Bayesian Analysis (2017) **12**, Number 4, pp. 1221–1274

Uncertainty Quantification for the Horseshoe (with Discussion)

Stéphanie van der Pas[*§], Botond Szabó[†§¶], and Aad van der Vaart[‡¶]

Abstract. We investigate the credible sets and marginal credible intervals resulting from the horseshoe prior in the sparse multivariate normal means model. We do so in an adaptive setting without assuming knowledge of the sparsity level (number of signals). We consider both the hierarchical Bayes method of putting a prior on the unknown sparsity level and the empirical Bayes method with the sparsity level estimated by maximum marginal likelihood. We show that credible balls and marginal credible intervals have good frequentist coverage and optimal size if the sparsity level of the prior is set correctly. By general theory honest confidence sets cannot adapt in size to an unknown sparsity level. Accordingly the hierarchical and empirical Bayes credible sets based on the horseshoe prior are not honest over the full parameter space. We show that this is due to over-shrinkage for certain parameters and characterise the set of parameters for which credible balls and marginal credible intervals do give correct uncertainty quantification. In particular we show that the fraction of false discoveries by the marginal Bayesian procedure is controlled by a correct choice of cut-off.

AMS 2000 subject classifications: Primary 62G15; secondary 62F15.

Keywords: credible sets, horseshoe, sparsity, nearly black vectors, normal means problem, frequentist Bayes.

1 Introduction

Despite the ubiquity of problems with sparse structures, and the large amount of research effort into finding consistent and minimax optimal estimators for the underlying sparse structures Tibshirani (1996); Johnstone and Silverman (2004); Castillo and Van der Vaart (2012); Castillo et al. (2015); Jiang and Zhang (2009); Griffin and Brown (2010); Johnson and Rossell (2010); Ghosh and Chakrabarti (2015); Caron and Doucet (2008); Bhattacharya et al. (2014); Bhadra et al. (2017); Ročková (2015), the number of options for uncertainty quantification in the sparse normal means problem is very limited. In this paper, we show that the horseshoe credible sets and intervals are effective tools for uncertainty quantification, unless the underlying signals are too close to the universal threshold in a sense that is made precise in this work. We first introduce the sparse normal means problem, and our measures of quality of credible sets.

[*]Leiden University, svdpas@math.leidenuniv.nl

[†]Leiden University and Budapest University of Technology and Economics, b.t.szabo@math.leidenuniv.nl

[‡]Leiden University, avdvaart@math.leidenuniv.nl

[§]Research supported by the Netherlands Organization for Scientific Research.

[¶]The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

 DOI: 10.1214/17-BA1065

The sparse normal means problem, also known as the sequence model, is frequently studied and considered as a test case for sparsity methods, and has some applications in, for example, image processing (Johnstone and Silverman (2004)). A random vector $Y^n = (Y_1, \ldots, Y_n)$ of observations, taking values in $\mathbb{R}^n$, is modelled as the sum of fixed means and noise:

$$Y_i = \theta_{0,i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where the ε_i follow independent standard normal distributions. The sparsity assumption made on the mean vector $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,n})$ is that it is nearly black, which stipulates that most of the means are zero, except for $p_n = \sum_{i=1}^n \mathbf{1}_{\theta_{0,i} \neq 0}$ of them. The sparsity level p_n is unknown, and assumed to go to infinity as n goes to infinity, but at a slower rate than n: $p_n \to \infty$ and $p_n = o(n)$.

This paper studies the Bayesian approach based on the *horseshoe prior* Carvalho et al. (2010, 2009); Scott (2011); Polson and Scott (2012a,b). The horseshoe prior is popular due to its good performance in simulations and under theoretical study (e.g. Carvalho et al. (2010, 2009); Polson and Scott (2012a, 2010); Bhattacharya et al. (2014); Armagan et al. (2013); van der Pas et al. (2014); Datta and Ghosh (2013)). The horseshoe prior is a scale mixture of normals, with a half-Cauchy prior on the variance. It is given by

$$\begin{aligned} \theta_i \,|\, \lambda_i, \tau &\sim \mathcal{N}(0, \lambda_i^2 \tau^2), \\ \lambda_i &\sim C^+(0,1), \qquad i = 1, \ldots, n. \end{aligned} \tag{2}$$

Across i the variables are assumed independent, with the exception of the hyperparameter τ if this is given a prior as well. The "global hyperparameter" τ was determined to be important towards the minimax optimality of the horseshoe posterior mean as an estimator of θ_0 (van der Pas et al. (2014)). The results in van der Pas et al. (2014) show that τ can be interpreted as the proportion of nonzero parameters, up to a logarithmic factor. If it is set at a value of the order $(p_n/n)\sqrt{\log(n/p_n)}$, then the horseshoe posterior contracts around the true θ_0 at the (near) minimax estimation rate for quadratic loss. Adaptive posterior contraction, where the number p_n is not assumed known but estimated by empirical Bayes or hierarchical Bayes as in this paper, was proven for estimators of τ that are bounded above by $(p_n/n)\sqrt{\log(n/p_n)}$ with high probability in van der Pas et al. (2017a).

The adaptive concentration of the horseshoe posterior is encouraging towards the usefulness of the horseshoe credible balls for uncertainty quantification, as in the Bayesian framework the spread of the posterior distribution over the parameter space is used as an indication of the error in estimation. It follows from general results of Li (1989); Robins and van der Vaart (2006); Nickl and van de Geer (2013) that honest uncertainty quantification is irreconcilable with adaptation to sparsity. Here *honesty* of confidence sets $\hat{C}_n = \hat{C}_n(Y^n)$ relative to a parameter space $\tilde{\Theta} \subset \mathbb{R}^n$ means that

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \tilde{\Theta}} P_{\theta_0}(\theta_0 \in \hat{C}_n) \geq 1 - \alpha,$$

for some prescribed confidence level $1 - \alpha$. Furthermore, *adaptation* to a partition $\tilde{\Theta} = \cup_{p \in P} \Theta_p$ of the parameter space into submodels Θ_p indexed by a hyper-parameter $p \in P$,

means that, for every $p \in P$ and for $r_{n,p}$ the (near) minimax rate of estimation relative to Θ_p,

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \Theta_p} P_{\theta_0}(\operatorname{diam}(\hat{C}_n) \leq r_{n,p}) = 1.$$

This second property ensures that the good coverage is not achieved by taking conservative, overly large confidence sets, but that these sets have "optimal" diameter. In our present situation we may choose the models Θ_p equal to nearly black bodies with p nonzero coordinates, in which case $r_{n,p}^2 \asymp p\log(n/p)$, if $p \ll n$. Now it is shown in Li (1989) that confidence regions that are honest over all parameters in $\tilde{\Theta} = \mathbb{R}^n$ cannot be of square diameter smaller than $n^{1/2}$, which can be (much) bigger than $p\log(n/p)$, if $p \ll n^{1/2}$. Similar restrictions are valid for honesty over subsets of $\mathbb{R}^n$, as follows from testing arguments (see the appendix in Robins and van der Vaart (2006)). Specifically, in Nickl and van de Geer (2013) it is shown that confidence regions that adapt in size to nearly black bodies of two different dimensions $p_{n,1} \ll p_{n,2}$ cannot be honest over the union of these two bodies, but only over the union of the smallest body and the vectors in the bigger body that are at some distance from the smaller body. As both the full Bayes and empirical Bayes horseshoe posteriors contract at the near square minimax rate $r_{n,p}$, adaptively over every nearly black body, it follows that their credible balls cannot be honest in the full parameter space.

In Bayesian practice credible balls are nevertheless used as if they were confidence sets. A main contribution of the present paper is to investigate for which parameters θ_0 this practice is justified. We characterise the parameters for which the credible sets of the horseshoe posterior distribution give good coverage, and the ones for which they do not. We investigate this both for the empirical and hierarchical Bayes approaches, both when τ is set deterministically, and in adaptive settings where the number of nonzero means is unknown. In the case of deterministically chosen τ, uncertainty quantification is essentially correct provided τ is chosen not smaller than $(p_n/n)\sqrt{\log(n/p_n)}$. For the more interesting full and empirical Bayes approaches, the correctness depends on the sizes of the nonzero coordinates in θ_0. If a fraction of the nonzero coordinates is detectable, meaning that they exceed the "threshold" $\sqrt{2\log(n/p_n)}$, then uncertainty quantification by a credible ball is correct up to a multiplicative factor in the radius. More generally, this is true if the sum of squares of the non-detectable nonzero coordinates is suitably dominated, as in Belitser and Nurushev (2015).

We show in this work that the uncertainty quantification given by the horseshoe posterior distribution is "honest" only under certain prior assumptions on the parameters. In contrast, interesting recent work within the context of the sparse linear regression model is directed at obtaining confidence sets that are honest in the full parameter set Zhang and Zhang (2014); van de Geer et al. (2014); Liu and Yu (2013). The resulting methodology, appropriately referred to as "de-sparsification", might in our present very special case of the regression model reduce to confidence sets for θ_0 based on the trivial pivot $Y^n - \theta_0$, or functions thereof, such as marginals. These confidence sets would have uniformly correct coverage, but be very wide, and not accommodate the presumed sparsity of the parameter. This seems a high price to pay; sacrificing some coverage so as to retain some shrinkage may not be unreasonable. Our contribution here is to investigate in what way the horseshoe prior makes this trade-off. In addition, we provide a specific

example of an estimator that meets our conditions for adaptive coverage: the maximum marginal likelihood estimator (MMLE). The MMLE is introduced in detail in van der Pas et al. (2017a). In this paper, we expand on the MMLE results in van der Pas et al. (2017a) by showing that it meets the imposed conditions for adaptive coverage as well.

Uncertainty quantification in the case of the sparse normal means model was addressed also in the recent paper Belitser and Nurushev (2015). These authors consider a mixed Bayesian-frequentist procedure, which leads to a mixture over sets $I \subset \{1, 2, \ldots, n\}$ of projection estimators $(Y_i \mathbf{1}_{i \in I})$, where the weights over I have a Bayesian interpretation and each projection estimator comes with a distribution. Treating this as a posterior distribution, the authors obtain credible balls for the parameter, which they show to be honest over parameter vectors θ_0 that satisfy an "excessive-bias restriction". This interesting procedure has similar properties as the horseshoe posterior distribution studied in the present paper. While initially we had derived our results under a stronger "self-similarity" condition, we present here the results under a slight weakening of the "excessive-bias restriction" introduced in Belitser and Nurushev (2015).

The performance of adaptive Bayesian methods for uncertainty quantification for the estimation of functions has been previously considered in Szabó et al. (2015a,b); Serra and Krivobokova (2017); Castillo and Nickl (2014); Ray (2014); Sniekers and van der Vaart (2015a,c,b); Belitser (2017); Rousseau and Szabo (2016). These papers focus on adaptation to functions of varying regularity. This runs into similar problems of honesty of credible sets, but the ordering by regularity sets the results apart from the adaptation to sparsity in the present paper.

For single coordinates $\theta_{0,i}$ uncertainty quantification by marginal credible intervals is quite natural. Credible intervals can be easily visualised by plotting them versus the index (cf. Figure 1). A simulation study in the context of the linear regression model is given in Bhattacharya et al. (2015). Marginal credible intervals may also be used as a testing device, for instance by declaring coordinates i for which the credible interval does not contain 0 to be *discoveries*. We show that the validity of these intervals depends on the value of the true coordinate. On the positive side we show that marginal credible intervals for coordinates $\theta_{0,i}$ that are either close to zero or above the detection boundary are essentially correct. In particular, the fraction of false discoveries tends to zero. On the negative side the horseshoe posteriors shrink intervals for intermediate values too much to zero for coverage. Different from the case of credible balls, these conclusions are hardly affected by whether the sparseness level τ is set by an oracle or adaptively, based on the data.

The paper is organized as follows. Section 2 is concerned with marginal credible intervals. Consequences for the false and true discoveries are explored in Section 3. Results for credible balls are collected in Section 4. In all cases, the results are given for deterministic and general empirical and hierarchical Bayes approaches. The coverage as well as model selection properties of the marginal credible sets are investigated in a simulation study in Section 5. Section 6 contains proofs for the marginal credible intervals. A supplement (van der Pas et al., 2017b) contains the proofs of the other results, as a sequence of appendices.

1.1 Notation

The posterior distribution of θ relative to the prior (2) given fixed τ is denoted by $\Pi(\cdot \,|\, Y^n, \tau)$, and the posterior distribution in the hierarchical setup where τ has received a prior is denoted by $\Pi(\cdot \,|\, Y^n)$. We use $\Pi(\cdot \,|\, Y^n, \hat{\tau})$ for the empirical Bayes "plug-in posterior", which is $\Pi(\cdot \,|\, Y^n, \tau)$ with a data-based variable $\hat{\tau}$ substituted for τ. To emphasize that $\hat{\tau}$ is not conditioned on, we alternatively use $\Pi_\tau(\cdot \,|\, Y^n)$ for $\Pi(\cdot \,|\, Y^n, \tau)$, and $\Pi_{\hat{\tau}}(\cdot \,|\, Y^n)$ for $\Pi(\cdot \,|\, Y^n, \hat{\tau})$.

The function φ denotes the density of the standard normal distribution. The class of nearly black vectors is given by $\ell_0[p] = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n \mathbf{1}_{\theta_i \neq 0} \leq p\}$, and we abbreviate

$$\zeta_\tau = \sqrt{2\log(1/\tau)}, \qquad \tau_n(p) = (p/n)\sqrt{\log(n/p)}, \qquad \tau_n = \tau_n(p_n).$$

The cardinality of the discrete set S is denoted by $\#(S)$.

2 Credible intervals

We study the coverage properties of credible intervals for the individual coordinates $\theta_{0,i}$. We show that the marginal credible intervals fall into three categories, dependent on τ. We show that coordinates $\theta_{0,i}$ that are either "small" or "large" will be covered, in the sense that within both categories the fraction of correct intervals is arbitrarily close to 1. On the other hand, none of the "intermediate" coordinates $\theta_{0,i}$ are covered. We show this first for the deterministic case, where the boundaries between the categories are at multiples of τ and ζ_τ respectively. Furthermore, we show that the results for deterministic marginal credible intervals extend to the adaptive situation for *any* true parameter θ_0, with slight modification of the boundaries between the three cases of small, intermediate and large coordinates. We elaborate on the implications for model selection in Section 3.

2.1 Definitions

Non-adaptive marginal credible intervals can be constructed from the *marginal posterior distributions* $\Pi(\theta : \theta_i \in \cdot \,|\, Y^n, \tau)$. By the independence of the pairs (θ_i, Y_i) given τ, the ith marginal depends only on the ith observation Y_i. We consider intervals of the form

$$\hat{C}_{ni}(L, \tau) = \big\{\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq L\hat{r}_i(\alpha, \tau)\big\}, \tag{3}$$

where $\hat{\theta}_i(\tau) = \mathrm{E}(\theta_i \,|\, Y_i, \tau)$ is the marginal posterior mean, L a positive constant, and $\hat{r}_i(\alpha, \tau)$ is determined so that, for a given $0 < \alpha \leq 1/2$,

$$\Pi\big(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq \hat{r}_i(\alpha, \tau) \,|\, Y_i, \tau\big) = 1 - \alpha.$$

Adaptive empirical Bayes marginal credible intervals are defined by plugging in an estimator $\widehat{\tau}_n$ for τ in the intervals $\hat{C}_{ni}(L, \tau)$ defined by (3).

Similarly full Bayes credible intervals $\hat{C}_{ni}(L)$ are defined from the full Bayes marginal posterior distributions, centered around the coordinates of the full posterior mean $\hat{\theta} = \mathrm{E}(\theta \,|\, Y)$ as

$$\hat{C}_{ni}(L) = \big\{\theta_i : |\theta_i - \hat{\theta}_i| \leq L\hat{r}_i(\alpha)\big\}, \tag{4}$$

for $\hat{r}_i(\alpha)$ determined so that $\hat{C}_{ni}(1)$ has posterior probability $1-\alpha$.

2.2 Credible intervals for deterministic τ

The coverage of the marginal credible intervals depends crucially on the value of the true coordinate $\theta_{0,i}$. For given $\tau \to 0$, positive constants k_S, k_M, k_L and numbers $f_\tau \uparrow \infty$ as $\tau \to 0$, we distinguish three regions (small, medium and large) of signal parameters:

$$\begin{aligned}
\mathcal{S} &:= \big\{1 \leq i \leq n : |\theta_{0,i}| \leq k_S\tau\big\},\\
\mathcal{M} &:= \big\{1 \leq i \leq n : f_\tau\tau \leq |\theta_{0,i}| \leq k_M\zeta_\tau\big\},\\
\mathcal{L} &:= \big\{1 \leq i \leq n : k_L\zeta_\tau \leq |\theta_{0,i}|\big\}.
\end{aligned}$$

The conditions on the constants and f_τ in the following theorem make it that these three sets may not cover all coordinates $\theta_{0,i}$, but their boundaries are almost contiguous. The following theorem shows that the fractions of coordinates contained in $\mathcal{S}$ and in $\mathcal{L}$ that are covered by the credible intervals are close to 1, whereas no coordinate in $\mathcal{M}$ is covered.

Theorem 1. *Suppose that $k_S > 0$, $k_M < 1$, $k_L > 1$, and $f_\tau \uparrow \infty$, as $\tau \to 0$. Then for $\tau \to 0$ and any sequence $\gamma_n \to c$ for some $0 \leq c \leq 1/2$, satisfying $\zeta_{\gamma_n} \ll \zeta_\tau$,*

$$P_{\theta_0}\Big(\frac{\#\big(\{i \in \mathcal{S} : \theta_{0,i} \in \hat{C}_{ni}(L_S,\tau)\}\big)}{\#(\mathcal{S})} \geq 1-\gamma_n\Big) \to 1, \tag{5}$$

$$P_{\theta_0}\big(\theta_{0,i} \notin \hat{C}_{ni}(L,\tau)\big) \to 1, \quad \textit{for any } L > 0 \textit{ and } i \in \mathcal{M}, \tag{6}$$

$$P_{\theta_0}\Big(\frac{\#\big(\{i \in \mathcal{L} : \theta_{0,i} \in \hat{C}_{ni}(L_L,\tau)\}\big)}{\#(\mathcal{L})} \geq 1-\gamma_n\Big) \to 1, \tag{7}$$

where $L_S = (2.1/z_\alpha)[k_S + (2/\gamma_n)\zeta_{\gamma_n/2}]$ and $L_L = (1.1/z_{\alpha/2})\zeta_{\gamma_n/2}$.

Proof. See Section 6. □

Remark 1. Statements (5) and (7) concern the *fractions* of intervals that cover. Under the conditions of Theorem 1 it is also true that the individual intervals satisfy

$$P_{\theta_0}(\theta_{0,i} \in \hat{C}_{ni}(L,\tau)) \geq 1-\gamma_n,$$

with $L = L_S$ and $L = L_L$ for $i \in \mathcal{S}$ and $i \in \mathcal{L}$, respectively. This is shown as part of the proof of Theorem 1 in Section 6.

