(0, \tau_0^2\big)$ (the conclusions of this example are not sensitive to this choice). For consistent sparsity assumptions, for the original and the regularized horseshoe, τ_0 was calculated from (3.12) using $p_0 = 2$, and for the hierarchical shrinkage by solving numerically $\mathrm{E}(m_{\text{eff}} \,|\, \tau = \tau_0) = p_0 = 2$.

Figure 7 shows the scatter plots of the posterior draws for β_1 and β_2 (top row) as well as the medians and 80%-intervals for $\beta_3, \ldots, \beta_{100}$ (middle row), for the four different priors. Because the data are separable using only x_2, this feature is a "solitary separator" and thus the mean for β_2 does not exist under the horseshoe prior due to its Cauchy tails (Ghosh, Li and Mitra, 2017). Although for β_1 the mean exist (x_1 is not a solitary separator), the posterior has substantial mass for very large values for this parameter also. Moreover, for the original horseshoe the NUTS produces almost 200 divergent transitions after the warmup showing clear problems with sampling. Using $\nu = 3$ for the local parameters cuts down the tails and reduces the number of divergent transitions to a few, but still yields quite fat posterior tails for these two coefficients and results in much less shrinkage for the coefficients of the irrelevant features (middle row). Finally, the regularized horseshoe exhibits the most satisfactory performance cutting down the tails for β_1 and β_2 while still being able to shrink the irrelevant coefficients as well as the horseshoe. In this case, fixing the scale of the Gaussian slab to $c = 2$ results in too strong regularization for the relevant coefficients and generally we would prefer a less informative choice such as $c \sim \text{Inv-Gamma}(2, 8)$, but our purpose here was to demonstrate how easy it is to specify different levels of regularization for the largest coefficients using the new prior (2.8).

The top row of Figure 7 clearly reveals the multimodality of the posterior for β_1 and β_2. This does not produce marked problems in this case (e.g. MCMC convergence problems) but in general the multimodality can be an issue both for the original and the regularized horseshoe, and we will discuss this further in Sections 4.2 and 6.

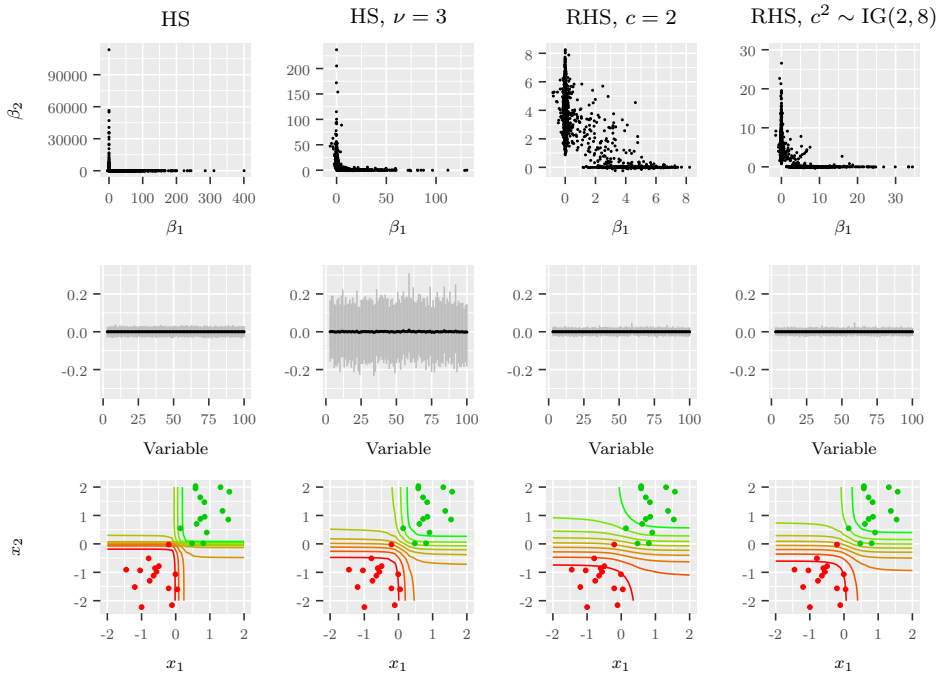

FIG 7. Separable classification. *Top row: draws for the regression coefficients of the two relevant features. Middle row: median and 80%-interval for the coefficients of the irrelevant features. Bottom row: the observed data (red and green denoting the different classes) and the predictive class probabilities for $\tilde{y} = 1$ as a function of the first two inputs, given that the rest of the inputs are set to zero. First two columns denote the results for the pure horseshoe and hierarchical shrinkage, and the last two columns for the regularized horseshoe.*

4.2. Real world data – microarray cancer classification

This section further illustrates the important concepts with some real world examples. We use the four microarray cancer classification datasets from our earlier paper (Piironen and Vehtari, 2017a). The datasets are summarized in Table 2 and can be found online.[2]

For all the datasets we used the standard logistic regression model with a vague prior $\beta_0 \sim \mathrm{N}\big(0, 10^2\big)$ for the intercept, and the original and regularized horseshoe priors for the regression coefficients to compare the differences between the two. For these problems, we reduced the number of draws per chain to 1000 to reduce the computation time.