Remark 2. The results of Theorem 1 can be extended to the class of global-local scale mixtures of normals introduced in Ghosh and Chakrabarti (2015) with density $\pi(\lambda_i^2)$ given as

$$\pi(\lambda_i^2) = K\frac{1}{\lambda_i^{2+2a}}L(\lambda_i^2),$$

where $a \geq 1/2$, $K > 0$ and the function $L : (0,\infty) \mapsto (0,\infty)$ satisfies that $\sup_{t>0} L(t) \leq M$ and $\inf_{t\geq t_0} L(t) \geq c_0$ for some $c_0, t_0 > 0$. This class of priors includes the horseshoe prior, normal-exponential-gamma priors, the three parameter beta normal mixtures, the generalized double Pareto, the inverse gamma and half-t priors. The resulting constants L_S and L_L will depend on the hyper-parameters c_0, t_0, M, K and a.

2.3 Adaptive credible intervals

We show that the adaptive credible intervals mimic the behaviour of the intervals for deterministic τ given in Theorem 2. The adaptive results require some conditions on either the empirical Bayes estimator of τ, or the hyperprior on τ. In the empirical Bayes case, one condition on the estimator of τ suffices, stated below. It is the same condition under which adaptive contraction of the empirical Bayes horseshoe posterior was proven in van der Pas et al. (2017a).

Condition 1. There exists a constant $C > 0$ such that $\widehat{\tau}_n \in [1/n, C\tau_n(p_n)]$, with P_{θ_0}-probability tending to one, uniformly in $\theta_0 \in \ell_0[p_n]$.

A natural choice of estimator $\hat{\tau}_n$ is the marginal maximum likelihood estimator (MMLE), defined as

$$\widehat{\tau}_M = \operatorname*{argmax}_{\tau\in[1/n,1]} \prod_{i=1}^{n} \int_{-\infty}^{\infty} \varphi(y_i - \theta) g_\tau(\theta)\, d\theta, \tag{8}$$

where $g_\tau(\theta) = \int_0^\infty \varphi(\frac{\theta}{\lambda\tau})\frac{1}{\lambda\tau}\frac{2}{\pi(1+\lambda^2)}\, d\lambda$. It is shown in van der Pas et al. (2017a) that Condition 1 holds for the MMLE.

The restriction of the MMLE to the interval $[1/n, 1]$ corresponds to an assumption that the number of signals is between 1 and n, following the interpretation of τ as (approximately) the proportion of signals. In van der Pas et al. (2017a), and in the simulation study in Section 5, the MMLE is compared to the "simple" estimator of van der Pas et al. (2014), which estimates p_n by counting the number of observations that are larger than (a constant multiple of) the universal threshold $\sqrt{2\log n}$ and its computation is discussed. It is proven that the MMLE meets Condition 1, and thus that the empirical Bayes procedure with the MMLE as a plug-in estimate of τ leads to adaptive posterior concentration results.

In the hierarchical Bayes procedure, we impose the same conditions on the hyperprior π_n as for adaptive posterior concentration in van der Pas et al. (2017a). We recall them below.

Condition 2. The prior density π_n is supported inside $[1/n, 1]$.

Condition 3. Let $t_n = C_u\pi^{3/2}\,\tau_n(p_n)$, with the constant C_u as in Lemma G.8(i). The prior density π_n satisfies

$$\int_{t_n/2}^{t_n} \pi_n(\tau)\, d\tau \gtrsim e^{-cp_n}, \quad \text{for some } c < C_u/2.$$

Condition 3 may be replaced by the weaker Condition 4, at the price of suboptimal rates.

Condition 4. For t_n as in Condition 3 the prior density π_n satisfies,

$$\int_{t_n/2}^{t_n} \pi_n(\tau)\, d\tau \gtrsim t_n.$$

Examples of priors meeting Conditions 2 and 4 are the Cauchy prior on the positive reals, or the uniform prior, both truncated to $[1/n, 1]$. They satisfy the stronger Condition 3 if $p_n \geq C \log n$, for a sufficiently large $C > 0$.

In the adaptive case, the three regions (small, medium and large) of signal parameters are defined as, for given positive constants k_S, k_M, k_L, and f_n:

$$\begin{aligned}
\mathcal{S}_a &:= \big\{1 \leq i \leq n : |\theta_{0,i}| \leq k_S/n\big\},\\
\mathcal{M}_a &:= \big\{1 \leq i \leq n : f_n \tau_n(p_n) \leq |\theta_{0,i}| \leq k_M \sqrt{2\log(1/\tau_n(p_n))}\big\},\\
\mathcal{L}_a &:= \big\{1 \leq i \leq n : k_L \sqrt{2\log n} \leq |\theta_{0,i}|\big\}.
\end{aligned}$$

Theorem 2. *Suppose that $k_S > 0$, $k_M < 1$, $k_L > 1$, and $f_n \uparrow \infty$. If $\hat{\tau}_n$ satisfies Condition 1, then for any sequence $\gamma_n \to c$ for some $0 \leq c \leq 1/2$ such that $\zeta_{\gamma_n}^2 \ll \log(1/\tau_n(p_n))$, we have that*

$$P_{\theta_0}\Big(\frac{\#\big(\{i \in \mathcal{S}_a : \theta_{0,i} \in \hat{C}_{ni}(L_S, \hat{\tau}_n)\}\big)}{\#(\mathcal{S}_a)} \geq 1 - \gamma_n\Big) \to 1, \tag{9}$$

$$P_{\theta_0}\big(\theta_{0,i} \notin \hat{C}_{ni}(L, \hat{\tau}_n)\big) \to 1, \quad \text{for any } L > 0 \text{ and } i \in \mathcal{M}_a, \tag{10}$$

$$P_{\theta_0}\Big(\frac{\#\big(\{i \in \mathcal{L}_a : \theta_{0,i} \in \hat{C}_{ni}(L_L, \hat{\tau}_n)\}\big)}{\#(\mathcal{L}_a)} \geq 1 - \gamma_n\Big) \to 1, \tag{11}$$

with L_S and L_L given in Theorem 1. Under Conditions 2 and 3 and in addition $p_n \gtrsim \log n$ the same statements hold for the hierarchical Bayes marginal credible sets. This is also true under Conditions 2 and 4 if $f_n \gg \log n$, with different constants L_S and L_L.

Proof. See Appendix A.1 in the supplement (van der Pas et al., 2017b). □

Remark 3. Under the self-similarity assumption (15) discussed in Section 4.3, the statements of Theorem 2 hold for the sets $\mathcal{S}$, $\mathcal{M}$ and $\mathcal{L}$ given preceding Theorem 1 with $\tau = \tau_n(p_n)$.

Remark 4. Statements (9) and (11) concern the *fractions* of intervals that cover. Under the conditions of Theorem 2 it is also true that the individual intervals satisfy

$$P_{\theta_0}(\theta_{0,i} \in \hat{C}_{ni}(L, \hat{\tau}_n)) \geq 1 - \gamma_n,$$

with $L = L_S$ and $L = L_L$ for $i \in \mathcal{S}_a$ and $i \in \mathcal{L}_a$, respectively. The same statement holds for the hierarchical Bayes marginal credible intervals. This is shown as part of the proof of Theorem 2 in Appendix A.1 in the supplement (van der Pas et al., 2017b).

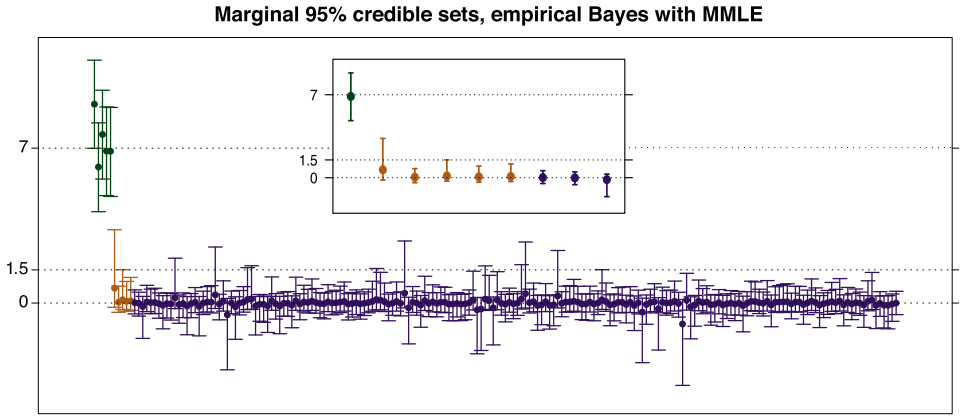


Figure 1: 95% marginal credible intervals based on the MMLE empirical Bayes method, constructed using the 2.5% and 97.5% quantiles, for a single observation Y^n of length $n = 200$ with $p_n = 10$ nonzero parameters, the first 5 (from the left) being 7 (green), the next 5 equal to 1.5 (orange); the remaining 190 parameters are coded (blue). The inserted plot zooms in on credible intervals 5 to 13, thus showing one large mean and all intermediate means.

Figure 1 illustrates Theorem 2 by showing the marginal credible sets for just a single simulated data set, in a setting with $n = 200$, and $p_n = 10$ nonzero coordinates. The value τ was chosen equal to the MMLE, which realised as approximately 0.11. The means were taken equal to 7, 1.5 or 0, corresponding to the three regions $\mathcal{L}, \mathcal{M}, \mathcal{S}$ listed in the theorem ($\sqrt{2 \log n} \approx 3.3$). All the large means (equal to 7) were covered; only 2 out of 5 of the medium means (equal to 1.5) were covered; and all small (zero) means were covered, in agreement with Theorem 2. It may be noted that intervals for zero coordinates are not necessarily narrow.

3 Model selection

Marginal credible sets give rise to a natural model selection procedure: a parameter is selected as a signal (or "is a discovery") if and only if the corresponding credible interval does not contain zero. We can study this procedure again both in the case of deterministic τ and in the adaptive case, where τ is estimated from the data or receives a hyperprior. For simplicity in this section we only state the result for the adaptive case, leaving the non-adaptive case to the supplement (van der Pas et al., 2017b), see Theorem B.1 in Appendix B .

For zero coordinates $\theta_{0,i}$ selection is the same as coverage, but for nonzero coordinates selection is easier. While coverage involves both the center and the spread of

the posterior distribution, selection depends only on the posterior probability that the signal is positive (or negative). This makes that the blow-up constant L in the definition (3) or (4) of a credible interval is unimportant. Thus we consider these intervals with an arbitrary constant $L > 0$, and say that a parameter $\theta_{0,i} = 0$ is falsely selected, or that a parameter $\theta_{0,i} \neq 0$ is correctly selected, if, in both cases, it is contained in the interval $\hat{C}_{ni}(L, \hat{\tau})$ or $\hat{C}_{ni}(L)$, in the empirical Bayes or full Bayes cases respectively. These are the false and true positives, respectively.

Now few zero parameters (the ones with index in $\mathcal{N} := \{1 \leq i \leq n : \theta_{0,i} = 0\}$) are falsely selected and most large signals (the ones with index in $\mathcal{L}_a$) are correctly selected. However most of the remaining parameters (the ones with index in $(\mathcal{N}^c \cap \mathcal{S}_a) \cup \mathcal{M}_a$) are not selected, and hence constitute false negatives. Thus the procedure is conservative, the good news being that discoveries tend to be true discoveries.

Theorem 3. *Suppose that $k_M < 1 < k_L$ and $f_n \uparrow \infty$ and let $L > 0$. For any sequence γ_n such that $\zeta^2_{\gamma_n} \ll \zeta^2_{\tau_n(p_n)}$, the following statements hold, with probability tending to one:*

(i) The number of selected parameters with $i \in \mathcal{N}$ divided by the total number $\#(\mathcal{N})$ of zero parameters is at most γ_n.

(ii) The number of selected parameters in $i \in \mathcal{L}_a$ divided by the total number of large parameters $\#(\mathcal{L}_a)$ is at least $1 - \gamma_n$, i.e.

$$P_{\theta_0}\Big(\frac{\#\big(\{i \in \mathcal{L}_a : 0 \notin \hat{C}_{ni}(L, \hat{\tau})\}\big)}{\#(\mathcal{L}_a)} \geq 1 - \gamma_n\Big) \to 1,$$

and the same for the hierarchical Bayes intervals $\hat{C}_{ni}(L)$.

(iii) At most a fraction γ_n of the parameters within $i \in (\mathcal{N}^c \cap \mathcal{S}_a) \cup \mathcal{M}_a$ will be selected.

Proof. See Appendix B in the supplement (van der Pas et al., 2017b). □

The assertion of the theorem are in the spirit of "false discovery rates". However, none of the statements concerns the usual false discovery rate, defined as the number of falsely selected parameters divided by the total number of selected parameters. Our current methods do not seem to provide realistic bounds on this quantity, partly because we are working under the assumption of sparsity.

An alternative method for model selection using the horseshoe was proposed by Carvalho et al. (2010). They proposed to select as nonzero coordinates the indices such that the ratio $\kappa_i(\tau) = \hat{\theta}_i(\tau)/Y_i$ exceeds a threshold (to be precise $\kappa_i(\tau) > 1/2$). This method has similar behaviour to the credible set based model selection approach, as proven in Theorem B.2 in Appendix B in the supplement (van der Pas et al., 2017b). We refer to Datta and Ghosh (2013) for theoretical properties of this procedure, and compare the credible interval and thresholding methods further through simulation in Section 5.

4 Credible balls

By their definition, credible sets contain a fixed fraction, e.g. 95 %, of the posterior mass. The diameter of such sets will be at most of the order of the posterior contraction rate. The upper bounds on the contraction rates of the horseshoe posterior distributions given in van der Pas et al. (2017a) imply that the horseshoe credible sets are narrow enough to be informative. However, these bounds do not guarantee that the credible sets will *cover* the truth. The latter is dependent on the *spread* of the posterior mass relative to its distance to the true parameter. For instance, the bulk of the posterior mass may be highly concentrated inside a ball of radius the contraction rate, but within a narrow area of diameter much smaller than its distance to the true parameter.

In this section we study coverage of credible balls, that is, credible sets for the full parameter vector $\theta_0 \in \mathbb{R}^n$ relative to the Euclidean distance. We do so first in the case of deterministic τ and next for the empirical and full Bayes posterior distributions.

4.1 Definitions

Given a deterministic hyperparameter τ, possibly depending on n and p_n, we consider a *credible ball* of the form

$$\hat{C}_n(L,\tau) = \big\{\theta : \|\theta - \hat{\theta}(\tau)\|_2 \leq L\hat{r}(\alpha,\tau)\big\}, \tag{12}$$

where $\hat{\theta}(\tau) = \mathrm{E}(\theta \,|\, Y^n, \tau)$ is the posterior mean, L a positive constant, and for a given $\alpha \in (0,1)$ the number $\hat{r}(\alpha,\tau)$ is determined such that

$$\Pi\big(\theta : \|\theta - \hat{\theta}(\tau)\|_2 \leq \hat{r}(\alpha,\tau) \,|\, Y^n, \tau\big) = 1-\alpha.$$

Thus $\hat{r}(\alpha,\tau)$ is the natural radius of a set of "Bayesian credible level" $1-\alpha$, and L is a constant, introduced to make up for a difference between credible and confidence levels, similarly as in Szabó et al. (2015b). Unlike in the latter paper the radii $\hat{r}(\alpha,\tau)$ do depend on the observation Y^n, as indicated by the hat in the notation.

In the empirical Bayes approach we define a credible set by plugging in an estimator $\widehat{\tau}_n$ of τ into the non-adaptive credible ball $\hat{C}_n(L,\tau)$ given in (12):

$$\hat{C}_n(L,\widehat{\tau}_n) = \big\{\theta : \|\theta - \hat{\theta}(\widehat{\tau}_n)\|_2 \leq L\hat{r}(\alpha,\widehat{\tau}_n)\big\}. \tag{13}$$

In the hierarchical Bayes case we use a ball around the full posterior mean $\hat{\theta} = \int \theta\, \Pi(d\theta \,|\, Y^n)$, given by

$$\hat{C}_n(L) = \big\{\theta : \|\theta - \hat{\theta}\|_2 \leq L\hat{r}(\alpha)\big\}, \tag{14}$$

where L is a positive constant and $\hat{r}(\alpha)$ is defined from the full posterior distribution by

$$\Pi\big(\theta : \|\theta - \hat{\theta}\|_2 \leq \hat{r}(\alpha) \,|\, Y^n\big) = 1-\alpha.$$

The question is whether these Bayesian credible sets are appropriate for uncertainty quantification from a frequentist point of view.

4.2 Credible balls for deterministic τ

The following lower bound for $\hat{r}(\alpha, \tau)$ in the case that $n\tau \to \infty$ is the key to the frequentist coverage. The assumption $n\tau/\zeta_\tau \to \infty$ is satisfied for τ of the order the "optimal" rate $\tau_n(p_n)$ provided $p_n \to \infty$ (as we assume).

Lemma 1. *If $n\tau/\zeta_\tau \to \infty$, then with P_{θ_0}-probability tending to one,*

$$\hat{r}(\alpha, \tau) \geq 0.5\sqrt{n\tau\zeta_\tau}.$$

Proof. See Appendix C.1 in the supplement (van der Pas et al., 2017b). □

Theorem 4. *If $\tau \geq \tau_n$ and $\tau \to 0$ and $p_n \to \infty$ with $p_n = o(n)$, then, there exists a large enough $L > 0$ such that*

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \ell_0[p_n]} P_{\theta_0}\big(\theta_0 \in \hat{C}_n(L, \tau)\big) \geq 1 - \alpha.$$

Proof. The probability of the complement of the event in the display is equal to $P_{\theta_0}(\|\theta_0 - \hat{\theta}(\tau)\|_2 > L\,\hat{r}(\alpha, \tau))$. In view of Lemma 1 this is bounded by $o(1)$ plus

$$P_{\theta_0}\big(\|\theta_0 - \hat{\theta}(\tau)\|_2 > 0.5L\sqrt{n\tau\zeta_\tau}\big) \lesssim \frac{\mathrm{E}_{\theta_0}\|\hat{\theta}(\tau) - \theta_0\|_2^2}{L^2 n\tau\zeta_\tau}.$$

By Theorem 3.2 of van der Pas et al. (2014) the numerator on the right is bounded by a multiple of $p_n \log(1/\tau) + n\tau\sqrt{\log 1/\tau}$. By the assumption $\tau \geq \tau_n \geq 1/n$ the quotient is smaller than α for appropriately large choice of L. □

Theorem 4 combined with the upper bound on the posterior contraction rate in van der Pas et al. (2014) show that a (slightly enlarged) credible ball centered at the posterior mean is of rate-adaptive size and covers the truth provided τ is chosen of the order of the "optimal" value $\tau_n(p_n)$. This is not possible in general, as it requires knowing the number of signals. In the next sections, we will show that if the empirical Bayes estimator of τ is "close" to $\tau_n(p_n)$, or if a hyperprior on τ places "enough" mass on a neighborhood of a quantity of order $\tau_n(p_n)$, then adaptation to the unknown number of signals is possible.

4.3 Adaptive credible balls

We now turn to credible sets in the more realistic scenario that the sparsity parameter p_n is not available. We investigate both the empirical Bayes and the hierarchical Bayes credible balls. We show that both empirical and hierarchical credible balls cover the true parameter θ_0, if θ_0 satisfies the "excessive-bias restriction", given below, under some conditions on the empirical Bayes plug-in estimate or the hierarchical Bayes hyperprior on τ.