We first consider the Ovarian dataset as a representative example of how the prior choice $p(\tau)$ and the use of the regularized horseshoe can affect the results. We fitted the model to the data with two hyperprior choices, $\tau \sim \mathrm{C}^+(0, 1)$ and $\tau \sim \mathrm{C}^+\big(0, \tau_0^2\big)$, where τ_0 is computed from (3.12) using $p_0 = 3$ as our prior

[2]Colon, Prostate and Leukemia: http://featureselection.asu.edu/datasets.php; Ovarian data: request from the first author if needed.

TABLE 2
Summary of the real world microarray cancer datasets; number of predictor variables D and dataset size n.

Dataset	Type	D	n
Ovarian	Binary classif.	1536	54
Colon	Binary classif.	2000	62
Prostate	Binary classif.	5966	102
Leukemia (ALL-AML)	Binary classif.	7129	72

guess for the number of relevant variables. For the regularized horseshoe we used hyperprior $c^2 \sim \text{Inv-Gamma}(2, 8)$ (as with the synthetic example, Sec. 4.1.2) which corresponds to a Student-$t_4\big(0, 2^2\big)$ slab (see Sec. 2.3).

Figure 8 shows prior and posterior draws for τ and m_{eff}, and the absolute values of the posterior means for the regression coefficients, for the different prior configurations. The results for $\tau \sim \text{C}^+(0, 1)$ illustrate how weakly τ is identified by the data: there is very little difference between the prior and posterior samples for τ and consequently for m_{eff}, and thus this "non-informative"

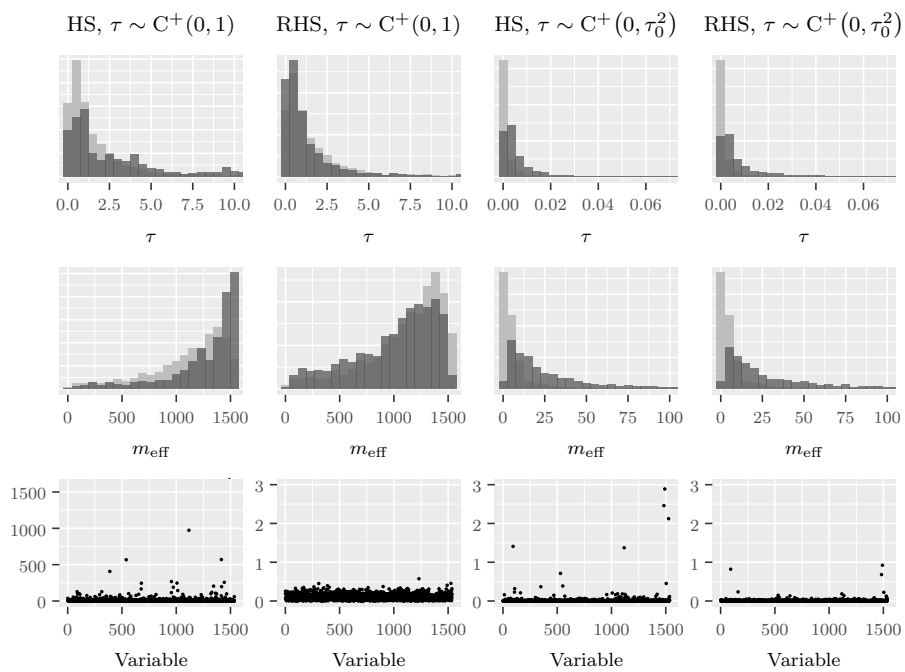

FIG 8. Ovarian dataset *: Histograms of prior (light gray) and posterior (dark gray) draws for τ (top row) and m_{eff} (middle row), and absolute values of the posterior means for the regression coefficients $|\bar{\beta}_j|$ (bottom row) imposed by different prior choices. The first two columns denote the results for the pure and regularized horseshoe with hyperprior $\tau \sim \text{C}^+(0, 1)$, and the last two columns the same but with $\tau \sim \text{C}^+\big(0, \tau_0^2\big)$, where τ_0 is computed from (3.12) with prior guess $p_0 = 3$.*

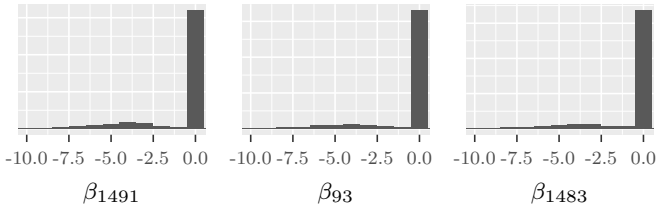

FIG 9. Ovarian dataset*: Histograms of the posterior draws under the regularized horseshoe prior for the three regression coefficients with the largest absolute mean. The histograms highlight the multimodality of the posterior.*

prior actually has a strong influence on the posterior. With the pure horseshoe this results in severe under-regularization and implausibly large magnitude for the regression coefficients. This happens because the effective model complexity is so large that the classes become separable, and therefore the coefficients are only weakly identified (as in the synthetic example in Sec. 4.1.2). With the regularized horseshoe, even with the complete separation, the magnitude of the coefficients remains sensible due to the regularization by the slab.

Replacing the scale of the half-Cauchy hyperprior with τ_0, reflecting a more sensible guess for the number of relevant variables, has a substantial effect on the posterior: the posterior mass for m_{eff} becomes concentrated on much smaller values and the magnitude of the regression coefficients more sensible. For this hyperprior choice the difference between the original and regularized horseshoe is much less severe, but also in this case the largest coefficients are smaller for the regularized horseshoe. How this affects the predictive accuracy will be discussed in a moment.

A potential explanation for why τ and therefore m_{eff} are not strongly identified here is that there are a lot of correlations in the data. For instance, the predictor $j = 1491$ which appears relevant based on its regression coefficient, has an absolute correlation of at least $|\rho| = 0.5$ with 314 other variables, out of which 65 correlations exceed $|\rho| = 0.7$. This indicates that there are a lot of potentially relevant but redundant predictors in the data, and thus similar fit could be obtained by models with varying levels of sparsity.

Another difficulty with correlating predictors is that they cause the posterior of the regression coefficients to become multimodal, which usually results in convergence issues. Figure 9 shows the posterior draws under the regularized horseshoe prior for the three coefficients with the largest absolute mean. All three coefficients have most of the draws near zero but some of the draws are far from zero, illustrating the multimodality. The $\hat{R}$-values (potential scale reduction factor, see e.g., Gelman et al., 2013, ch. 11) for these parameters were 1.09, 1.20 and 1.05 indicating problems with the convergence of the MCMC chains. Although these convergence issues do not seem to have a drastic negative impact on the predictive accuracy (see the discussion below), the multimodality dramatically reduces the sampling efficiency and is therefore a major practical

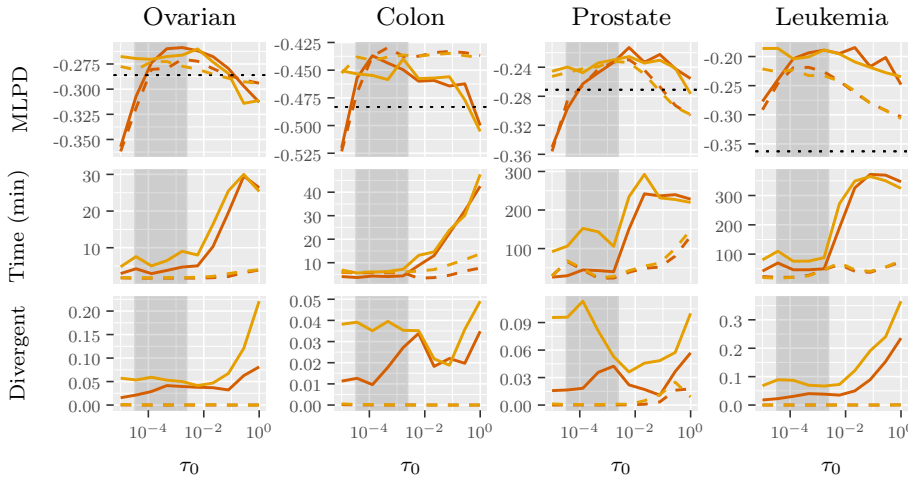

FIG 10. Microarray datasets: *Mean log predictive density (MLPD) on test data (top row), computation time (middle row), and the fraction of divergent transitions after warmup (bottom row) as a function of τ_0 for two hyperpriors: $\tau \sim \mathrm{N}^+\big(0, \tau_0^2\big)$ (red), and $\tau \sim \mathrm{C}^+\big(0, \tau_0^2\big)$ (yellow). Solid lines denote the pure horseshoe and dashed lines the regularized horseshoe with $c^2 \sim \text{Inv-Gamma}(2, 8)$. The shaded area denotes the values for τ_0 that correspond to sparsity guesses between $p_0 = 1$ and $p_0 = n$ (Eq. (3.12)). The black dotted line in the top row plots shows the MLPD for Lasso. All the results are averaged over 50 random splits into training and test sets.*

and theoretical concern and deserves more attention in the future research (see also the discussion in Section 6).