The excessive-bias restriction

Unfortunately, coverage can be guaranteed only for a selection of true parameters θ_0. The problem is that a data-based estimate of sparsity may lead to *over-shrinkage*, due to a too small value of the plug-in estimator or concentration of the posterior distribution of τ too close to zero. Such over-shrinkage makes the credible sets too small and close to zero. A simple condition preventing over-shrinkage is that a sufficient number of nonzero parameters $\theta_{0,i}$ are above the "detection boundary". The minimum threshold for detection required in our proof is $\sqrt{2\log(n/p_n)}$. This leads to the following condition.

Assumption 1 (self-similarity)**.** A vector $\theta_0 \in \ell_0[p]$ is called *self-similar* if

$$\#\big(i : |\theta_{0,i}| \geq A\sqrt{2\log(n/p)}\big) \geq \frac{p}{C_s}. \tag{15}$$

The two constants C_s and A will be fixed to universal values, where necessarily $C_s \geq 1$ and it is required that $A > 1$.

The problem of over-shrinkage is comparable to the problem of over-smoothing in the context of nonparametric density estimation or regression, due to the choice of a too large bandwidth or smoothness level. The preceding self-similarity condition plays the same role as the assumptions of "self-similarity" or "polished tail" used by Picard and Tribouley (2000); Giné and Nickl (2010); Bull (2012); Nickl and Szabo (2016); Szabó et al. (2015b); Sniekers and van der Vaart (2015c); Rousseau and Szabo (2016) in their investigations of confidence sets in nonparametric density estimation and regression, or the "excessive-bias" restriction in Belitser (2017) employed in the context of Besov-regularity classes in the normal mean model.

The self-similarity condition is also reminiscent of the *beta-min condition* for the adaptive Lasso van de Geer et al. (2011); Bühlmann and van de Geer (2011), which imposes a lower bound on the nonzero signals in order to achieve consistent selection of the set of nonzero coordinates of θ_0. However, the present condition is different in spirit both by the size of the cut-off and by requiring only that a fraction of the nonzero means is above the threshold.

For ensuring coverage of credible balls the condition can be weakened to the following more technical condition.

Assumption 2 (excessive-bias restriction)**.** A vector $\theta_0 \in \ell_0[p]$ satisfies the *excessive-bias restriction* for constants $A > 1$ and $C_s, C > 0$, if there exists an integer $q \geq 1$ with

$$\sum_{i:|\theta_{0,i}|<A\sqrt{2\log(n/q)}} \theta_{0,i}^2 \leq Cq\log(n/q), \qquad \#\big(i : |\theta_{0,i}| \geq A\sqrt{2\log(n/q)}\big) \geq \frac{q}{C_s}. \tag{16}$$

The set of all such vectors θ_0 (for fixed constants A, C_s, C) is denoted by $\Theta[p]$, and $\tilde{p} = \tilde{p}(\theta_0)$ denotes $\#(i : |\theta_{0,i}| \geq A\sqrt{2\log(n/q)})$, for the smallest possible q.

If $\theta_0 \in \ell_0[p]$ is self-similar, then it satisfies the excessive-bias restriction with $q = p$, $C = 2A^2$ and the same constants A and C_s. This follows, because the sum in (16) is trivially bounded by $\#(i : \theta_{0,i} \neq 0)\, A^2 2\log(n/q)$.

In the following example we show that the excessive-bias restriction is also implied by a condition with the same name introduced in Belitser and Nurushev (2015). The latter condition motivated Assumption 2, which is more suited to our investigation of the horseshoe credible sets.

Example 1. For a given θ_0 and any subset $I \subset \{1, 2, \ldots, n\}$ let

$$G(I) = \sum_{i \in I^c} \theta_{0,i}^2 + 2A^2 \#(I) \log \frac{ne}{\#(I)}.$$

In Belitser and Nurushev (2015) θ_0 is defined to satisfy the *excessive-bias restriction* if G takes its minimum at a nonempty set $\tilde{I}$ such that $G(\tilde{I}) \leq C\#(\tilde{I}) \log(ne/\#(\tilde{I}))$.

We now show that in this case θ_0 also satisfies Assumption 2, with $q = \#(\tilde{I})$. Let $\theta_{0,i}$ be a coordinate with $i \in \tilde{I}$ of minimal absolute value $|\theta_{0,i}| = \min\{|\theta_{0,j}| : j \in \tilde{I}\}$. From $G(\tilde{I}) \leq G(\tilde{I} - \{i\})$ we obtain that $\theta_{0,i}^2 \geq 2A^2\#(\tilde{I}) \log(ne/\#(\tilde{I})) - 2A^2(\#(\tilde{I}) - 1) \log(ne/(\#(\tilde{I}) - 1)) \geq 2A^2 \log(n/\#(\tilde{I}))$, since the derivative of $x \mapsto x \log(ne/x)$ is $\log(n/x)$. Consequently, first $\#(j : \theta_{0,j}^2 \geq 2A^2 \log(n/\#(\tilde{I}))) \geq \#(j : \theta_{0,j}^2 \geq \theta_{0,i}^2) \geq \#(\tilde{I})$, by the minimising property of $\theta_{0,i}$, verifying the second inequality in (16). Second $\{j : \theta_{0,j}^2 < 2A^2 \log(n/q)\} \subset \{j : \theta_{0,j}^2 < \theta_{0,i}^2\} \subset \tilde{I}^c$, again by the minimising property of $\theta_{0,i}$. Thus the first inequality of (16) follows by the fact that $G(\tilde{I}) \leq C\#(\tilde{I}) \log(ne/\#(\tilde{I}))$.

Empirical Bayes condition and the MMLE

To obtain coverage in the empirical Bayes setting, we replace Condition 1 by the following.

Condition 5. The estimator $\widehat{\tau}_n$ satisfies, for a given sequence p_n and some constant $C > 1$, with $\tilde{p} = \tilde{p}(\theta_0)$,

$$\inf_{\theta_0 \in \Theta[p_n]} P_{\theta_0}\big(C^{-1}\tau_n(\tilde{p}) \leq \widehat{\tau}_n \leq C\tau_n(\tilde{p})\big) \to 1.$$

The lower bound of order $\tau_n(\tilde{p})$ instead of $1/n$ prevents over-shrinkage. Although this condition may appear more restrictive than Condition 1, Condition 5 may not be more stringent than Condition 1, because it only needs to hold for vectors θ_0 that meet the excessive-bias restriction.

For the coverage results in this paper, we need the additional result that the MMLE is of the order $\tau_n(\tilde{p}(\theta_0))$ for all vectors θ_0 satisfying the excessive-bias restriction.

Lemma 2. *For $p_n \to \infty$ such that $p_n = o(n)$, the MMLE $\widehat{\tau}_n$ satisfies Condition 5.*

Proof. See Appendix D.1 in the supplement (van der Pas et al., 2017b). □

The relative performances of the empirical Bayes procedures with the MMLE or the "simple" estimator are studied further in Section 5.

Main result on adaptive credible balls

Under the excessive-bias restriction, both the empirical and hierarchical Bayes credible balls are honest and adaptive. In the hierarchical Bayes setting, the hyperprior is assumed to be supported on $[1/n, 1]$, similar to the MMLE.

Theorem 5. *Let $\tilde{p}_n \leq p_n$ be given sequences with $\tilde{p}_n \to \infty$ and $p_n = o(n)$. If the estimator $\widehat{\tau}_n$ of τ satisfies Condition 5, then for a sufficiently large constant L (depending on A, C_s, C) the empirical Bayes credible ball $\hat{C}_n(L, \hat{\tau}_n)$ has honest coverage and rate adaptive (oracle) size:*

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \Theta[p_n], \tilde{p}(\theta_0) \geq \tilde{p}_n} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L, \hat{\tau}_n)\Big) \geq 1 - \alpha,$$
$$\inf_{\theta_0 \in \Theta[p_n]} P_{\theta_0}\Big(\operatorname{diam}\big(\hat{C}_n(L, \widehat{\tau}_n)\big) \lesssim \sqrt{\tilde{p}\log(n/\tilde{p})}\Big) \to 1.$$

In particular, these assertions are true for the MMLE. Furthermore, if $\tilde{p}_n \geq C \log n$ for a sufficiently large constant C, then the hierarchical Bayes method with $\tau \sim \pi_n$ for π_n probability densities on $[1/n, 1]$ that are bounded away from zero also yields adaptive and honest confidence sets: for sufficiently large L,

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \Theta[p_n], \tilde{p}(\theta_0) \geq \tilde{p}_n} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L)\Big) \geq 1 - \alpha,$$
$$\inf_{\theta_0 \in \Theta[p_n], \tilde{p}(\theta_0) \geq \tilde{p}_n} P_{\theta_0}\Big(\operatorname{diam}\big(\hat{C}_n(L)\big) \lesssim \sqrt{\tilde{p}\log(n/\tilde{p})}\Big) \to 1.$$

Proof. See Appendix E.1 in the supplement (van der Pas et al., 2017b). □

It may be noted that for self-similar θ_0 the square diameter of the credible balls is of the order $p \log(n/p)$, improving on the square contraction rate $p \log n$ obtained in van der Pas et al. (2017a). For parameters satisfying the excessive-bias restriction, this may further improve to $\tilde{p} \log(n/\tilde{p})$.

The size of the required blow-up factor L in the radius of the credible ball depends on the constants A, C_s, C in the excessive-bias restriction, Assumption 2. As argued in the introduction no statistical procedure can simultaneously adapt to sparsity and give uniform coverage over all parameters, so that with every given L necessarily some parameters will not be covered. In practice the validity of the excessive-bias restriction for particular A, C_s, C cannot be verified, but one must be satisfied with coverage for a set of "reasonable" parameters. The choice of L operationalises "reasonable"; it will be hard to define an optimal choice. In our simulations in the next section we used $L = 1$, which is the natural Bayesian choice and seems to work well, at least for the parameters considered in the simulation.

5 Simulation study

In the first simulation study in Section 5.1, we compare four versions of the horseshoe (empirical Bayes with two different estimators and hierarchical Bayes with two different

priors) and evaluate the coverage properties and interval lengths of the resulting credible intervals. In addition, we include an approximation to the credible intervals based on the normal distribution.

In the simulation study in Section 5.2, we compare the model selection properties of the method based on credible intervals resulting from the horseshoe with the MMLE, as discussed in Section 3, to the thresholding method introduced by Carvalho et al. (2010), with the MMLE of τ plugged in. We use the MMLE because the best results are obtained for the horseshoe with MMLE in the first simulation in Section 5.1. All simulations were carried out using the R package 'horseshoe' (van der Pas et al., 2016b).

5.1 Coverage, interval length, and τ

Several Markov Chain Monte Carlo (MCMC) samplers and software packages are available for computation of the posterior distribution (Scott, 2010; Makalic and Schmidt, 2016; Gramacy, 2014; van der Pas et al., 2016b; Hahn et al., 2016).

We study the relative performances of the empirical Bayes and hierarchical Bayes approaches further through simulation studies, extending the simulation study in van der Pas et al. (2014). We consider empirical Bayes combined with either (i) the simple estimator (with $c_1 = 2, c_2 = 1$) or (ii) the MMLE, and for hierarchical Bayes with either (iii) a Cauchy prior on τ, or (iv) a Cauchy prior truncated to $[1/n, 1]$ on τ. We study the coverage and average lengths of the marginal credible intervals resulting from these four methods, as well as intervals based solely on the posterior mean and variance. In addition, we study intervals of the form $\hat{\theta}_i(y_i, \widehat{\tau}_M) \pm 1.96\sqrt{\mathrm{var}(\theta_i \,|\, y_i, \widehat{\tau}_M)}$, based on a normal approximation to the posterior, where $\hat{\theta}_i(y_i, \widehat{\tau}_M)$ is the posterior mean and $\mathrm{var}(\theta_i \,|\, y_i, \widehat{\tau}_M)$ refers to the posterior variance, both with the MMLE plugged in. We include the approximation because it offers a computational advantage over the other methods, as no MCMC is required.

We consider a mean vector of length $n = 400$, with $p_n \in \{20, 200\}$. We draw the nonzero means from a $\mathcal{N}(A, 1)$-distribution, with $A = c\sqrt{2 \log n}$ for $c \in \{1/2, 1, 2\}$, corresponding to most nonzero means being below the universal threshold, close to the universal threshold, or well past the universal threshold, respectively. Instead of the symmetric intervals studied in our theorems, we computed the practically appealing quantile-based 95% marginal credible sets for the hierarchical and empirical Bayes methods, taking the 2.5%- and 97.5%-quantiles of the MCMC samples as the endpoints. We did not include a blow-up factor. The procedure was repeated $N = 500$ times.

Figure 2 gives the coverage results averaged over the 500 iterations, for all parameters, and separately for the p_n nonzero means and the $(n - p_n)$ zero means. The average lengths of the credible sets, again for all signals and separately for the nonzero and zero means, are displayed in Figure 3. Figure 4 gives the mean value of τ - in the hierarchical Bayes settings, the posterior mean of τ was recorded for each iteration. No value is given for the normal approximation, as it uses the MMLE as a plug-in value for τ.

We remark on some aspects of the results. First, we see that the zero means are nearly perfectly covered by all methods in all settings, and the main differences lie in

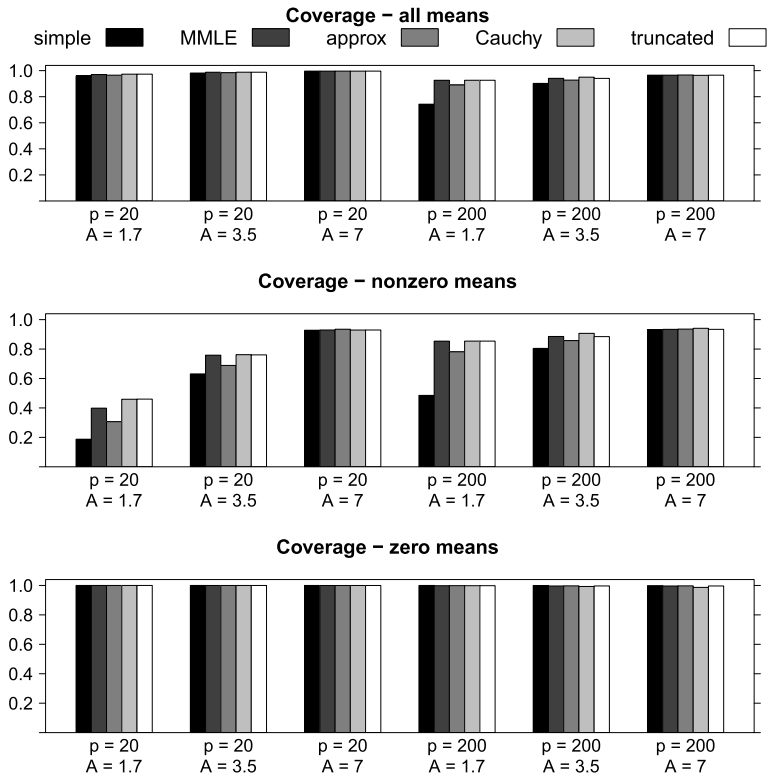


Figure 2: Average coverage of all parameters (top), the nonzero means (middle) and the zero means (bottom) for the five methods, from left to right: empirical Bayes with simple estimator ($c_1 = 2, c_2 = 1$) and MMLE, normal approximation, hierarchical Bayes with Cauchy prior on τ and with Cauchy prior truncated to $[1/n, 1]$. The p_n nonzero means were drawn from a $\mathcal{N}(A, 1)$ distribution. Results are based on averaging over 500 iterations.

the nonzero means. Secondly, coverage of the nonzero means improves as their values increase. Thirdly, the lengths of the credible intervals adapt to the signal size. They are smaller for the zero means than for the nonzero means, and smaller for the nonzero means corresponding to $A = (1/2)\sqrt{2\log n}$ than for the nonzero means corresponding to $A = \sqrt{2\log n}$ and $A = 2\sqrt{2\log n}$, while there is not much difference between the interval lengths in those latter two settings, suggesting that the interval length does not increase indefinitely with the size of the nonzero mean.

Furthermore, empirical Bayes with the simple estimator achieves the lowest overall coverage, and especially bad coverage of the nonzero means. This appears to be due to smaller interval lengths caused by lower estimates of τ compared to the other methods. The normal approximation leads to better coverage than the simple estimator, and has the highest coverage of the nonzero means, even though the corresponding intervals are

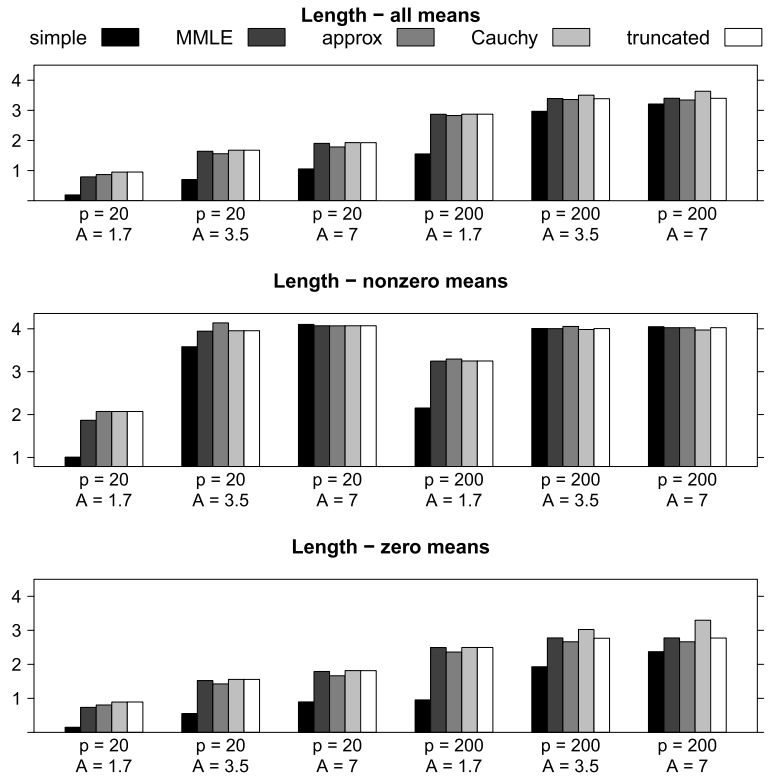


Figure 3: Average length of the credible sets of all parameters (top), the nonzero means (middle) and the zero means (bottom) for the five methods, from left to right: empirical Bayes with simple estimator ($c_1 = 2, c_2 = 1$) and MMLE, normal approximation, hierarchical Bayes with Cauchy prior on τ and with Cauchy prior truncated to $[1/n, 1]$. The p_n nonzero means were drawn from a $\mathcal{N}(A, 1)$ distribution. Results are based on averaging over 500 iterations.

slightly shorter than those of empirical Bayes with the MMLE and the hierarchical Bayes approaches. However, its coverage of nonzero means is worse than that of those three methods, while the corresponding intervals are longer, except in the case where A is largest. The normal approximation appears to be reasonable for very large signals only.