To investigate the effect of the prior choices on the prediction accuracy, we split each dataset into two halves, using one fifth of the data as a test set. All the results were then averaged over 50 such random splits into training and test sets. We carried out the tests for priors $\tau \sim \mathrm{C}^+\big(0, \tau_0^2\big)$ and $\tau \sim \mathrm{N}^+\big(0, \tau_0^2\big)$ with various τ_0. The half-normal prior was included in the tests to investigate whether it sometimes could be beneficial to have a strong control over τ by using a short-tailed prior. To get a baseline for the comparisons, we also computed the prediction accuracies to Lasso with the regularization parameter tuned by 10-fold cross-validation. The Lasso results were computed with the default settings of the R-package `glmnet` (Friedman, Hastie and Tibshirani, 2010).

The top row of Figure 10 shows the effect of the prior choice on the test prediction accuracy, and the other two rows the computation time[3] and the fraction of divergent transitions produced by the NUTS after the warm-up. For the original horseshoe (solid lines) the results illustrate a clear benefit from using even a crude prior guess for the number of relevant variables p_0: transforming any guess between $p_0 = 1$ and $p_0 = n$ into a value of τ_0 using Equation (3.12) (shaded region) and using this as the scale for the half-Cauchy prior instead of $\tau_0 = 1$ yields improved prediction accuracy and reduced computation time

[3]Wall time when the four chains are run in parallel using four cores.

in all the datasets. The regularized horseshoe seems in general less sensitive to the prior choice for τ (Ovarian and Colon datasets), but in some cases clearly benefits from a more carefully chosen prior (Prostate and Leukemia datasets). For both priors, using half-Cauchy hyperprior for τ is clearly less sensitive to the prior guess τ_0, and yields better results than the half-normal especially when τ_0 was chosen to be too small.

In terms of the predictive accuracy, the regularized horseshoe performs overall comparably to the original horseshoe: for the Colon dataset it performs slightly better, for Leukemia slightly worse, and for Ovarian and Prostate about the same as the pure horseshoe. Only for the Colon and Leukemia datasets the difference is statistically significant (the errorbars are left out from the plot to avoid mess). In Leukemia dataset where the regularized horseshoe loses to the pure horseshoe the chosen prior for the slab width c is probably unnecessarily restrictive leading to too large regularization for the largest coefficients. It is evident that in this case using a looser prior for c could improve the results (because when $c \to \infty$ we recover the original horseshoe), but we feel that these results are enough to convince the reader that a reasonably chosen hyperprior $p(c)$ does not necessarily compromise and in some cases can even improve the accuracy compared to the pure horseshoe.

On the other hand, the regularized horseshoe clearly improves the sampling robustness of the posterior. The regularized horseshoe produces only very few divergent transitions after the warm-up if any (the fraction is non-negligible only for the Prostate dataset with a poorly chosen global prior $p(\tau)$), whereas for the original horseshoe the fraction varies between 1–30% which is really a lot. With this many divergences, there is always a concern about a biased inference and this cannot be taken too lightly (see the next Section for recommendations). Moreover, also the computation times are systematically either smaller or similar to the original horseshoe, which is due to better behaving posterior.

For any reasonably selected prior (shaded region), the (regularized) horseshoe consistently outperforms Lasso in terms of predictive accuracy, but in some cases the difference is not very large. A clear advantage for Lasso, on the other hand, is that it is hugely faster, with computation time of less than a second for these problems, whereas even for the regularized horseshoe, the computation starts to get quite involved for the two biggest problems (around 30–60 minutes for Prostate and Leukemia datasets), which is the price for the full Bayesian inference.

5. Recommendations

Based on the theoretical considerations and the experimental results, instead of the original horseshoe, we recommend using the regularized horseshoe with a weakly informative prior on c, such as (2.11) with appropriately chosen scale s and degrees of freedom ν. As illustrated in Section 4, this does not compromise the predictive accuracy but greatly improves the sampling robustness of the prior and is resistant to problems originating from weak identifiability of the

parameters. Still, it is worthwhile to compare the results to a model with a relatively loose prior on c using standard model assessment techniques (Vehtari and Ojanen, 2012; Vehtari, Gelman and Gabry, 2017) to get an indication if the slab width was chosen to be too restrictive. However, even if a very large value for c yields better predictive fit, if it produces a large number of divergent transitions, it cannot be recommended because in such cases the inference is likely to have a bias with an unknown magnitude (Betancourt, 2017a). In our experience, for the regularized horseshoe it is often possible to get rid off the divergences by tuning Stan's adaptation routine as explained by Betancourt (2017b).

Also, instead of using the simple $\tau \sim \mathrm{C}^+(0,1)$, for linear regression we generally recommend a weakly informative default choice $\tau \,|\, \sigma \sim \mathrm{C}^+\big(0, \tau_0^2\big)$, where τ_0 is computed from (3.12) using the prior guess p_0 for the number of relevant variables. For the other generalized linear models (GLMs), an approximately equivalent choice is obtained by replacing σ with a appropriate plug-in value as explained in Section. 3.5. Based on the results on the real world data, this choice seems to perform well unless p_0 is chosen to be much too large. However, the toy example shows that sometimes even better results can be obtained by more informative prior.