The hierarchical Bayes approach with a non-truncated Cauchy on τ leads to the highest overall coverage and coverage of the nonzero means, albeit by a small margin. The price is slightly larger intervals compared to the other methods, mostly for the zero means. These larger intervals are most likely due to the larger values of τ that are employed, this being the only approach that allows for estimates of τ larger than one, and it avails itself of the opportunity in the non-sparse setting. Finally, we again observe that the results for empirical Bayes with the MMLE and hierarchical Bayes with a truncated Cauchy lead to highly similar results. Their coverage is comparable

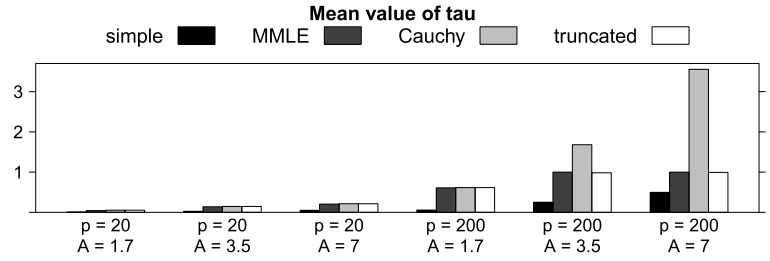


Figure 4: Average value of τ for four methods, from left to right: empirical Bayes with simple estimator ($c_1 = 2, c_2 = 1$) and MMLE, hierarchical Bayes with Cauchy prior on τ and with Cauchy prior truncated to $[1/n, 1]$. For the hierarchical Bayes approaches, the posterior mean of τ was recorded for each iteration. The p_n nonzero means were drawn from a $\mathcal{N}(A, 1)$ distribution. Results are based on averaging over 500 iterations.

to that of hierarchical Bayes with a non-truncated Cauchy in all settings except when $p_n = 200$ and A is at least at the threshold, in which case the non-truncated Cauchy has slightly better coverage. Their intervals are shorter on average, because τ is not allowed to be larger than one.

In conclusion, empirical Bayes with the simple estimator should not be used for uncertainty quantification. The normal approximation is faster to compute than the marginal credible sets, but leads to worse coverage of the nonzero compared to the empirical Bayes with the MMLE and the hierarchical Bayes approaches, unless the nonzero means are very large. The results of those latter three methods are very similar to each other. All these results can be understood in terms of the behaviour of the estimate of τ: larger values lead to larger intervals and better coverage, which may lead to worse estimates however (as seen in the previous section). Empirical Bayes with the MMLE, or hierarchical Bayes with a truncated Cauchy, appear to be the best choices when considering both estimation and coverage. Those two approaches yield highly similar results and the choice for one over the other may be based on other considerations such as computational ones.

5.2 Model selection

We compare the procedure based on credible intervals studied in Section 3 to the thresholding method introduced in Carvalho et al. (2010). Two scenarios are considered. In the first, the signals are either "small", "intermediate" or "large", as defined in Section 2.3. In the second, all signals are drawn from a distribution.

In the credible interval method, a parameter is selected as a signal if zero is not contained in the corresponding credible interval. For the thresholding method of Carvalho et al. (2010), the posterior mean is divided by the observation. The result is a number between zero and one, which indicates the amount of shrinkage of that particular observation. If this number is larger than 0.5, the corresponding parameter is considered a signal. For both methods, we estimate τ by the MMLE.

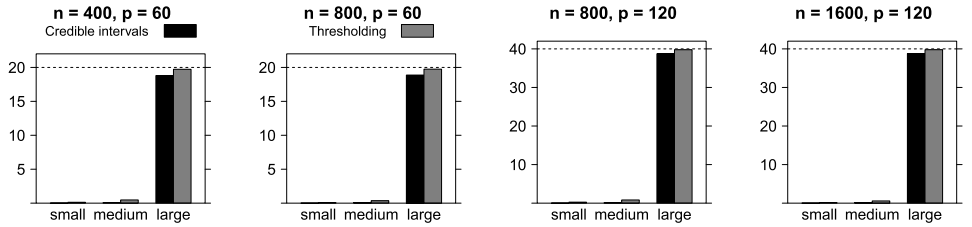


Figure 5: Number of true discoveries, split up by signal size, in scenario 1. The true number of signals in each category is indicated by the dotted line.

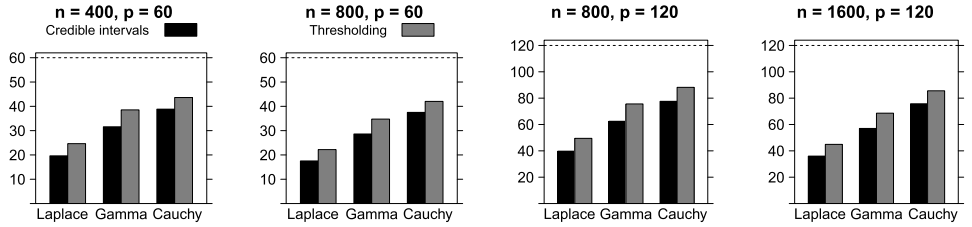


Figure 6: Number of true discoveries in scenario 2. The true number of signals in each category is indicated by the dotted line.

In the first scenario, we have n observations, with p_n signals. The p_n signals are divided into three groups, corresponding to the three intervals of Section 2.3. The small ones are equal to $1/n$, the intermediate ones are $0.5\sqrt{2\log(1/\tau_n(p_n))}$, and the large ones are equal to $1.5\sqrt{2\log n}$. We study four combinations of n and p_n: $n = 400, p_n = 60$; $n = 800, p_n = 60$; $n = 800, p_n = 120$ and $n = 1600, p_n = 120$. We count the number of false positives, that is the noise signals that are incorrectly selected as signals, and the number of correctly selected signals in each group. The number of true discoveries, averaged over $N = 500$ iterations, are in Figure 5, and the false discovery rate (FDR) is in the upper left panel of Figure 7.

In the second scenario, all signals are drawn from a distribution: the Laplace distribution with dispersion parameter equal to 3, the Gamma distribution with shape and scale equal to 2, or the Cauchy distribution with scale equal to 5. The number of false positives and the number of correctly selected variables are counted. The number of true discoveries, averaged over $N = 100$ iterations, are in Figure 6, and the FDRs are in Figure 7.

Both simulation scenarios tell a consistent story: the thresholding method results in more discoveries, both true and false, than the credible interval method. The findings of Figures 5 are as expected based on the theoretical results of Section 3: almost none of the small and medium signals are detected, while the large signals are nearly perfectly detected by both methods. In scenario 2, where the signals are drawn from a distribution, the thresholding method finds more of the signals. Comparing the left and right columns of Figure 6, we see that both methods detect more of the signals when the truth is less

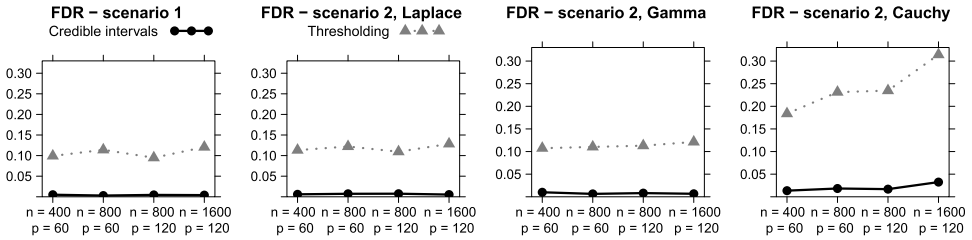


Figure 7: False discovery rate in scenarios 1 and 2. (i) $n = 400$, $p_n = 60$; (ii) $n = 800$, $p_n = 60$; (iii) $n = 800$, $p_n = 120$; (iv) $n = 1600$, $p_n = 120$.

sparse. This may be due to the behaviour of the MMLE, which is likely to be larger in the less sparse settings, leading to less shrinkage of the true signals.

The FDR of the credible interval method remains well below 0.05 in all settings (Figure 7). In contrast, the FDR of the thresholding method exceeds 0.10 in all cases, and is much larger still when the observations are drawn from a Cauchy distribution. The FDR of the thresholding method can of course be lowered by taking a different cut-off than 0.5, but no guidelines are available at the moment, and a decrease of the FDR will come at the cost of the number of true discoveries. The credible intervals have low FDR, but fail to detect small and medium observations. We speculate that improvement might be possible by combining the information contained in the posterior mean and variance.

6 Proof of Theorem 1

The posterior distribution of θ_i given (Y_i, τ, λ_i) is normal with mean and variance

$$\hat{\theta}_i(\tau, \lambda_i) := \mathrm{E}(\theta_i \,|\, Y_i, \tau, \lambda_i) = \frac{\lambda_i^2 \tau^2}{1 + \lambda_i^2 \tau^2} Y_i, \tag{17}$$

$$r_i^2(\tau, \lambda_i) := \mathrm{var}(\theta_i \,|\, Y_i, \tau, \lambda_i) = \frac{\lambda_i^2 \tau^2}{1 + \lambda_i^2 \tau^2}. \tag{18}$$

Furthermore, the posterior distribution of λ_i given (Y_i, τ) possesses a density function given by

$$\pi(\lambda_i \,|\, Y_i, \tau) \propto e^{-\frac{Y_i^2}{2(1+\lambda_i^2\tau^2)}} (1 + \tau^2 \lambda_i^2)^{-1/2} (1 + \lambda_i^2)^{-1}.$$

The parameter $\theta_{0,i}$ is contained in $C_{ni}(L, \tau)$ if and only if $|\theta_{0,i} - \hat{\theta}_i(\tau)| \leq L \hat{r}_i(\alpha, \tau)$. We show that this is true, or not, for $\theta_{0,i}$ belonging to the three regions separately for $\mathcal{S}$, $\mathcal{L}$ and $\mathcal{M}$.

Case $\mathcal{S}$: proof of (5). If $i \in \mathcal{S}$, then $|\theta_{0,i} - \hat{\theta}_i(\tau)| \leq k_S \tau + \tau |Y_i| e^{Y_i^2/2}$, by the triangle inequality and Lemma 3(iii). Below we show that $\hat{r}_i(\alpha, \tau) \geq \tau z_\alpha c$, with probability tending to one, for z_α the standard normal upper α-quantile and every $c < 1/2$. Hence $\theta_{0,i} \in C_{ni}(L, \tau)$ as soon as $|Y_i| e^{Y_i^2/2} \leq L z_\alpha c - k_S$.

For $i \in \mathcal{S}$ the variable $|Y_i|$ is stochastically bounded by $|\theta_{0,i}| + |\varepsilon_i| \leq k_S\tau + |\varepsilon_i|$. Since the variables $|\varepsilon_i|$ are i.i.d. with quantile function $u \mapsto \Phi^{-1}((u+1)/2) \leq \sqrt{2\log(2/(1-u))}$, a fraction $1-\gamma$ of the variables Y_i with $i \in \mathcal{S}$ is bounded above by $k_S\tau + \sqrt{2\log(2/\gamma)} + \delta = k_S\tau + \zeta_{\gamma/2} + \delta$, with probability tending to 1, for any $\delta > 0$. Then the corresponding fraction of parameters $\theta_{0,i}$ is contained in their credible interval if L is chosen big enough that

$$Lz_\alpha c - k_S \geq (k_S\tau + \zeta_{\gamma/2} + \delta)e^{(k_S\tau+\zeta_{\gamma/2}+\delta)^2/2}.$$

As the right hand side of the above inequality is bounded above by $\frac{2}{\gamma}\zeta_{\gamma/2}(1+\varepsilon)$, where $\varepsilon \to 0$ if $\gamma, \tau, \delta \to 0$, this is certainly true for L_S as in the theorem. We also note that the above argument implies that for every given $i \in \mathcal{S}$ the variable Y_i is bounded from above by $k_S\tau + \zeta_{\gamma/2} + \delta$ with probability at least $1-\gamma$.

We finish by proving the lower bound for the radius $\hat{r}_i(\alpha, \tau)$. Because the conditional distribution of θ_i given (Y_i, τ, λ_i) is normal with mean $\hat{\theta}_i(\tau, \lambda_i)$ it follows by Anderson's lemma that $\Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| > r \,|\, Y_i, \tau, \lambda_i) \geq \Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau, \lambda_i)| > r \,|\, Y_i, \tau, \lambda_i)$, for any $r > 0$. Furthermore, by the monotonicity of the variance in λ_i of this conditional distribution, the last function is increasing in λ_i. If $\tilde{\pi}(\cdot \,|\, \tau)$ is the probability density given by $\tilde{\pi}(\lambda_i \,|\, \tau) \propto (\lambda_i^2\tau^2 + 1)^{-1/2}(1+\lambda_i^2)^{-1}$, then $\lambda_i \mapsto \pi(\lambda_i \,|\, Y_i, \tau)/\tilde{\pi}(\lambda_i \,|\, \tau)$ is increasing. Combining the preceding observations with Lemma G.2, we see that

$$\begin{aligned}\alpha &= \int_0^\infty \Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| > \hat{r}_i(\alpha, \tau) \,|\, Y_i, \tau, \lambda_i)\pi(\lambda_i \,|\, Y_i, \tau)\, d\lambda_i \\ &\geq \int_0^\infty \Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau, \lambda_i)| > \hat{r}_i(\alpha, \tau) \,|\, Y_i, \tau, \lambda_i)\tilde{\pi}(\lambda_i \,|\, \tau)\, d\lambda_i. \end{aligned} \tag{19}$$

On the other hand, since $\mathrm{sd}(\theta_i \,|\, Y_i, \tau, \lambda_i) \geq \tau/2(1+o(1))$, for $\lambda_i \geq 1/2$, by (18), the normality of the conditional distribution of θ_i given (Y_i, τ, λ_i) gives that

$$\begin{aligned}&\int_0^\infty \Pi\big(\theta_i : |\theta_i - \hat{\theta}_i(\tau, \lambda_i)| > z_\alpha\tau/2(1+o(1))) \,|\, Y_i, \tau, \lambda_i\big)\tilde{\pi}(\lambda_i \,|\, \tau)\, d\lambda_i \\ &\quad \geq 2\alpha\, \tilde{\Pi}(\lambda_i \geq 1/2 \,|\, \tau) \geq 2\alpha \times 2/3 > \alpha. \end{aligned} \tag{20}$$

Here the second last inequality follows from

$$\frac{\int_0^{1/2}(\lambda_i^2\tau^2+1)^{-1/2}\,(1+\lambda_i^2)^{-1}\, d\lambda}{\int_0^\infty(\lambda_i^2\tau^2+1)^{-1/2}\,(1+\lambda_i^2)^{-1}\, d\lambda_i} \to \frac{\int_0^{1/2}(1+\lambda_i^2)^{-1}\, d\lambda_i}{\int_0^\infty(1+\lambda_i^2)^{-1}\, d\lambda_i} < \frac{1}{3}, \tag{21}$$

as $\tau \to 0$, by two applications of the dominated convergence theorem. Combination of (19) and (20) shows that $\hat{r}_i(\alpha, \tau) \geq z_\alpha\tau/2(1+o(1))$.

Case $\mathcal{L}$: proof of (7). If $i \in \mathcal{L}$, then

$$|\theta_{0,i} - \hat{\theta}_i(\tau)| \leq |\theta_{0,i} - Y_i| + |Y_i - \hat{\theta}_i(\tau)| \leq |\varepsilon_i| + 2\zeta_\tau^{-1}, \tag{22}$$

eventually, provided $|Y_i| \geq A\zeta_\tau$ for some constant $A > 1$, by the triangle inequality and Lemma 3(i). Below we show that $\hat{r}_i(\alpha, \tau) \geq z_{\alpha'/2} + o(1)$, for every $\alpha' > \alpha$ with probability tending to one. It then follows that $\theta_{0,i} \in C_{ni}(L, \tau)$ as soon as $|Y_i| \geq A\zeta_\tau$ and $|\varepsilon_i| \leq Lz_{\alpha'/2} + o(1) - 2\zeta_\tau^{-1} = Lz_{\alpha'/2} + o(1)$.

For $i \in \mathcal{L}$ the variable $|Y_i|$ is lower bounded by $|\theta_{0,i}| - |\varepsilon_i| \geq k_L\zeta_\tau - |\varepsilon_i|$ and hence $|Y_i| \geq A\zeta_\tau$ if $|\varepsilon_i| \leq (k_L - A)\zeta_\tau$. This is automatically satisfied if $|\varepsilon_i| \leq Lz_{\alpha'/2} + o(1)$, for constants L with $L \ll \zeta_\tau$. As for the proof of Case $\mathcal{S}$ we have that $|\varepsilon_i| \leq Lz_{\alpha'/2} + o(1)$ with probability tending to one for a fraction $1-\gamma$ of the indices $i \in \mathcal{S}$ if $L \geq z_{\alpha'/2}^{-1}\zeta_{\gamma/2}+\delta$, for some $\delta > 0$. This is satisfied by L_L. Also note that for every $i \in \mathcal{L}$ the inequality $|\varepsilon_i| \leq Lz_{\alpha'/2} + o(1)$ holds with probability at least $1 - \gamma$ for $L \geq z_{\alpha'/2}^{-1}\zeta_{\gamma/2} + \delta$.

The proof that $\hat{r}_i(\alpha, \tau) \geq z_{\alpha'/2} + o(1)$ follows the same lines as the proof of the corresponding result in Case $\mathcal{S}$, expressed in (19) and (20), but with the true density π instead of $\tilde{\pi}$. Inequality (19) with π instead of $\tilde{\pi}$ is valid by Anderson's lemma, while in (20) we replace $z_\alpha\tau/2(1 + o(1))$ by $z_{\alpha'/2} + o(1)$. Since $\operatorname{var}(\theta_i \,|\, Y_i, \tau, \lambda_i) \geq g_\tau/(1 + g_\tau) = 1 + o(1)$ for every $\lambda_i \geq g_\tau/\tau$ and $g_\tau \to \infty$, the desired result follows if $\Pi(\lambda_i \geq g_\tau/\tau \,|\, Y_i, \tau) = 1+o(1)$, for every i such that $|Y_i| \geq A\zeta_\tau$. Now by the form of $\pi(\lambda_i \,|\, Y_i, \tau)$, for any $c, d > 0$,

$$\begin{aligned}\Pi(\lambda_i \leq g_\tau/\tau \,|\, Y_i, \tau) &\leq \frac{e^{-\frac{Y_i^2}{2(1+c^2)}}\int_0^{c/\tau}(1+\lambda^2)^{-1}\,d\lambda + e^{-\frac{Y_i^2}{2(1+g_\tau^2)}}\int_{c/\tau}^{g_\tau/\tau}(1+c^2/\tau^2)^{-1}\,d\lambda}{e^{-\frac{Y_i^2}{2(1+d^2g_\tau^2)}}\int_{dg_\tau/\tau}^{2dg_\tau/\tau}(1+4d^2g_\tau^2)^{-1/2}(1+4d^2g_\tau^2/\tau^2)^{-1}\,d\lambda} \\ &\lesssim \frac{\exp\Big[-\frac{Y_i^2}{2}\Big(\frac{1}{1+c^2} - \frac{1}{1+d^2g_\tau^2}\Big)\Big] + \exp\Big[-\frac{Y_i^2}{2}\Big(\frac{1}{1+g_\tau^2} - \frac{1}{1+d^2g_\tau^2}\Big)\Big]g_\tau\tau}{(g_\tau/\tau)(1/g_\tau)(\tau^2/g_\tau^2)}.\end{aligned} \tag{23}$$

For $|Y_i| > A\zeta_\tau$ and $A > 1$ we can choose c sufficiently close to zero so that the first exponential is of order $\tau^{A'}$ for some $A' > 1$. Then it is much smaller than the denominator, which is of order τ/g_τ^2, provided g_τ tends to infinity slowly. If we choose $d > 1$, then the term involving the second exponential will also tend to zero for $|Y_i| > A\zeta_\tau$ as soon as $e^{-c\zeta_\tau^2/g_\tau^2}g_\tau^3 \to 0$, for a sufficiently small constant c. This is true (for any $c > 0$) for instance if $g_\tau = \sqrt{\zeta_\tau}$. Then the quotient tends to zero, whence the analogon of (20) is lower bounded by $\alpha'(1 - o(1)) > \alpha$, eventually.