6. Discussion

This paper has discussed the use of the horseshoe prior for sparse Bayesian generalized linear models. We considered two methodological advances. First, we proposed a generalization of the horseshoe – called the regularized horseshoe – that operates otherwise similarly as the horseshoe but allows specifying the regularization to the coefficients that are far from zero. Second, we introduced a new concept – effective number of nonzero parameters – which is useful for guiding the hyperprior choice for the global shrinkage parameter.

The experiments demonstrated the benefit of both of these approaches. The ability to regularize those parameters that are far from zero is useful especially when the parameters are only weakly identified by the data. As an example we discussed the logistic regression with data separation. Adding a small regularization to the largest coefficients ensures that the posterior mean will exist, leading to more reasonable parameter estimates and faster posterior exploration. The regularized horseshoe solves also the problems with the divergent transitions that have previously been an indication about problems in posterior simulation and possibly biased inference (Betancourt, 2017b).

Regarding the hyperprior choice for the global shrinkage parameter, we argued that the previous default choices are often dubious based on their tendency to favor solutions with too many parameters unshrunk. The experiments show that for many datasets, one can obtain clear improvements – in terms of better parameter estimates, prediction accuracy and faster computation – by coming up even with a crude guess for the number of relevant variables and transforming this knowledge into a prior for τ using our proposed framework. Based on our results, for a reasonably selected global hyperprior, the (regularized) horseshoe

outperforms Lasso in terms of predictive accuracy. A notable difference is that Lasso produces a truly sparse solution with exact zeros for some coefficients, whereas horseshoe does not. After fitting the full model, a truly sparse solution without losing predictive accuracy could be obtained using the projective variable selection (Piironen and Vehtari, 2017b)[4].

Although these advances improve the overall performance and applicability of the horseshoe prior to various problems, some challenges still persist. When using the horseshoe prior with correlating predictors, the major concern is always the multimodality of the posterior, which can lead to difficulties in sampling and especially to slow convergence of the MCMC. The experiments indicated that the regularized horseshoe can improve the sampling robustness but does not remove the multimodality. It must be noted though, that the multimodality is a direct consequence of the sparsifying prior assumption that favors solutions where only one in a group of correlating predictors would have a nonzero coefficient. Whether this assumption is reasonable in practice is a more fundamental question. It could make sense – both computationally and theoretically – to employ a horseshoe or other sparsifying prior on some transformed set of features – for instance principal components of the original predictors – instead of the original variables themselves. This would yield a unimodal posterior but the effects on the other aspects such as the predictive accuracy remain to be explored. We leave these ideas for future investigation.

Appendix A: Derivation of the regularized horseshoe

The regularized horseshoe from Section 2.3 can be derived by designing a new prior with shrinkage profile otherwise similar to the horseshoe, but so that instead of favoring $\kappa_j = 0$ and $\kappa_j = 1$, the prior favors $\kappa_j = b_j$ and $\kappa_j = 1$. This is achieved by defining the shrinkage factor for the new prior as

$$\tilde{\kappa}_j = (1 - b_j)\kappa_j + b_j, \tag{A.1}$$

where κ_j is the shrinkage factor for the original horseshoe. Now $\tilde{\kappa}_j \to 1$ when $\kappa_j \to 1$, but $\tilde{\kappa}_j \to b_j$ when $\kappa_j \to 0$, so we shift the shrinkage profile of the horseshoe from interval $(0, 1)$ to $(b_j, 1)$. By plugging in the expression for the shrinkage factor $\kappa_j = 1/(1+a_j^2\lambda_j^2)$ where $a_j^2 = n\sigma^{-2}\tau^2\, s_j^2$, with a straightforward manipulation we get

$$\tilde{\kappa}_j = \frac{1 + b_j\, a_j^2\lambda_j^2}{1 + a_j^2\lambda_j^2}. \tag{A.2}$$

We want to write this in the form $\tilde{\kappa}_j = 1/(1 + a_j^2\tilde{\lambda}_j^2)$. Solving $\tilde{\lambda}_j^2$ from this equation yields

$$\tilde{\lambda}_j^2 = \frac{(1 - b_j)\lambda_j^2}{1 + b_j\, a_j^2\lambda_j^2}. \tag{A.3}$$

[4]Code available at https://github.com/stan-dev/projpred.

For convenience, we require that $b_j = \frac{1}{1+n\sigma^{-2}s_j^2c^2}$ which corresponds to shrinkage by Gaussian with variance c^2. By plugging this into (A.3), after a few lines of straightforward algebra we are left with

$$\tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{\frac{\sigma^2}{ns_j^2} + c^2 + \tau^2\lambda_j^2}. \tag{A.4}$$

Thus we have shown that the prior

$$\beta_j \mid \lambda_j, \tau \sim \mathrm{N}\Big(0, \tau^2\tilde{\lambda}_j^2\Big), \quad \lambda_j \sim \mathrm{C}^+(0,1), \tag{A.5}$$

with $\tilde{\lambda}_j^2$ defined by (A.4) has the shrinkage profile of the horseshoe shifted from the interval $(0,1)$ to $(b_j,1)$, where $b_j = \frac{1}{1+n\sigma^{-2}s_j^2c^2}$. As discussed in Section 2.3, the term $\frac{\sigma^2}{ns_j^2}$ is typically small compared to c^2 so by leaving this out from (A.5), we get the prior (2.8).