Case $\mathcal{M}$: proof of (6). We show below that $\hat{r}_i(\alpha, \tau) \lesssim U_\tau := \tau(1 \vee |Y_i|e^{Y_i^2/2})$, with probability tending to one, whenever $i \in \mathcal{M}$. By Lemma 3(iii) exactly the same bound is valid for $|\hat{\theta}_i(\tau)|$. If $|\hat{\theta}_i(\tau)|+\hat{r}_i(\alpha, \tau) \lesssim U_\tau$, but $|\theta_{0,i}| \gg U_\tau$ then $\theta_{0,i} \notin C_{ni}(L, \tau)$ eventually, and hence it suffices to prove that the probability of the event that $|\theta_{0,i}| \gg U_\tau$ tends to one whenever $i \in \mathcal{M}$. Consider two cases. If $|\theta_{0,i}| \leq 1$, then $|Y_i| \leq 1 + |\varepsilon_i| = O_P(1)$ and hence $U_\tau = O_P(\tau)$. For $i \in \mathcal{M}$, we have $|\theta_{0,i}| \gg \tau$ and hence $|\theta_{0,i}| \gg U_\tau$ with probability tending to one. On the other hand, if $|\theta_{0,i}| \geq 1$ but $|\theta_{0,i}| \leq k_M\zeta_\tau$, then $|Y_i| \leq k\zeta_\tau$ with probability tending to one for any $k > k_M$, and hence $U_\tau \lesssim \tau\zeta_\tau e^{k^2\zeta_\tau^2/2} = \tau^{1-k^2}\zeta_\tau$. Since $k_M < 1$ we can choose $k < 1$, so that $\tau^{1-k^2}\zeta_\tau \to 0$, and again we have $|\theta_{0,i}| \gg U_\tau$ with probability tending to one.

We finish by proving that $\hat{r}_i(\alpha, \tau) \lesssim U_\tau$, with probability tending to one. As a first step we show that, for $k < 1$,

$$\lim_{M\to\infty}\sup_{|y|\leq k\zeta_\tau}\Pi(\lambda_i \geq M \,|\, Y_i = y, \tau) \to 0. \tag{24}$$

By the explicit form of the posterior density of λ_i we have

$$\Pi(\lambda_i \geq M \,|\, Y_i = y, \tau) \leq \frac{\int_M^\infty e^{-\frac{y^2}{2(1+\lambda_i^2\tau^2)}} (1+\lambda_i^2\tau^2)^{-1/2}(1+\lambda_i^2)^{-1}\, d\lambda_i}{\int_1^2 e^{-\frac{y^2}{2(1+\lambda_i^2\tau^2)}} (1+\lambda_i^2\tau^2)^{-1/2}(1+\lambda_i^2)^{-1}\, d\lambda_i}$$
$$\leq e^{y^2/2} 5\sqrt{2} \int_M^\infty e^{-\frac{y^2}{2(1+\lambda_i^2\tau^2)}} (1+\lambda_i^2\tau^2)^{-1/2}(1+\lambda_i^2)^{-1}\, d\lambda_i. \tag{25}$$

We split the remaining integral over the intervals $[M, \tau^{-a})$ and $[\tau^{-a}, \infty)$, for some $a < 1$. On the first interval we use that $y^2/(1+\lambda_i^2\tau^2) = y^2 + o(1)$, uniformly in $|y| \lesssim \zeta_\tau$ and $\lambda_i \leq \tau^{-a}$, while on the second we simply bound the factor $e^{-y^2/(2(1+\lambda_i^2\tau^2))}$ by 1, to see that the preceding display is bounded above by

$$e^{y^2/2} 5\sqrt{2}\Big[e^{-y^2/2} e^{o(1)} \int_M^{\tau^{-a}} (1+\lambda_i^2)^{-1}\, d\lambda_i + \int_{\tau^{-a}}^\infty (1+\lambda_i^2)^{-1}\, d\lambda_i\Big].$$

The first term in square brackets (times the leading term) contributes less than a multiple of $\int_M^\infty \lambda^{-2}\, d\lambda = 1/M$, while the second term contributes less than $e^{y^2/2}\tau^a \leq \tau^{-k^2+a}$, for $|y| \leq k\zeta_\tau$, which tends to zero if $a > k^2$. This concludes the proof of (24).

By the reverse triangle inequality, for any $M > 0$,

$$\int_0^\infty \Pi(\theta_i : |\theta_i - \hat\theta_i(\tau)| \geq r + |\hat\theta_i(\tau, \lambda_i) - \hat\theta_i(\tau)| \;\,|\, Y_i, \lambda_i, \tau)\pi(\lambda_i \,|\, Y_i, \tau)\, d\lambda_i$$
$$\leq \int_0^M \Pi(\theta_i : |\theta_i - \hat\theta_i(\tau, \lambda_i)| \geq r \,|\, Y_i, \lambda_i, \tau)\pi(\lambda_i \,|\, Y_i, \tau)\, d\lambda_i + \Pi(\lambda_i \geq M \,|\, Y_i, \tau).$$

For sufficiently large M the second term on the far right is smaller than $\alpha/2$ by the preceding paragraph and for $r = z_{\alpha/4} \sup_{\lambda \leq M} r_i(\tau, \lambda)$ the first term on the right is smaller than $\alpha/2$ as well, by the normality of θ_i given (Y_i, λ_i, τ) and the definition of $r_i(\tau, \lambda_i)$. The inequality remains valid if $|\hat\theta_i(\tau, \lambda_i) - \hat\theta_i(\tau)|$ in the first line is replaced by $\sup_{\lambda_i \leq M} |\hat\theta_i(\tau, \lambda_i)| + |\hat\theta_i(\tau)|$. It follows that

$$\hat r_i(\alpha, \tau) \leq z_{\alpha/4} \sup_{\lambda_i \leq M} r_i(\tau, \lambda_i) + \sup_{\lambda_i \leq M} |\hat\theta_i(\tau, \lambda_i)| + |\hat\theta_i(\tau)|.$$

The first term is bounded above by $M\tau$, and the second by $M\tau|Y_i|$, by the definitions of $r_i(\tau, \lambda)$ and $\hat\theta_i(\tau, \lambda)$, while $|\hat\theta_i(\tau)| \leq \tau|Y_i|e^{Y_i^2/2}$, by Lemma 3(iii). This concludes the proof that $\hat r_i(\alpha, \tau) \lesssim U_\tau$.

Remark 5. The proof of the more general statement of Remark 2 follows similar lines of reasoning as the proof of Theorem 1. The main differences are that in the computation of the marginal posterior probabilities in (21), (23) and (25) the term $(1+\lambda^2)^{-1}$ is replaced by $\lambda^{-1-2a}L(\lambda^2)$ and that the upper bound $|\hat\theta_i(\tau)| \leq \tilde C\tau|Y_i|e^{Y_i^2/2}$ can be derived using (19) of van der Pas et al. (2016a) instead of Lemma 3(iii).

The following lemma and its proof can be found in van der Pas et al. (2017a).

Lemma 3. *For $A > 1$ and every $y \in \mathbb{R}$,*

(i) $|\mathrm{E}(\theta_i \,|\, Y_i = y, \tau) - y| \leq 2\zeta_\tau^{-1}$, *for* $|y| \geq A\zeta_\tau$, *as* $\tau \to 0$.

(ii) $|\mathrm{E}(\theta_i \,|\, Y_i = y, \tau)| \leq |y|$.

(iii) $|\mathrm{E}(\theta_i \,|\, Y_i = y, \tau)| \leq \tau |y| e^{y^2/2}$, *as* $\tau \to 0$.

(iv) $|\operatorname{var}(\theta_i \,|\, Y_i = y, \tau) - 1| \leq \zeta_\tau^{-2}$, *for* $|y| \geq A\zeta_\tau$, *as* $\tau \to 0$.

(v) $\operatorname{var}(\theta_i \,|\, Y_i = y, \tau) \leq 1 + y^2$.

(vi) $\operatorname{var}(\theta_i \,|\, Y_i = y, \tau) \lesssim \tau e^{y^2/2}(y^{-2} \wedge 1)$, *as* $\tau \to 0$.

(vii) $|\mathrm{E}(\theta_i \,|\, Y_i = y, \tau) - y| \lesssim (\log |y|)/|y|$, *uniformly in* $\tau \geq \tau_0 > 0$ *and* $|y| \to \infty$.

Supplementary Material

Supplement to: Uncertainty quantification for the horseshoe
(DOI: 10.1214/17-BA1065SUPP; .pdf). The remaining proofs are given in the supplement.

References

Armagan, A., Dunson, D. B., and Lee, J. (2013). "Generalized Double Pareto Shrinkage." *Statistica Sinica*, 23: 119–143. MR3076161. 1222

Belitser, E. (2017). "On coverage and local radial rates of credible sets." *Annals of Statistics*, 45(3): 1124–1151. MR3662450. doi: http://dx.doi.org/10.1214/16-AOS1477. 1224, 1233

Belitser, E. and Nurushev, N. (2015). "Needles and straw in a haystack: empirical Bayes confidence for possibly sparse sequences." *ArXiv e-prints*. 1223, 1224, 1233

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). "The Horseshoe+ Estimator of Ultra-Sparse Signals." Advance publication. http://dx.doi.org/10.1214/16-BA1028 1221

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2015). "Fast sampling with Gaussian scale-mixture priors in high-dimensional regression." *ArXiv e-prints*. MR3620452. doi: http://dx.doi.org/10.1093/biomet/asw042. 1224

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2014). "Dirichlet-Laplace Priors for Optimal Shrinkage." ArXiv:1401.5398. MR3449048. doi: http://dx.doi.org/10.1080/01621459.2014.960967. 1221, 1222

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg. MR2807761. doi: http://dx.doi.org/10.1007/978-3-642-20192-9. 1233

Bull, A. (2012). "Honest adaptive confidence bands and self-similar functions." *Electronic Journal of Statistics*, 6: 1490–1516. http://projecteuclid.org/euclid.ejs/1346421602 MR2988456. doi: http://dx.doi.org/10.1214/12-EJS720. 1233

Caron, F. and Doucet, A. (2008). "Sparse Bayesian Nonparametric Regression." In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 88–95. New York, NY, USA: ACM. 1221

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling Sparsity via the Horseshoe." *Journal of Machine Learning Research, W&CP*, 5: 73–80. 1222

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The Horseshoe Estimator for Sparse Signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: http://dx.doi.org/10.1093/biomet/asq017. 1222, 1230, 1235, 1238, 1239

Castillo, I. and Nickl, R. (2014). "On the Bernstein von Mises phenomenon for nonparametric Bayes procedures." *Annals of Statistics*, 42(5): 1941–1969. MR3262473. doi: http://dx.doi.org/10.1214/14-AOS1246. 1224

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: http://dx.doi.org/10.1214/15-AOS1334. 1221

Castillo, I. and Van der Vaart, A. W. (2012). "Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences." *Annals of Statistics*, 40(4): 2069–2101. MR3059077. doi: http://dx.doi.org/10.1214/12-AOS1029. 1221

Datta, J. and Ghosh, J. K. (2013). "Asymptotic Properties of Bayes Risk for the Horseshoe Prior." *Bayesian Analysis*, 8(1): 111–132. MR3036256. doi: http://dx.doi.org/10.1214/13-BA805. 1222, 1230

Ghosh, P. and Chakrabarti, A. (2015). "Posterior Concentration Properties of a General Class of Shrinkage Estimators around Nearly Black Vectors." ArXiv:1412.8161v2. 1221, 1226

Giné, E. and Nickl, R. (2010). "Confidence bands in density estimation." *Annals of Statistics*, 38(2): 1122–1170. MR2604707. doi: http://dx.doi.org/10.1214/09-AOS738. 1233

Gramacy, R. B. (2014). *monomvn: Estimation for multivariate normal and Student-t data with monotone missingness*. R package version 1.9-5. http://CRAN.R-project.org/package=monomvn 1235

Griffin, J. E. and Brown, P. J. (2010). "Inference with Normal-Gamma Prior Distributions in Regression Problems." *Bayesian Analysis*, 5(1): 171–188. MR2596440. doi: http://dx.doi.org/10.1214/10-BA507. 1221

Hahn, R. P., He, J., and Lopes, H. (2016). *fastHorseshoe: The Elliptical Slice Sampler for Bayesian Horseshoe Regression*. R package version 0.1.0. https://cran.r-project.org/package=fastHorseshoe 1235

Jiang, W. and Zhang, C.-H. (2009). "General maximum likelihood empirical Bayes estimation of normal means." *Annals of Statistics*, 37(4): 1647–1684. MR2533467. doi: http://dx.doi.org/10.1214/08-AOS638. 1221

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(2): 143–170. MR2830762. doi: http://dx.doi.org/10.1111/j.1467-9868.2009.00730.x. 1221

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences." *Annals of Statistics*, 32(4): 1594–1649. MR2089135. doi: http://dx.doi.org/10.1214/009053604000000030. 1221, 1222

Li, K.-C. (1989). "Honest confidence regions for nonparametric regression." *Annals of Statistics*, 17(3): 1001–1008. MR1015135. doi: http://dx.doi.org/10.1214/aos/1176347253. 1222, 1223

Liu, H. and Yu, B. (2013). "Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression." *Electronic Journal of Statistics*, 7: 3124–3169. MR3151764. 1223

Makalic, E. and Schmidt, D. F. (2016). "A Simple Sampler for the Horseshoe Estimator." *IEEE Signal Processing Letters*, 23(1): 179–182. 1235

Nickl, R. and Szabo, B. (2016). "A sharp adaptive confidence ball for self-similar functions." *Stochastic Processes and their Applications*, 126(12): 3913–3934. http://www.sciencedirect.com/science/article/pii/S0304414916300394 MR3565485. doi: http://dx.doi.org/10.1016/j.spa.2016.04.017. 1233

Nickl, R. and van de Geer, S. (2013). "Confidence sets in sparse regression." *Annals of Statistics*, 41(6): 2852–2876. MR3161450. doi: http://dx.doi.org/10.1214/13-AOS1170. 1222, 1223

Picard, D. and Tribouley, K. (2000). "Adaptive confidence interval for pointwise curve estimation." *Annals of Statistics*, 28(1): 298–335. http://projecteuclid.org/euclid.aos/1016120374 MR1762913. doi: http://dx.doi.org/10.1214/aos/1016120374. 1233

Polson, N. G. and Scott, J. G. (2010). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 9*. Oxford University Press. MR3204017. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0017. 1222

Polson, N. G. and Scott, J. G. (2012a). "Good, Great or Lucky? Screening for Firms with Sustained Superior Performance Using Heavy-Tailed Priors." *The Annals of Applied Statistics*, 6(1): 161–185. 1222

Polson, N. G. and Scott, J. G. (2012b). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, 7(4): 887–902. 1222

Ray, K. (2014). "Adaptive Bernstein-von Mises theorems in Gaussian white noise." *ArXiv e-prints*. 1224

Robins, J. and van der Vaart, A. (2006). "Adaptive nonparametric confidence sets." *Annals of Statistics*, 34(1): 229–253. MR2275241. doi: http://dx.doi.org/10.1214/009053605000000877. 1222, 1223

Ročková, V. (2015). "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." 1221

Rousseau, J. and Szabo, B. (2016). "Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors in general settings." *ArXiv e-prints*. 1224, 1233

Scott, J. G. (2010). "Parameter Expansion in Local-Shrinkage Models." ArXiv: 1010.5265. 1235

Scott, J. G. (2011). "Bayesian Estimation of Intensity Surfaces on the Sphere via Needlet Shrinkage and Selection." *Bayesian Analysis*, 6(2): 307–328. 1222

Serra, P. and Krivobokova, T. (2017). "Adaptive Empirical Bayesian Smoothing Splines." *Bayesian Analysis*, 12(1): 219–238. MR3597573. doi: http://dx.doi.org/10.1214/16-BA997. 1224

Sniekers, S. and van der Vaart, A. (2015a). "Adaptive Bayesian credible sets in regression with a Gaussian process prior." *Electronic Journal of Statistics*, 9(2): 2475–2527. MR3425364. doi: http://dx.doi.org/10.1214/15-EJS1078. 1224

Sniekers, S. and van der Vaart, A. (2015b). "Adaptive credible bands in nonparametric regression with Brownian motion prior." *preprint*. 1224

Sniekers, S. and van der Vaart, A. (2015c). "Credible sets in the fixed design model with Brownian motion prior." *Journal of Statistical Planning and Inference*, 166: 78–86. MR3390135. doi: http://dx.doi.org/10.1016/j.jspi.2014.07.008. 1224, 1233

Szabó, B., van der Vaart, A., and van Zanten, H. (2015a). "Honest Bayesian confidence sets for the L2-norm." *Journal of Statistical Planning and Inference*, 166: 36–51. Special Issue on Bayesian Nonparametrics. http://www.sciencedirect.com/science/article/pii/S0378375814001244 MR3390132. doi: http://dx.doi.org/10.1016/j.jspi.2014.06.005. 1224

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015b). "Frequentist coverage of adaptive nonparametric Bayesian credible sets." *Annals of Statistics*, 43(4): 1391–1428. MR3357861. doi: http://dx.doi.org/10.1214/14-AOS1270. 1224, 1231, 1233

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 58(1): 267–288. MR1379242. 1221

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models." *Annals of Statistics*, 42(3): 1166–1202. MR3224285. doi: http://dx.doi.org/10.1214/14-AOS1221. 1223

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). "The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso)." *Electronic Journal of Statistics*, 5: 688–749. MR2820636. doi: http://dx.doi.org/10.1214/11-EJS624. 1233

van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016a). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10(1): 976–1000. MR3486423. doi: http://dx.doi.org/10.1214/16-EJS1130. 1244

van der Pas, S., Scott, J., Chakraborty, A., and Bhattacharya, A. (2016b). *horseshoe: Implementation of the Horseshoe Prior*. R package version 0.1.0. https://CRAN.R-project.org/package=horseshoe 1235

van der Pas, S., Szabó, B., and van der Vaart, A. (2017a). "Adaptive posterior contraction rates for the horseshoe." To appear in Electronic Journal of Statistics. 1222, 1223, 1227, 1230, 1235, 1244

van der Pas, S., Szabó, B., and van der Vaart, A. (2017b). "Supplement to: Uncertainty quantification for the horseshoe". *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/17-BA1065SUPP. 1224, 1228, 1229, 1230, 1231, 1234, 1235

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(2): 2585–2618. MR3285877. doi: http://dx.doi.org/10.1214/14-EJS962. 1222, 1227, 1232, 1235

Zhang, C.-H. and Zhang, S. S. (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 76(1): 217–242. MR3153940. doi: http://dx.doi.org/10.1111/rssb.12026. 1223

Invited comment on Article by van der Pas, Szabó, and van der Vaart

Ismaël Castillo*

The presently discussed paper by Stéphanie van der Pas, Botond Szabó and Aad van der Vaart is a third of a series of very interesting works on convergence properties of posterior distributions associated to the horseshoe prior in the sparse normal means model. The horseshoe prior distribution as considered in the paper is a specific scale mixture of normal distributions. Given τ, it is the distribution of θ_1 obtained from

$$\theta_1 \,|\, \lambda, \tau \sim \mathcal{N}(0, \lambda^2\tau^2), \qquad \lambda \sim C^+(0,1). \tag{1}$$

Convergence rates are obtained in van der Pas et al. (2014) and adaptive counterparts are derived in van der Pas et al. (2017). In the present paper the authors make an important step further and study uncertainty quantification: they demonstrate that under certain conditions credible sets derived from the horseshoe posterior distribution, either local marginal credible intervals or global ℓ^2 credible balls, can be used as confidence sets, asymptotically in the number of observations. This is, after Belitser and Nurushev (2015), one of the first works on the subject using Bayesian methods in sparse settings.

I really enjoyed reading this paper and the previous ones. Below I discuss two main points and then close my discussion with a couple of more specific questions. The first comment draws some analogies with spike and slab priors with sparsity parameter calibrated by empirical Bayes (EB) and asks for possibly more general horseshoe-type distributions. In a second comment, we discuss model selection properties and credible sets for the horseshoe.

Some of the comments below are inspired by current work in progress with Romain Mismer Castillo and Mismer (2017) and Botond Szabó Castillo and Szabó (2017), in which we consider related questions for spike and slab prior distributions

$$\theta_1 \sim (1-\alpha)\delta_0 + \alpha G, \tag{2}$$

for some absolutely continuous distribution G and α calibrated by an empirical Bayes approach: following the steps of Johnstone and Silverman (2004), who studied risks of a class of point estimators derived from the EB approach, we consider the convergence of the full EB-posterior and related credible sets properties.

1. More flexible horseshoe prior distributions?

Following Carvalho et al. (2010), if $\pi(\theta_1)$ denotes the marginal density of θ_1 in (1),

$$\frac{1}{\tau}\log\left(1+\frac{4\tau^2}{\theta_1^2}\right) \lesssim \pi(\theta_1) \lesssim \frac{1}{\tau}\log\left(1+\frac{2\tau^2}{\theta_1^2}\right). \tag{3}$$

*Université Pierre et Marie Curie – Paris 6; Laboratoire Probabilités et Modèles Aléatoires (LPMA); UMR 7599; 5, place Jussieu; 75005 Paris, France, ismael.castillo@upmc.fr

This implies that the horseshoe prior, given τ, has a pole at zero and Cauchy tails. The pole at zero guarantees shrinkage of small signals while heavy tails avoid over-shrinkage of large signals.

There seems to be a striking correspondance between the tuning parameter τ of the horseshoe and the success probability α in the spike and slab prior, especially when G is taken to be a distribution with Cauchy tails. For instance, when using a marginal maximum likelihood empirical Bayes (MMLE) method to estimate α for such a spike and slab prior with Cauchy tails, one can show Castillo and Mismer (2017) (thereby slightly improving, in the case one restricts to $\ell_0[p_n]$ classes, upon the estimate from Lemma 10 and (101) in Johnstone and Silverman (2004)) the estimate $\hat{\alpha}$ is such that, as $n \to \infty$,

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{\theta_0}\left[\hat{\alpha} > (p_n/n)\sqrt{\log(n/p_n)}\right] = o(1),$$

where p_n is the sparsity parameter. This is the same as the upper boundary for τ obtained by the authors who established in van der Pas et al. (2017) that the MMLE $\hat{\tau}_n$ verifies

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{\theta_0}\left[\hat{\tau}_n > \tau(p_n)\right] = o(1),$$

which is part of Condition 1 of the present paper. This suggests that tails of the horseshoe and tails of the slab distribution play a similar role, also at level of precise conditions arising in the proofs.

This naturally leads to the question of whether it is possible to allow for other tail distributions for the marginal distribution of θ_1 for horseshoe-type priors. Another reason why we mention this is that it appears from Castillo and Mismer (2017)-Castillo and Szabó (2017) that in the spike and slab case, tails of G are particularly critical in obtaining optimal adaptive rates and confidence sets when using an empirical Bayes method. While Cauchy tails are fine in the spike and slab case when the squared ℓ^2 loss $\|\theta - \theta'\|^2 = \sum_i (\theta_i - \theta_i')^2$ is considered, they presumably lead to suboptimal rates if the loss is measured in terms of d_q-distances $d_q(\theta, \theta) := \sum_i |\theta_i - \theta_i'|^q$ (as in Castillo and van der Vaart (2012)) when $q < 1$ (we note here that we are talking about results for the full EB posterior distribution, not aspects of it such as the median or mode as in Johnstone and Silverman (2004), for which this phenomenon does not arise).

Perhaps heavier tails, such as $\theta_1^{-1-\delta}$ with $\delta < 1$ could be obtained by considering one of the other mixture priors mentioned in the paper such as the normal-exponential-gamma or the more general global-local scale mixture of normals, although we could not find any explicit results on tails of the marginal distribution in the mentioned references.

2. Model selection: 'sparsifying' the horseshoe?

By construction, a draw from the posterior distribution associated to the horseshoe prior does not set any component exactly to zero. In a sense, again by construction, the horseshoe prior is not exactly 'made for' $\ell_0[p_n]$ classes. Still, as the authors nicely prove, it leads to very good results for estimation and confidence sets for the squared ℓ^2 loss and $\ell_0[p_n]$ classes.

When one looks at a different type of results, such as model selection, or results for loss functions that are more sensible to missing the exact zeros, such as d_q–losses, something must be done, and the authors propose an additional *selection rule* to set some of the coefficients to zero.

The selection rule consists in looking at marginal credible intervals for individual coefficients θ_i and to select the given index i if the credible interval does not contain zero. This rule is very intuitive, but is there a qualitative justification of this specific choice? For instance, can something be said about its corresponding 'threshold' in the sense of the smallest signal strength that gives detection?

Part of the interesting message from Sections 2 (credible intervals) and 3 (model selection) from the paper is that, after the selection rule is applied, the resulting procedure does qualitatively something similar to what priors with a built-in selection procedure, such as spike and slab, would do: most true zero parameters are set to zero, large enough signals are always detected, while 'intermediate' signals are often set to zero.

One can wonder whether it is possible to recover some results obtained for priors with built-in selection with the horseshoe combined with the selection rule, for instance in the following two directions

a) Number of non-zero coefficients. From (i) of Theorem 3.1, it follows that the total number of selected coefficients is no larger than $p_n + (n - p_n)\gamma_n$ (I believe γ_n should be read $(n - p_n)\gamma_n$ in point (i) of the statement). The condition on γ_n implies that $n\gamma_n$ is of larger order than p_n. Could one prove that the bound is close to p_n, or rather here, say, a constant times $p_n\sqrt{\log(n/p_n)}$?

b) d_q–losses. In principle, one could also expect that, once some of the smallest coefficients of the horseshoe estimator are set to zero, the resulting 'after selection'-estimate would perform well also in terms of d_q–distances, at least for some qs in $(0, 2)$. This question arises for estimation as in van der Pas et al. (2017) but also for credible sets as in Section 4 of the present paper.

Specific questions

(i) Adaptive minimax rate with precise logarithmic term.

In the companion paper van der Pas et al. (2017), the authors obtain a nearly optimal minimax rate $Cp_n \log n$ for the horseshoe posterior, which may miss the minimax rate of the order $p_n \log(n/p_n)$ for signals that are nearly dense (e.g. $p_n = n/\log n$ or $p_n = n/e^{\sqrt{\log n}}$). It would be interesting to see whether the precise logarithmic term can be obtained.

(ii) Simulations.

In principle, when looking at classes of sparse vectors that do not specifically contain zeros, such as strong or weak ℓ^p classes $(0 < p < 2)$, the horseshoe estimator should perform even better, in the sense that it is not 'penalised' by the fact of not setting some coefficients to zero. Did the authors do some simulations in this type of setting?

Also, how does one choose in practice the blow-up factor L of the credible intervals or credible balls? Is there a recommended rule to chose it in simulations?

References

Belitser, E. and Nurushev, N. (2015). "Needles and straw in a haystack: empirical Bayes confidence for possibly sparse sequences." Preprint. 1250

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. 1250

Castillo, I. and Mismer, R. (2017). "Empirical Bayes analysis of Spike and Slab posterior distributions." In preparation. 1250, 1251

Castillo, I. and Szabó, B. T. (2017). "Spike and Slab empirical Bayes sparse credible sets." In preparation. 1250, 1251

Castillo, I. and van der Vaart, A. W. (2012). "Needles and straw in a haystack: posterior concentration for possibly sparse sequences." *Annals of Statistics*, 40(4): 2069–2101. 1251

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in a haystacks: empirical Bayes estimates of possibly sparse sequences." *Annals of Statistics*, 32(4): 1594–1649. 1250, 1251

van der Pas, S., Kleijn, B., and van der Vaart, A. (2014). "The horseshoe estimator: posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8(2): 2585–2618. 1250

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). "Adaptive posterior contraction rates for the horseshoe." *Electronic Journal of Statistics*, 11(2): 3196–3225. 1250, 1251, 1252

Invited comment on Article by van der Pas, Szabó, and van der Vaart

Ryan Martin*

1 Introduction

Since its first appearance, not quite 10 years ago, the horseshoe prior has come a very long way. When I first learned about it as a graduate student in 2008,[1] I remember thinking that building a hierarchical model around a prior with both a huge spike at zero and very heavy tails was a clever way to induce this kind of shrinkage needed for structured high-dimensional problems. What I didn't realize at the time was that the horseshoe was more than just a clever idea; it was a *game-changer*, motivating the now very active research on general classes of global–local priors. While I may have missed my chance in 2008 to get in on the ground floor of the horseshoe enterprise, it is great to have the opportunity here in 2017 to reflect a bit on these developments.

The normal mean model discussed here has $X_i \sim N(\theta_i, \sigma^2)$, $i = 1, \ldots, n$, independent, and the goal is inference on the mean vector $\theta = (\theta_1, \ldots, \theta_n)$. Let Π^n denote the posterior distribution for θ based on the horseshoe prior, with either a plug-in estimator or a hyper-prior for the global scale factor τ. The theoretical questions that Bayesians are currently interested in revolve around the asymptotic concentration, as $n \to \infty$, of the posterior Π^n at certain "true" mean vectors $\theta_0 \in \mathbb{R}^n$, in particular, at θ_0 which are p-sparse. The most precise of such questions concern the properties of marginal posterior credible intervals or posterior credible balls; that is, does a credible set have the right size and the right coverage probability? The impossibility theorem (e.g., Li, 1989) says that no confidence sets—Bayesian or otherwise—are both *adaptive*, in the sense that their size corresponds to the minimax rate for the true-but-unknown sparsity level $p = o(n)$, and *honest*, in the sense that they attain the nominal frequentist coverage probability, uniformly over all θ_0. It is intuitively clear that the troublemaker θ_0's, the so-called "inconvenient truths" in Szabó et al. (2015), have too many $|\theta_{0,i}|$ close to a certain detectability boundary. The authors of the present paper, namely, van der Pas, Szabó, and van der Vaart, are to be congratulated for their efforts to give a precise mathematical characterization of these troublemakers in the context of the horseshoe model. A natural next step would be to develop similar results for other horseshoe-like models, such as those in Ghosh et al. (2016) and Ghosh and Chakrabarti (2017). While I am excited to see how far these efforts can go, I also have some reservations.

For my discussion here, I'll focus on two higher-level points, namely, the choice of a one-group model, like the horseshoe, versus a two-groups model, and coming to terms with the inevitable dishonesty of posterior credible sets.

*Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu

[1]The 2008 technical report I read is, as of the time I'm writing this, still available on Prof. Polson's website: http://faculty.chicagobooth.edu/nicholas.polson/research/papers/Horse.pdf.

2 One group or two?

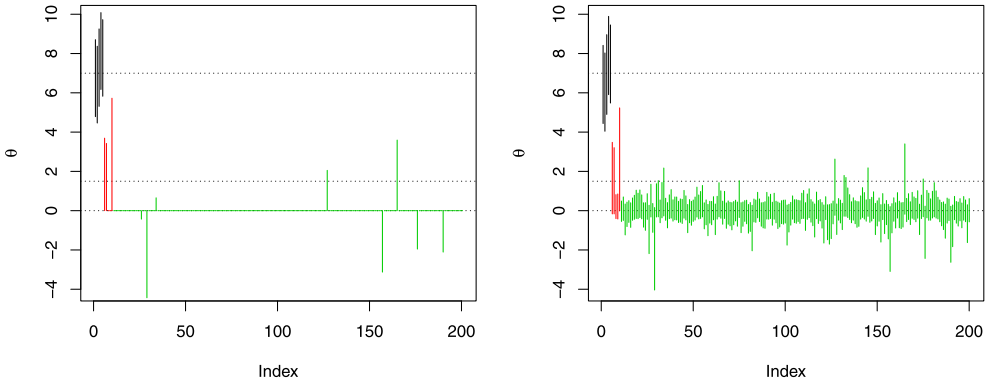
The horseshoe is an example of a one-group model since it does not treat the zero and non-zero θ_i's differently *a priori*. A two-groups model, on the other hand, treats the zero and non-zero θ_i differently, usually with a discrete mass at zero and continuous density away from zero, respectively. If there is reason to believe that the true θ_0 is sparse in the sense that it contains many exact zeros, then the two-groups formulation is clearly more natural. But the one-group approach has an apparent computational advantage in that it does not require posterior exploration over the very large space of zero/non-zero configurations. This advantage is further boosted by the theoretical fact that a two-groups formulation with conjugate, fixed-center normal priors for the non-zero θ_i's can lead to sub-optimal posterior concentration properties (e.g., Castillo and van der Vaart, 2012, Theorem 2.8). The heavier-than-normal-tailed priors that have the desired theoretical properties, such as Laplace, make computation more expensive. Of course, the computational benefits afforded to these one-group models would be of little relevance if there was no supporting theory, but see the present paper, as well as Ghosh and Chakrabarti (2015) and van der Pas et al. (2017).

Despite the nice results for one-group priors, I still would lean towards a two-groups model in this case: the sparsity assumption is genuine prior information, "genuine" in the sense that I'm willing to make that assumption in the supporting theory, so my prior ought to allow, and even encourage, many of the means to be exact zeros. But how to address the challenges presented in the previous paragraph? It turns out that both the computational and theoretical obstacles can be overcome by taking a conjugate normal prior for the non-zero means but with an informative choice of center. In the context of this normal means model, Martin and Walker (2014) suggest centering the (conditional) prior for the non-zero means at the observations; that is,

$$(\theta_i \mid \theta_i \neq 0) \sim N(\mu_i, \gamma^{-1}),$$

where μ_i is chosen to be X_i and $\gamma > 0$ is some (small) fixed constant. Then, of course, a prior is assigned to the configurations of zero and non-zero means, which is where the sparsity assumption is incorporated. This prior has some intuitive appeal: we are taking an informative prior on the configuration, the aspect of θ that we have genuine prior information about, and a data-driven "non-informative" prior on the actual values of the non-zero means, about which we have no genuine prior information. From a computational point of view, the data-dependence does not affect conjugacy, so posterior computations are relatively simple. Theoretically, Stephen and I showed that the corresponding "empirical Bayes" posterior has the same adaptive and asymptotically minimax concentration rate as has been demonstrated for the horseshoe. I won't say any more here about our specific formulation, but I'll note that Stephen and I have since extended these results to sparse high-dimensional regression (Martin et al., 2017) and even to some nonparametric problems (Martin and Walker, 2017).

To conclude this part of my discussion, I want to revisit van der Pas, Szabó, and van der Vaart's example in Section 2.2 of their paper. There they have $n = 200$ means, $p = 10$ of which are non-zero, with five means equal to 7 and the other five equal to 1.5. The point is that 7 easily exceeds their detection boundary but 1.5 is right near



(a) Two-group empirical Bayes (b) One-group horseshoe

Figure 1: Marginal equal-tailed 95% credible intervals based on (a) the two-groups empirical Bayes approach in Martin and Walker (2014) and (b) the horseshoe, for the same experiment summarized in Figure 1 of van der Pas, Szabó, and van der Vaart.

it; therefore, according to their theory, the means equal to 1.5 ought to be difficult to cover. For comparison, Figure 1 shows the naive equal-tailed 95% marginal empirical Bayes credible intervals from the same example but based on the two-groups empirical Bayes formulation in Martin and Walker (2014) and our simple Gibbs sampler. Just like van der Pas, Szabó, and van der Vaart, we cover all the zero means (green) and all the size-7 means (black), but miss a couple of the size-1.5 means (red). However, some of the marginal credible intervals are actually *singletons*, whereas the horseshoe always returns intervals. I wonder if the two-groups formulation has some efficiency advantage over the one-group version in the sense that the coverage rates ought to be the same but the former's "overall length" is shorter? In this particular instance, the average length of the 200 one- and two-groups intervals were about 1.5 and 0.25, respectively.

To be fair, coverage properties for our credible sets have not yet been worked out, but given the results in Belitser and Nurushev (2017), we have every reason to think that this can be done, and we'll report these details elsewhere.

3 "Honesty is the best policy"

My Google search results attribute this saying to Benjamin Franklin, and its long lifetime is, I think, a testament to the quality of the advice. I remember my parents telling me this as a child. But I didn't expect it to be relevant in both life and statistics!

Indeed, just like life presents circumstances where we, as individuals, have to choose between honesty and dishonesty, the impossibility theorem says that we, as statisticians, must choose between honesty and adaptation. These adaptation properties are mathematically elegant, of practical importance in terms of efficiency of estimation, and, frankly, fun to work out. Having a characterization of those troublemaker param-

eter values at which adaptivity holds but honesty fails, as van der Pas, Szabó, and van der Vaart, among others, have given, is theoretically valuable, primarily for the insights it provides. Unfortunately, these insights don't translate to practical guidance; for example, for fixed n, it's impossible to tell if a particular θ_0 satisfies the excessive-bias restriction. Moreover, it's exactly those "intermediate" parameters carved about by the theorem's conditions for which a precise uncertainty quantification is needed. In any case, I think many users of Bayesian methods are sold by the often-spoken but rarely-written claim that "Bayes provides automatic uncertainty quantification." But the impossibility theorem says that if the posterior is good, i.e., adaptive, then this rationale breaks down. Of course, the impossibility theorem applies to Bayes and non-Bayes approaches alike, but Bayes isn't needed to construct an adaptive estimator so, if the posterior doesn't provide honest uncertainty quantification, then what does it have to offer? Call me over-dramatic, but does taking the "honesty is the best policy" advice, as I think we should, require a change of perspective? Our theoretical efforts thus far have focused primarily on adaptation-driven priors; is the next *game-changer* a class of honesty-driven priors? If so, then what will be the horseshoe's fate?