Appendix B: Pseudo variance for Non-Gaussian observations

As discussed in Section 3.5, assuming y has a distribution in the exponential family with natural parameter θ and dispersion ϕ, the log likelihood has the form

$$L = \frac{y\theta - B(\theta)}{A(\phi)} - C(y_i, \phi),$$

for some specific functions $A(\cdot)$, $B(\cdot)$ and $C(\cdot)$. From the well-known relations

$$\mathrm{E}\left(\frac{\partial L}{\partial \theta}\right) = 0, \quad \text{and} \quad \mathrm{E}\left(\frac{\partial^2 L}{\partial \theta^2}\right) + \mathrm{E}\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right) = 0,$$

we can easily derive

$$\begin{aligned} \mathrm{E}(y) &= B'(\theta), \\ \mathrm{Var}(y) &= B''(\theta)A(\phi). \end{aligned}$$

We denote $\mu = B'(\theta)$ and $V(\mu) = B''(\theta)$, from which we can also deduce $\frac{\partial \mu}{\partial \theta} = V(\mu)$. The first derivative is given by the chain rule

$$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial f} = \frac{y-\mu}{A(\phi)}\frac{1}{V(\mu)}\frac{\partial \mu}{\partial f}.$$

The second derivative is then

$$\frac{\partial^2 L}{\partial f^2} = \frac{1}{A(\phi)}\left[(y-\mu)\frac{\partial}{\partial f}\left(\frac{1}{V(\mu)}\frac{\partial \mu}{\partial f}\right) - \frac{1}{V(\mu)}\left(\frac{\partial \mu}{\partial f}\right)^2\right].$$

Thus the pseudo variance becomes

$$\tilde{\sigma}^2 = -\left(\frac{\partial^2 L}{\partial f^2}\right)^{-1} = A(\phi)\left[\frac{1}{V(\mu)}\left(\frac{\partial \mu}{\partial f}\right)^2 - (y-\mu)\frac{\partial}{\partial f}\left(\frac{1}{V(\mu)}\frac{\partial \mu}{\partial f}\right)\right]^{-1}.$$

Appendix C: Stan codes

C.1. Simple parametrization

The following shows the Stan code for the linear Gaussian model with the regularized horseshoe prior using a straightforward parametrization. In our experience this code works fine, but in Appendix C.2 we also provide another code using different parametrization (with which we generated our results). This is worth trying if the simple code has issues with sampling (produces divergences).

In the code, both τ and λ_j are given half-t priors with the degrees of freedom and the scale defined by the user (the scale can be adjusted only for τ). Setting `nu_local` $= 1$ corresponds to the horseshoe. `nu_global` $= 1$ gives τ a half-Cauchy prior. The scale for τ is `scale_global*sigma`, so if we want to set this to be $\tau_0 = \frac{p_0}{D-p_0}\frac{\sigma}{\sqrt{n}}$ (Eq. (3.12)), we should set `scale_global` $= \frac{p_0}{(D-p_0)\sqrt{n}}$. The code assumes a Student-t slab for regularizing the largest coefficients, and the scale and degrees of freedom can be specified using `slab_scale` and `slab_df` arguments.

```
data {
  int<lower=0> n;              # number of observations
  int<lower=0> d;              # number of predictors
  vector[n] y;                 # outputs
  matrix[n,d] x;               # inputs
  real<lower=0> scale_icept;   # prior std for the intercept
  real<lower=0> scale_global;  # scale for the half-t prior for tau
  real<lower=1> nu_global;     # degrees of freedom for the half-t prior
                               # for tau
  real<lower=1> nu_local;      # degrees of freedom for the half-t priors
                               # for lambdas
  real<lower=0> slab_scale;    # slab scale for the regularized horseshoe
  real<lower=0> slab_df;       # slab degrees of freedom for the regularized
                               # horseshoe
}

parameters {
  real logsigma;
  real beta0;
  vector[d] z;
  real<lower=0> tau;           # global shrinkage parameter
  vector<lower=0>[d] lambda;   # local shrinkage parameter
  real<lower=0> caux;
}

transformed parameters {
  real<lower=0> sigma;                # noise std
  vector<lower=0>[d] lambda_tilde;    # 'truncated' local shrinkage parameter
  real<lower=0> c;                    # slab scale
  vector[d] beta;                     # regression coefficients
  vector[n] f;                        # latent function values
  sigma = exp(logsigma);
  c = slab_scale * sqrt(caux);
  lambda_tilde = sqrt( c^2 * square(lambda) ./ (c^2 + tau^2*square(lambda)) );
  beta = z .* lambda_tilde*tau;
  f = beta0 + x*beta;
}
```

```
model {
  # half-t priors for lambdas and tau, and inverse-gamma for c^2
  z ~ normal(0, 1);
  lambda ~ student_t(nu_local, 0, 1);
  tau ~ student_t(nu_global, 0, scale_global*sigma);
  caux ~ inv_gamma(0.5*slab_df, 0.5*slab_df);
  beta0 ~ normal(0, scale_icept);
  y ~ normal(f, sigma);
}
```

The code for the logistic regression model is very similar, we simply remove the lines related to the noise deviation `sigma`, and change the observation model and the type of the target variable data `y`. Notice also that now the scale for τ is simply `scale_global`. Thus, to follow our recommendation, we set `scale_global` $= \tau_0 = \frac{p_0}{D-p_0}\frac{\sigma}{\sqrt{n}}$ (Eq. (3.12)), by using a suitable plug-in value for σ (Sec. 3.5). The lines that need to be changed (in addition to removing definitions related to `sigma`) are shown below.

```
data {
  ...
  int<lower=0,upper=1> y[n];  # outputs
  ...
}

transformed parameters {
  ...
  tau = aux1_global * sqrt(aux2_global) * scale_global;
  ...
}

model {
  ...
  y ~ bernoulli_logit(f);
}
```

C.2. Alternative parametrization

This parameterization was proposed by Peltola et al. (2014) (codes at https://github.com/to-mi/stan-survival-shrinkage). In practice we have not observed problems with the code presented in Appendix C.1 but if it has issues with sampling, it is worth trying the following code (using which we ran our experiments). Below is the code for the Gaussian observation model.

```
data {
  int<lower=0> n;              # number of observations
  int<lower=0> d;              # number of predictors
  vector[n] y;                 # outputs
  matrix[n,d] x;               # inputs
  real<lower=0> scale_icept;   # prior std for the intercept
  real<lower=0> scale_global;  # scale for the half-t prior for tau
  real<lower=1> nu_global;     # degrees of freedom for the half-t prior
                               # for tau
  real<lower=1> nu_local;      # degrees of freedom for the half-t priors
                               # for lambdas
  real<lower=0> slab_scale;    # slab scale for the regularized horseshoe
  real<lower=0> slab_df;       # slab degrees of freedom for the regularized
                               # horseshoe
}
```

```
parameters {
  real logsigma;
  real beta0;
  vector[d] z;
  real<lower=0> aux1_global;
  real<lower=0> aux2_global;
  vector<lower=0>[d] aux1_local;
  vector<lower=0>[d] aux2_local;
  real<lower=0> caux;
}

transformed parameters {
  real<lower=0> sigma;              # noise std
  real<lower=0> tau;                # global shrinkage parameter
  vector<lower=0>[d] lambda;        # local shrinkage parameter
  vector<lower=0>[d] lambda_tilde;  # 'truncated' local shrinkage parameter
  real<lower=0> c;                  # slab scale
  vector[d] beta;                   # regression coefficients
  vector[n] f;                      # latent function values
  sigma = exp(logsigma);
  lambda = aux1_local .* sqrt(aux2_local);
  tau = aux1_global * sqrt(aux2_global) * scale_global*sigma;
  c = slab_scale * sqrt(caux);
  lambda_tilde = sqrt( c^2 * square(lambda) ./ (c^2 + tau^2*square(lambda)) );
  beta = z .* lambda_tilde*tau;
  f = beta0 + x*beta;
}

model {
  # half-t priors for lambdas and tau, and inverse-gamma for c^2
  z ~ normal(0, 1);
  aux1_local ~ normal(0, 1);
  aux2_local ~ inv_gamma(0.5*nu_local, 0.5*nu_local);
  aux1_global ~ normal(0, 1);
  aux2_global ~ inv_gamma(0.5*nu_global, 0.5*nu_global);
  caux ~ inv_gamma(0.5*slab_df, 0.5*slab_df);
  beta0 ~ normal(0, scale_icept);
  y ~ normal(f, sigma);
}
```

Again, the code for the logistic regression model is very similar. Below are the lines that need to be changed (in addition to removing the definitions related to `sigma`).

```
data {
  ...
  int<lower=0,upper=1> y[n];  # outputs
  ...
}

transformed parameters {
  ...
  tau = aux1_global * sqrt(aux2_global) * scale_global;
  ...
}

model {
  ...
  y ~ bernoulli_logit(f);
}
```

Appendix D: rstanarm code

The horseshoe prior is implemented in the R-package **rstanarm**[5], which contains a lot of precompiled Stan code and provides an easy-to-use interface to the most commonly used regression models. At the time of writing this, the package does not implement the regularized horseshoe but it will be added in the near future.