To conclude, I want to mention that this conflict between concentration properties and uncertainty quantification is not unique to the type of high-dimensional problems considered in van der Pas, Szabó, and van der Vaart's paper and the ensuing discussion. For example, even in apparently simple fixed-dimensional problems, Fraser (2011) and Fraser et al. (2016) describe situations where Bayesian uncertainty quantification is less than fully satisfactory. More generally, there are known concerns about uncertainty quantification via the marginal posterior distribution for non-linear interest parameters, including the extreme cases in Gleser and Hwang (1987) where marginal posterior credible intervals could have zero coverage probability for all n. My co-authors and I have been writing on these points recently (e.g., Martin and Liu, 2016b; Balch et al., 2017; Martin, 2017), advocating for a stronger property called *validity* that can stand up to these problematic cases. Efforts to develop a framework for valid statistical inference are underway (e.g., Martin and Liu, 2016a), and I am excited to see how the new approach can handle these high-dimensional problems.

References

Balch, M. S., Martin, R., and Ferson, S. (2017). "Coverage probability fails to ensure reliable inference." arXiv:1706.08565. 1257

Belitser, E. and Nurushev, N. (2017). "Needles and straw in a haystack: robust confidence for possibly sparse sequences." arXiv:1511.01803. 1256

Castillo, I. and van der Vaart, A. (2012). "Needles and straw in a haystack: posterior concentration for possibly sparse sequences." *Annals of Statistics*, 40(4): 2069–2101. MR3059077. doi: http://dx.doi.org/10.1214/12-AOS1029. 1255

Fraser, D. A. S. (2011). "Is Bayes posterior just quick and dirty confidence?" *Statistical Science*, 26(3): 299–316. 1257

Fraser, D. A. S., Bédard, M., Wong, A., Lin, W., and Fraser, A. M. (2016). "Bayes, reproducibility and the quest for truth." *Statistical Science*, 31(4): 578–590. 1257

Ghosh, P. and Chakrabarti, A. (2015). "Posterior concentration properties of a general class of shrinkage estimators around nearly black vectors." arXiv:1412.8161. 1255

Ghosh, P. and Chakrabarti, A. (2017). "Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems." *Bayesian Analysis*, to appear. doi: http://dx.doi.org/10.1214/16-BA1029. 1254

Ghosh, P., Tang, X., Ghosh, M., and Chakrabarti, A. (2016). "Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity." *Bayesian Analysis*, 11(3): 753–796. 1254

Gleser, L. J. and Hwang, J. T. (1987). "The nonexistence of $100(1-\alpha)\%$ confidence sets of finite expected diameter in errors-in-variables and related models." *Annals of Statistics*, 15(4): 1351–1362. 1257

Li, K.-C. (1989). "Honest confidence regions for nonparametric regression." *Annals of Statistics*, 17(3): 1001–1008. 1254

Martin, R. (2017). "A mathematical characterization of confidence as valid belief." arXiv:1707.00486. 1257

Martin, R. and Liu, C. (2016a). *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. 1257

Martin, R. and Liu, C. (2016b). "Validity and the foundations of statistical inference." arXiv:1607.05051. 1257

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. 1255

Martin, R. and Walker, S. G. (2014). "Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector." *Electronic Journal of Statistics*, 8(2): 2188–2206. 1255, 1256

Martin, R. and Walker, S. G. (2017). "Empirical priors for target posterior concentration rates." arXiv:1604.05734. 1255

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). "Frequentist coverage of adaptive nonparametric Bayesian credible sets." *Annals of Statistics*, 43(4): 1391–1428. 1254

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). "Adaptive posterior contraction rates for the horseshoe." *Electronic Journal of Statistics*, 11(2):3196–3225. 1255

Acknowledgments

This work is partially supported by NSF grant DMS–1507073.

Invited comment on Article by van der Pas, Szabó, and van der Vaart

Nicholas G. Polson*

Let me first congratulate the authors on an impressive paper that solves an open problem on uncertainty quantification for the horseshoe estimator. These are challenging problems and of great importance to our understanding of uncovering sparse signals. Most of my comments are based on their main result (Theorem 2) and the ensuing Figure 1 which illustrates the marginal credible sets in a simple simulation.

The sparse normal means problem is concerned with inference for the parameter vector $\theta = (\theta_1, \ldots, \theta_p)$ where we observe data $y_i = \theta_i + \epsilon_i$ where the level of sparsity might be unknown. From both a theoretical and empirical viewpoint, regularised estimators have won the day. This still leaves open the question of how does specify a penalty, denoted by π_{HS}, (a.k.a. log-prior, $-\log p_{HS}$)? Lasso simply uses an L^1-norm, $\sum_{i=1}^{K} |\theta_i|$, as opposed to the horseshoe which (essentially) uses the penalty

$$\pi_{HS}(\theta_i|\tau) = -\log p_{HS}(\theta_i|\tau) = -\log\log\left(1 + \frac{2\tau^2}{\theta_i^2}\right). \tag{1}$$

The motivation for the horseshoe penalty arises from the analysis of the prior mass and influence on the posterior in **both** the tail and behaviour at the origin. The latter is the key determinate of the sparsity properties of the estimator. See Bhadra et al. (2017) for a recent review that compares and contrasts Lasso and Horseshoe. The "choice" of τ depends on how much one is willing to assume *a priori* about the sparsity properties of the underlying vector. Among other things, the authors propose a marginal maximum likelihood estimator (MMLE), defined in (8). This has been discussed by many authors, e.g. Gelman (2006), personally I like to relate this to a similar problem in the Bayesian analysis of variance, see Tiao and Tan (1965), Stein (1969) which I discuss below.

From an applied perspective, many of the authors' results can be inferred from their Figure 1. This illustrates their main theoretical result in Theorem 2.4. The marginal credible sets for uncovering the parameter vector are shown for a single simulated normal means problem with $n = 200$ and $p = 10$ non-zero coordinates. The true means are taken to be $0, 1.5, 7$, corresponding to the three regions analyzed theoretically to provide bounds in Theorem 2. As predicted by theory, all the means equal to 7 are recovered nicely—as an aside, this would no be true for lasso.The sparse zeroes are also recovered by design. The "honesty" of the estimator, as defined by the authors, can be seen by behavior at recovering the means equal to 1.5 where only 2 out of 5 succeeded for this particular simulation. Lasso would have done a better job! As there is no free lunch for admissible estimators, maybe this result is not that surprising. Assessing the magnitudes (a.k.a. uncertainty quantification) is the goal of the authors' analysis.

*5807 S Woodlawn Avenue, Chicago, IL 60637, USA, ngp@chicagobooth.edu

From a historical perspective, James–Stein (a.k.a L^2-regularisation) is only a global shrinkage rule—there are no local parameters—to learn about sparsity. A simple sparsity example shows the issue with L^2-regularisation. Consider the sparse r-spike problem where focusing solely on rules with the same global shrinkage weight (albeit benefiting from pooling of information) has an issue. Let the true parameter value be $\theta_p = (\sqrt{d/p}, \ldots, \sqrt{d/p}, 0, \ldots, 0)$. James–Stein is equivalent to the hierarchical model

$$y_i = \theta_i + \epsilon_i \text{ and } \theta_i \sim \mathcal{N}\left(0, \tau^2\right) \text{ all } \lambda_i \equiv 1.$$

This dominates the plain MLE but loses admissibility! This is due to the fact that a "plug-in" estimate of global shrinkage $\hat{\tau}$ is used. Tiao and Tan's original "closed-form" analysis is particularly relevant here. They point out that the mode of $p(\tau^2|y)$ is zero exactly when the shrinkage weight turns negative (their condition 6.6). From a risk perspective $E\|\hat{\theta}^{JS} - \theta\| \leq p, \forall\theta$ showing the inadmissibility of the MLE. At origin the risk is 2, **but!**

$$\frac{p\|\theta\|^2}{p + \|\theta\|^2} \leq R\left(\hat{\theta}^{JS}, \theta_p\right) \leq 2 + \frac{p\|\theta\|^2}{p + \|\theta\|^2}$$

implying $R(\hat{\theta}^{JS}, \theta_p) \geq (p/2)$. Simple thresholding rule beats this with a risk, $\sqrt{\log p}$. This simple historical example merely shows that the choice of penalty should not be taken for granted. The same thought applies to lasso—the credible sets implicit in lasso are not optimal and the horseshoe approach achieves much large gains both theoretically and empirically—which the optimality properties and caveats revealed by the authors' paper.

There are still many fruitful areas of research in Bayesian sparsity and horseshoe estimation. One avenue is to further understand Tiao and Tan's (1965) condition for the posterior on τ to have a mode at the origin in the case of sparsity.

On the empirical side, it is pleasing to see many `R` packages available to implement horseshoe estimation in a variety of situations, see Bhadra et al. (2017) for further discussion. Bhattacharya et al. (2016) provide one such package and also shows how horseshoe can vastly outperform lasso in typical applied contexts. These packages are closing the gap on computational speed that lasso enjoys. On the theoretical side, hyper-parameter selection stills seems an interested area to me where many of the old discussions are still relevant. Recent applications of these methods are in evermore complicated models, such as deep learning (Polson and Sokolov, 2017), and understanding the theoretical underpinnings is as important as ever.

References

Bhadra, A., J. Datta, N. G. Polson and B. T. Willard (2017). Lasso Meets Horseshoe. arXiv:1706.10179. 1259

Bhadra, A., J. Datta, N. G. Polson and B. T. Willard (2017). Horseshoe Regularisation for Feature Subset Selection. arXiv:1702.07440. 1260

Bhattacharya, A., A. Chakraborty and B. K. Mallick (2016). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *Biometrika*, 103(4), 985–991. MR3620452. doi: http://dx.doi.org/10.1093/biomet/asw042. 1260

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515–534. 1259

Polson, N. G. and V. O. Sokolov (2017). Deep Learning: A Bayesian Perspective. arXiv:1706.00473. 1260

Stein, C. (1969). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, 16, 155–160. 1259

Tiao, G. C. and W. Y. Tan (1965). Bayesian Analysis of random effects in the analysis of variance I: Posterior distribution of variance components. *Biometrika*, 52, 37–53. 1259, 1260

Contributed comment on Article by van der Pas, Szabó, and van der Vaart

William Weimin Yoo*

Abstract. We begin by introducing the main ideas of the paper under discussion. We discuss some interesting issues regarding adaptive component-wise credible intervals. We then briefly touch upon the concepts of self-similarity and excessive bias restriction. This is then followed by some comments on the extensive simulation study carried out in the paper.

Keywords: horseshoe, half-Cauchy, MMLE, credible intervals, model selection, credible balls, adaptive, excessive bias restriction.

I would like to congratulate the authors for such an comprehensive and interesting paper on the horseshoe prior and its use in Bayesian uncertainty quantification. Let me first summarize key ideas of van der Pas et al. (2017b) in order to set the tone for my discussion. The horseshoe prior is a scale mixture of normals with the half-Cauchy as the mixing distribution. The half-Cauchy is in turn the absolute value of a standard Cauchy distribution. Working under the normal means model $Y_i = \theta_i + \varepsilon_i, i = 1, \ldots, n$, this hierarchical prior takes the form $\theta_i|\lambda_i, \tau \sim \mathrm{N}(0, \lambda_i^2\tau^2)$ and $\lambda_i \sim$ Half-Cauchy. The paper considers two methods of estimating τ, namely by empirical Bayes through the maximum marginal likelihood estimator (MMLE) or by endowing another layer of hyper-prior.

The true signal θ_0 is assumed to be sparse, and the task of recovering these nonzero values using the horseshoe prior was studied by the same authors previously in van der Pas et al. (2017a). The present paper however deals with the issue of accessing the quality of this recovery procedure. This is accomplished through the construction of (adaptive) component-wise credible intervals and ℓ_2-credible balls. Moreover, the authors introduced a simple model selection procedure by declaring that a signal component is unimportant if its corresponding credible interval includes 0 within its span.

My discussion will focus on the case of component-wise credible interval, since I find that its results are the most interesting and these intervals are the ones used in the simulations. The most prominent feature regarding results for these intervals is the division of the true signal components into three regimes corresponding to small, intermediate and large signals. The horseshoe interval was able to provide adequate coverage for small and large signals but not the intermediate ones. The existence of this intermediate layer and the gaps in between these regimes made me wonder whether this is due to the intrinsic nature of component-wise credible intervals, or other more extrinsic factors such as the horseshoe prior used or perhaps an artefact of the proof techniques employed.

*Mathematical Institute, Leiden University, The Netherlands, yooweimin0203@gmail.com

It is now well known that adaptive credible sets cannot do honest uncertainty quantification over all possible true signals, and some of these signals must be permanently excluded. To this end, the authors discussed two criteria for removal, one based on the concept of self-similarity and the other based on the excessive bias restriction introduced by Belitser and Nurushev (2017). In the present setting of sparse signals, the key insight into these conditions is that the true signals must be at some distance away from the zero signal in a sense made precise in the paper. Interestingly as mentioned in Remark 3, the three regimes become more "contiguous" under self-similarity when compared to the situation where self-similarity was not assumed, as this is evident by comparing $\mathcal{S}_a, \mathcal{M}_a, \mathcal{L}_a$ with $\mathcal{S}, \mathcal{M}, \mathcal{L}$ for $\tau = \tau_n(p_n)$. This in turn suggests that throwing away troublesome truths enables the horseshoe credible interval to fill in the gaps between the three regions.

For the sake of discussion, let us continue working under the self-similarity or excessive bias restriction. From the simulation results, it is clear that the horseshoe credible intervals have the best performance in terms of high coverage and shortest lengths when the means (or the true signals) are zero. Two settings of p_n the number of nonzero signals were used, i.e., $p_n = 20$ and $p_n = 200$ when $n = 400$. Now let us increase the proportion of zero means $(n - p_n)$ to a point that the self-similarity condition is violated, will the horseshoe still enjoy this near-perfect performance? By looking at the bar charts on coverage and lengths, it is conceivable that we can still get good performance even if this condition is violated slightly. My question concerns whether it is possible to observe empirically what will happen when sparse signals become non self-similar in the sense discussed in the paper.

Uncertainty quantification is undoubtedly one of the most active research areas in Bayesian statistics, and as the present paper shows, it involves resolving many delicate technical and practical issues. The horseshoe prior has proven itself to be optimal in sparse signal recovery, and we are able to access the quality of this recovery thanks to the theories and methods developed in the paper. It would be interesting also to consider other classes of priors, e.g., spike-and-slab types, and I hope that there will be more papers on Bayesian uncertainty quantification for sparse models in the future.

References

Belitser, E. and Nurushev, N. (2017). "Needles and straw in a haystack: robust confidence for possibly sparse sequences." arXiv preprint: 1511.01803. 1263

van der Pas, S., Szabó, B., and van der Vaar, A. (2017a). "Adaptive posterior contraction rates for the horseshoe." *Electronic Journal of Statistics*, 11(2): 3196–3225. MR3705450. doi: http://dx.doi.org/10.1214/17-EJS1316. 1262

van der Pas, S., Szabó, B., and van der Vaart, A. (2017b). "Uncertainty quantification for the horseshoe." *Bayesian Analysis*, 1–29. Advance publication. 1262

Contributed comment on Article by van der Pas, Szabó, and van der Vaart

Juho Piironen*, Michael Betancourt†, Daniel Simpson‡, and Aki Vehtari§

The authors present a detailed analysis of the asymptotic frequentist properties of credible sets derived from posteriors with normal-linear measurement models and horseshoe priors. Although we disagree with the claim that *"In Bayesian practice credible balls are nevertheless used as if they were confidence sets"*, the results in the paper are important for identifying where the horseshoe priors are fragile asymptotically, and hence particularly dangerous in the non-asymptotic regimes more typical in the applied problems where sparse models are needed.

One clarification we believe is warranted is that the horseshoe family of prior distributions does not encode sparsity as is typically interpreted. Instead of partitioning parameters into those that are zero and non-zero, the horseshoe priors actually separate parameters into those that are resolvable by measurements and those that are not. In particular, as with any model the horseshoe priors cannot be interpreted outside of the context of a particular likelihood (Gelman et al., 2017). Consequently the statement that *"τ can be interpreted as the proportion of nonzero parameters, up to a logarithmic factor"* is not quite true.

Piironen and Vehtari (2017b; 2017c) demonstrate that the effects of τ in horseshoe priors are intimately related to the measurement variability σ, even for the simple normal-linear measurement model. Figure 6 of Piironen and Vehtari (2017c), for example, clearly illustrates that rescaling the data changes the impact of the horseshoe prior unless τ is scaled by σ, even with an oracle prior information about the true number of significant parameters, $p_0 = p_n$. In particular, the resolution threshold $\sqrt{2\log(n/p_n)}$ arising in the paper implicitly assumes that the measurement variability σ is equal to 1, but a more realistic threshold has to take into account the value of σ, which is typically unknown a priori. We are very curious as to how robust the results presented in the paper are to these circumstances where also σ must be inferred.

Additionally, we find that the focus on marginal credible intervals is a significant limitation. One of the defining features of the family of horseshoe priors, and indeed a strong reason for their utility, is that they do not regularize each parameter independently but rather induce a joint regularization over the entire parameter space. In particular, joint credible intervals can behave much differently from marginal intervals. Figure 1 illustrates that with a linear model that employs a uniform prior over the slopes of two correlating predictors x_1 and x_2 it may happen that the joint posterior concentrates away from the origin without either of the marginals clearly distinguished from zero.

*Dept. of Computer Science, Aalto University, Finland, Juho.Piironen@aalto.fi
†Dept. of Statistics, Columbia University, New York, NY, betanalpha@gmail.com
‡Dept. of Statistical Sciences, University of Toronto, Canada, dp.simpson@gmail.com
§Dept. of Computer Science, Aalto University, Finland, Aki.Vehtari@aalto.fi

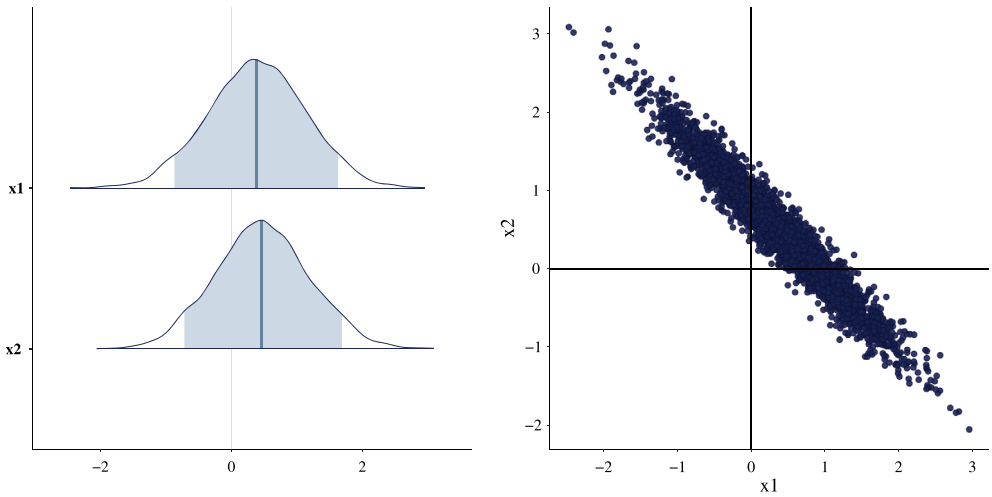


Figure 1: The left plot shows marginal posteriors of effects which overlap zero. The right plot shows the corresponding joint distribution which reveals strong posterior dependency and the fact that zero is not included in the joint credible region.