Assuming the predictor matrix and the corresponding targets are loaded into variables x and y, the linear Gaussian model with the horseshoe prior can be fitted as follows:

```
# use all available cores
options(mc.cores = parallel::detectCores())

# set up the prior, use hyperprior tau ~ half-Cauchy(0,tau0^2)
D <- ncol(x)
n <- nrow(x)
p0 <- 5 # prior guess for the number of relevant variables
tau0 <- p0/(D-p0) / sqrt(n) # rstanarm will scale this by sigma automatically
prior_coeff <- hs(df=1, global_df=1, global_scale=tau0)

# fit the model
fit <- stan_glm(y ~ x, family = gaussian(), data = data.frame(I(x),y),
                prior = prior_coeff)
```

The other generalized linear models can be fitted in a similar manner. Here is an example of fitting a logistic regression model (assume now that values y are binary):

```
# set up the prior, use hyperprior tau ~ half-Cauchy(0,tau0^2)
D <- ncol(x)
n <- nrow(x)
p0 <- 5 # prior guess for the number of relevant variables
sigma <- 1 / sqrt(mean(y)*(1-mean(y))) # pseudo sigma
tau0 <- p0/(D-p0) * sigma/sqrt(n)
prior_coeff <- hs(df=1, global_df=1, global_scale=tau0)

# fit the model
fit <- stan_glm(y ~ x, family = binomial(), data = data.frame(I(x),y),
                prior = prior_coeff)
```

Acknowledgments

We thank Andrew Gelman, Michael Betancourt and Daniel Simpson for helpful comments about the manuscript. We also acknowledge the computational resources provided by the Aalto Science-IT project.

References

Betancourt, M. (2017a). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

[5] https://github.com/stan-dev/rstanarm

Betancourt, M. (2017b). Diagnosing biased inference with divergences. Case study notebook. http://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html, accessed 15.6.2017.

Betancourt, M. and Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In *Current trends in Bayesian methodology with applications* (S. K. Upadhyay, U. Singh, D. K. Dey and A. Loganathan, eds.) 79–101. Chapman & Hall.

Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*. First Online, DOI: 10.1214/16-BA1028. MR3449048

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (D. van Dyk and M. Welling, eds.). *Proceedings of Machine Learning Research* **5** 73–80. PMLR. MR2650751

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR3036256

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis* **8** 111–132.

Faulkner, J. R. and Minin, V. N. (2017). Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis*. First Online, DOI: 10.1214/17-BA1050.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**. MR2221284

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1** 515–533. MR3235677

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, Third ed. Chapman & Hall.

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

Ghosh, S. and Doshi-Velez, F. (2017). Model selection in Bayesian neural networks via horseshoe priors. *arXiv:1705.10388*.

Ghosh, J., Li, Y. and Mitra, R. (2017). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*. First Online, DOI: 10.1214/17-BA1051. MR3616141

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity*. Chapman & Hall.

Hernández-Lobato, D., Hernández-Lobato, J. M. and Suárez, A. (2010). Expectation propagation for microarray data classification. *Pattern Recognition Letters* **31** 1618–1626. MR3347601

Hernández-Lobato, J. M., Hernández-Lobato, D. and Suárez, A. (2015). Expectation propagation in linear regresssion models with spike-and-slab priors. *Machine Learning* **99** 437–487. MR1765176

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14** 382–417. MR1765176 (2001a:62033)

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593–1623. MR2089135

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* **32** 1594–1649. MR2089135

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, second ed. *Monographs on Statistics and Applied Probability*. Chapman & Hall. MR0997578

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** 1023–1036. MR997578 (90f:62217)

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103** 681–686.

Peltola, T., Havulinna, A. S., Salomaa, V. and Vehtari, A. (2014). Hierarchical Bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Bayesian Modeling Applications Workshop. CEUR Workshop Proceedings* **1218** 79–88.

Piironen, J. and Vehtari, A. (2015). Projection predictive variable selection using Stan+R. *arXiv:1508.02502*.

Piironen, J. and Vehtari, A. (2017a). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (A. Singh and J. Zhu, eds.). *Proceedings of Machine Learning Research* **54** 905–913. PMLR. MR3613594

Piironen, J. and Vehtari, A. (2017b). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* **27** 711–735. MR3204017

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 501–538. Oxford University Press, Oxford. MR3204017

Stan Development Team (2017). Stan modeling language users guide and reference manual, Version 2.15.0. `http://mc-stan.org`. MR1379242

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288.

Titsias, M. K. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems 24* 2339–2347. MR3285877

van der Pas, S. L., Kleijn, B. J. K. and van der Vaart, A. W. (2014). The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8** 2585–2618. MR3285877

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27** 1413–1432. MR3011074

Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6** 142–228.

Zhang, Y., Reich, B. J. and Bondell, H. D. (2016). High dimensional linear regression via the R2-D2 shrinkage prior. *arXiv:1609.00046*.