The situation becomes even more difficult with a large number of correlating predictors when utilizing the horseshoe prior. In this case even for the most relevant variables most of the posterior mass can concentrate around zero, see for example Figure 9 in Piironen and Vehtari (2017c), which makes a reliable variable selection based on the posterior intervals challenging. Moreover, the method by Carvalho et al. (2010) of including all variables with $\kappa_j > 1/2$ would fail in this case because none of the predictors have $\kappa_j > 1/2$. Consequently, we believe the only reliable variable selection strategy in these situations is based on the estimated effect on the predictive distribution, for example using the projection predictive variable selection (Piironen and Vehtari, 2017a). This framework has the added benefit that it provides guidance on how to select out significant parameters jointly, instead of one by one as discussed in the paper.

Finally, we advise caution with regard to the recommendation of the maximum marginal likelihood estimator (MMLE) for τ in practical problems. The large p, small n applications where horseshoe priors are most needed lie far away from the asymptotic regime that stabilizes the MMLE. Any complexity of the measurement model beyond the normal-linear model only makes the matter worse.

References

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. 1265

Gelman, A., Simpson, D., and Betancourt, M. (2017). "The prior can generally only be understood in the context of the likelihood." *arXiv preprint arXiv:1708.07487*. 1264

Piironen, J. and Vehtari, A. (2017a). "Comparison of Bayesian predictive methods for model selection." *Statistics and Computing*, 27(3): 711–735. 1265

Piironen, J. and Vehtari, A. (2017b). "On the hyperprior choice for the global shrinkage parameter in the horseshoe prior." In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 905–913. 1264

Piironen, J. and Vehtari, A. (2017c). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*. Accepted for publication. arXiv preprint arXiv:1707.01694. 1264, 1265

Contributed comment on Article by van der Pas, Szabó, and van der Vaart

Eduard Belitser* and Nurzhan Nurushev†

We congratulate the authors for this very interesting article focused on the frequentist coverage of the credible sets resulting from the horseshoe prior in the sparse multivariate normal means model both for the empirical and hierarchical Bayes approaches. In paper Belitser and Nurushev (2015) (last version from 2017), we studied the empirical Bayes approach to the same problem but using a different (mixture of normals) prior. For brevity, we refer to the discussion paper as paper PSV and our paper as paper BN. We skip (because of space limitations) some computations that might be needed to back up some claims we state below, these can be provided upon request to an interested reader.

The main results of PSV are adaptive over the sparsity scale within a grand space $\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n 1\{\theta_i \neq 0\} \leq p_n\}$ for some $p_n = o(n)$. Although this excludes some "almost sparse" parameters that are formally non-sparse (with many very small, but nonzero, entries), this is a mild assumption (as p_n is not assumed to be known) and in fact necessary to ensure the asymptotic regime $n \to \infty$ considered in PSV. In BN we obtain local results without relating to any sparsity scale, e.g., the true parameter θ may be not $\ell_0[p_n]$-sparse at all. For example, as a consequence we derive the results not only for $\ell_0[p_n]$, but also for other sparsity scales, such as *weak* ℓ_s*-balls* $m_s[p_n]$. Besides, in BN we derive local non-asymptotic exponential concentration bounds, which give a refined characterization of the quality of coverage and size relation results (finer, than, e.g., Theorem 5 from PSV, which is asymptotic in $n \to \infty$) and allow subtle analysis for various asymptotic regimes. We should mention that the derivation of our somewhat stronger results in BN relies on certain explicit posterior expressions resulting from our choice of prior (mixture of normals, although the model is not assumed to be normal), whereas the horseshoe prior studied in PSV leads to only implicit posterior quantities so that the authors had to overcome complicated technical issues in the proofs.

It is well known and understood that in the studied model it is impossible to construct a confidence set simultaneously with a good coverage and optimal size adaptively to sparsity scales. Insisting on the uniform coverage would necessarily lead to a "big" size of resulting confidence set (therefore uninteresting, although optimal among all sets of uniform coverage), while pursuing the optimal size results in bad coverage for some "deceptive" parameters. In the both papers, PSV and BN, the second situation is studied, where the non-deceptive parameters are described by the so called "excessive-bias restriction" (EBR). This condition was introduced in BN and slightly weakened in PSV. Let us give an intuition behind the EBR. For $\sigma^2 = 1$ and any $\theta \in \mathbb{R}^n$, the oracle rate introduced in EBR is $\min_{I \subseteq [n]} G(I) = G(I_o)$ where $G(I) = G(I, \theta) = \sum_{i \in I^c} \theta_i^2 + \tau |I| \log(en/|I|) = B(I^c) + V(I)$. It is always smaller than the minimax rate

*Department of Mathematics, VU Amsterdam, e.n.belitser@vu.nl
†Department of Mathematics, VU Amsterdam, n.nurushev@vu.nl

over any sparsity class containing θ: $G(I_o, \theta) \lesssim p_n \log(en/p_n)$ for all $\theta \in \ell_0[p_n]$. Think of I_o as the set of the (oracle) significant coordinates, $B(I_o^c)$ as the approximation term (or "bias") and $V(I_o)$ as the complexity term (or "variance") of the oracle rate. Because of the non-asymptotic study, we need $\log(en/|I_o|)$ instead of $\log(n/|I_o|)$ (as compared with PSV) in the "variance" $V(I_o)$, which are of the same order if $|I_o| \le p_n = o(n)$.

The EBR from BN basically means that the bias of the oracle rate is of the order of the variance: $B(I_o^c) \le tV(I_o)$ for some $t > 0$. Consider now the weakened version of the EBR introduced in PSV. Since $p_n = o(n)$, without loss of generality one can set $C_s = 1$ in the EBR condition of PSV. If the EBR from BN is fulfilled, then the EBR from PSV (the two relations of display (16)) is also fulfilled which is in our notation as follows: for $\tau > 2$, $C > 0$, $B(I_o^c) \le CV(I_o)$ and $\#\{i : \theta_i^2 \ge \tau \log(n/|I_o|)\} \ge |I_o|$. The first relation is exactly the EBR from BN and the second relation is implied by the definition of the oracle I_o. The EBR from PSV is based on the fact that it is always possible to enlarge the oracle set I_o to the *smallest* $I_* \supseteq I_o$ such that the relation $B(I_*^c) \le CV(I_*)$ is fulfilled. At worst, I_* is the support of θ: $\{i \in [n] : \theta_i \ne 0\}$. It is easy to show that $V(I_*) \asymp G(I_*) \asymp G(I_o)$, i.e., the "variance" is the main term in the rate $G(I_*)$ that is of the oracle rate order. Then the EBR in PSV gives $\tilde{p} \asymp |I_*|$ and the size $\tilde{p} \log(n/\tilde{p}) \asymp G(I_*)$ of the confidence set $\hat{C}_n(L)$ in Theorem 5 is actually of the oracle rate order. The first relation of the EBR from PSV (in display (16)) can always be made satisfied, if not for the oracle set I_o, then for an enlarged version I_* of it. But this does not mean that the problem of adaptive confidence is solved without any price: "there is no free lunch". The point is that the second relation $\#\{i : \theta_i^2 \ge \tau \log(n/|I_*|)\} \ge |I_*|$ in display (16) from PSV is not automatically fulfilled for I_* as it was for the oracle set I_o. It is this relation that becomes the actual restriction on the parameter space, and the (uniform) lower bound on the constant τ is crucial here. In paper PSV, $\tau > 2$ and this has to do with the fact that $\varepsilon_i \overset{\text{ind}}{\sim} N(0,1)$. In BN, we assume that the error vector ε satisfies only certain mild *exchangeable exponential moment condition* (allowing non-normal and dependent coordinates) and the constant τ depends on the parameters of that condition. The main massage here is that the second relation in (16) must hold for some sufficiently large (depending on the distribution of ε) constant τ.

Motivated by the above discussion, we further weaken the EBR condition.
EBR CONDITION. For some (fixed) $C > 0$ and $\varrho \in (0,1)$ there exists a $k \in [n]$ such that $\sum_{i=1}^{n-k} \theta_{(i)}^2 \le Ck \log(en/k)$ and $\sum_{i=n-\lfloor \varrho k \rfloor + 1}^{n} \theta_{(i)}^2 \ge (1-\varrho)\tau k \log(en/k)$ for sufficiently large τ, where $\theta_{(1)}^2 \le \theta_{(2)}^2 \le \ldots \le \theta_{(n)}^2$. Let i_* be the smallest possible such k.

The above EBR condition follows from the EBR version of PSV with $i_* = \tilde{p}$ and $G(I_*) \asymp i_* \log(en/i_*) \asymp G(I_o)$ where $I_* = \{i \in [n] : \theta_i^2 \ge \theta_{(n-i_*+1)}^2\}$ becomes the structure of the parameter θ. The EBR condition ensures that the structure I_* of θ becomes "identifiable", which is the necessary ingredient in constructing the adaptive confidence sets. Under this weaker EBR version, we can prove exactly the same results as in BN with slightly modified constants.

We finish with two remarks/questions to the authors of PSV. First, the size relation for the constructed confidence set in Theorem 5 of PSV holds uniformly only over the EBR class (intersected with $\ell_0[p_n]$), whereas in BN the size relation holds uniformly

over the whole space $\mathbb{R}^n$. It would be interesting to know whether this is an artefact of the proofs or of the Bayesian procedure based on the horseshoe prior. Second, it might be interesting to study whether the (weakest version of) EBR is minimal for the existence of adaptive confidence sets, and what would even be a proper formulation of this property.

References

Belitser, E. and Nurushev, N. (2015). "Needles and straw in a haystack: robust confidence for possibly sparse sequences." https://arxiv.org/abs/1511.01803v5. 1267

Rejoinder

Stéphanie van der Pas[*¶], Botond Szabó[†,‡¶,‖], and Aad van der Vaart[§‖]

We sincerely thank all discussants for their generous discussions.

Yoo appears to raise the question if the noncoverage of intermediate values of the parameters is intrinsic or due to the horseshoe or due to our proof. Unless our proof is in error, the third possibility can be safely discarded. As we have pointed out, there are absolute restrictions on confidence sets, which cannot be overcome by any methods, Bayesian or nonBayesian, so the horseshoe seems not to blame either. But perhaps Yoo is referring more to the small gaps between the regions of small, medium and large parameter values in our presentation. There is indeed room for further research on parameters near the boundaries of the regions. This is likely to be delicate, and may not yield essentially more insight than our current treatment, the gaps between the regions being small. Yoo also asks what would happen for a steadily increasing number of zero parameters. As shown by our results, this will depend on the values of the nonzero parameters, not just on their proportion of the whole.

Polson contrasts the LASSO and the horseshoe. We are also not a fan of the LASSO, perhaps for other reasons. As pointed out in Castillo et al. (2015), in the sparse case the Bayesian LASSO posterior, as opposed to the LASSO as a posterior mode, lacks accurate uncertainty quantification through the spread of the posterior. This fact is not surprising, as the LASSO prior does not model sparsity in any way. Polson makes a link to the James–Stein estimator and notes that the choice of penalty (i.e. prior density) has a big effect on the type of shrinkage one gets. We agree. For a sparse model one needs a prior that allows for sparsity. Both the Laplace prior of the LASSO and the Gaussian prior behind James–Stein are incapable to induce sparsity in the posterior, although the LASSO at least has a sparse mode. One might ask if a nonparametric prior, for instance a hierarchical one where the means come from a distribution that itself receives a Dirichlet process prior, or the empirical Bayes version of this, might work in both sparse and non-sparse situations. The simulation study in Koenker and Mizera (2014) seems to say that at least recovery is good. We do not know about uncertainty quantification. We note that in a true Bayesian spirit a comparison between various priors should always take into account the full posterior distribution, not just a measure of its center. Thus the prior plays a bigger role than just suggesting a penalty that is added on to the likelihood.

Martin contrasts continuous priors with a peak at zero, such as the horseshoe, with "two-group" priors that have a pointmass at zero. If many parameters are thought to be

[*]Leiden University, svdpas@math.leidenuniv.nl
[†]Leiden University
[‡]Budapest University of Technology and Economics, b.t.szabo@math.leidenuniv.nl
[§]Leiden University, avdvaart@math.leidenuniv.nl
[¶]Research supported by the Netherlands Organization for Scientific Research.
[‖]The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

exactly zero and not just "small", then such a two-group parameter is indeed attractive as a model. The exact zero values in Ryan's Figure 1 are a boon (although one gets almost the same picture by shrinking the zero intervals of the horseshoe to a point; in fact in this example this gives an even more correct result). We have shown elsewhere Castillo and Van der Vaart (2012); Castillo et al. (2015) that the recovery obtained with the spike-and-slab and horseshoe (and many other priors) is comparable, and expect the uncertainty quantification of the full posterior distributions also to be similar. (Preliminary work by Castillo appears to confirm this.) We have shown in Castillo and Van der Vaart (2012) that a heavier-tailed distribution than Gaussian for the non-zero parameters is useful here, as the Gaussian shrinks too much to zero (unless the noise variance is very small relative to the prior Gaussian variance). Martin proposes to solve the latter problem by using Gaussian "priors" centered at the observations, a device that it is similar to the one used in Belitser (2017); Belitser and Nurushev (2015). This will indeed get rid of the shrinkage effect, but is perhaps not elegant from a Bayesian point of view. Regrettably the device does not remedy the computational disadvantage of the two-group priors, which is the main motivation for continuous priors such as the horseshoe. Martin concludes with an intriguing suggestion to look for a "change of perspective" going against honesty of credible intervals. As a contribution to this discussion we like to highlight that the failures of credible sets, which we observed in function estimation Szabó et al. (2015) and in sparse estimation in the present paper, appear to be of different natures. In function estimation the problem is the extrapolation from more or less observable features of an unknown response function to clearly unobserved ones that cannot be known. The data-analyst is not forced to extrapolate; the trouble is that hierarchical Bayesian procedures (and other adaptive schemes) automatically do so. In sparse estimation parameters are shrunk to zero, with the amount of shrinkage determined by a hierarchical or other adaptive scheme. In both cases we like hierarchical procedures, even if they necessarily bring trouble for credible sets. A difference between the cases is that in sparse estimation we know the type of trouble: spurious near-zeros. A possible change-of-perspective here is to interpret credible sets in the sense of false discovery rates: a non-zero is a true discovery, but a zero may be a false nondiscovery. In the case of function estimation the distortion is much less predictable. However, the response functions for which distortion occurs, the "inconvenient truths" of Szabó et al. (2015), are perhaps so unrealistic that they can be a-priori ruled out (don't worry be happy). The "a-priori" may be taken literally here, as Szabó et al. (2015) show that the priors used there do never produce them. This seems also a difference between the two cases: the intermediate values that are shrunk too much to zero in sparse estimation, seem very real (and their location depends on sample size and error variance and in regression on number of parameters and presumably the design matrix). We feel that honesty is a good policy, and do not feel that this lofty goal is at odds with adaptivity. A honest description of the honesty of adaptive credible intervals, and admission that there is a gray area where we run into a detectability limit, offers actionable insight into the results of a data-analysis, and is surely not in conflict with a statistician's integrity.

Belitser and Nurushev give a valuable summary of their interesting paper. They are interested in finding an optimal procedure, in some sense, under minimal conditions, in some sense, whereas our interest was to study the coverage properties of a reasonable

and popular Bayesian procedure. It is interesting that the horseshoe comes close to optimality in the sense of Belitser and Nurushev, and also that their optimal procedure, although not Bayesian, uses empirical Bayesian thinking. At the end of their contribution Belitser and Nurushev note that their procedure possesses a certain uniformity property over the full parameter space (restricted to nearly black bodies), whereas our stated results on the horseshoe do not. One might indeed look further into uniformity. However, in our current formulation the parameter $\tilde{p}$ is defined by θ_0 and so without the restriction we impose our statement would not make any sense. So this is built into formulation, and not an 'artefact of our proof'.

Castillo notes parallels between the performance of the horseshoe procedure and the procedure based on the spike-and-slab prior. For instance, there is a strong correspondence between the horseshoe tuning parameter τ and the weight of the spike. We are very much looking forward to the work in progress by him and his co-authors. Conceptually the spike-and-slab is attractive, perhaps more so than the horseshoe. The performance of the recovery procedure (as opposed to the uncertainty quantification) has been studied in more detail for the spike-and-slab than for the horseshoe, for instance relative to other loss functions than the square Euclidean norm and for other test classes than the class $\ell_0[p]$ of nearly black bodies. We agree that it will be interesting to study these for the horseshoe as well. We do not know of work or simulation studies in this direction. Our qualitative justification of using marginal credible intervals as a selection rule of nonzero parameters is purely Bayesian: it seems reasonable to use posterior uncertainty in this manner. We do provide 'some justification' by studying the frequentist properties and do mention some thresholds. One might wish to refine these results. Regarding the logarithmic factors in the case of signals with very few zeros, which we do not really cover, we agree that it would be interesting to refine the results, although we note that a purely asymptotic treatment without getting hold of the constants may not be very informative about performance. The suggested refinement of the bound on the number of selected coefficients would indeed be interesting. We would also welcome results on the false discovery rate, which do not follow from our results (although some of our results have that flavour). Regarding the choice of the blow-up factors L in practice, we recommend to set them equal to 1, as we did in our simulations, as this yields the natural Bayesian credible sets. Bayesian and frequentist coverage are not the same, and we see no compelling reason to correct Bayesian credible sets for exact frequentist coverage, in particular in situations where full frequentist coverage is known to be impossible. For instance, joint credible balls will have large frequentist coverage only for parameters that satisfy an 'excessive bias condition', and the coverage probability will depend both on the constants in the latter condition and the blow-up factor. Then there is no universally good blow-up factor. For credible intervals the case appears to be more favourable in that at least the large intervals seem to have a universal constant, for which we have set a concrete value L_L. (In the final version of the paper, we have improved this to $1.1\zeta_{\gamma/2}/z_{\alpha/2} \approx 1$ if $\gamma = \alpha$ are small; in the asymptotics 1.1 can be replaced by any constant strictly bigger than 1.)

Piironen, Betancourt, Simpson and Vehtari contest our claim that τ can be interpreted as the 'proportion of nonzero parameters, up to a logarithmic factor'. Their argument against it seems to be that one can make such a statement only relative to

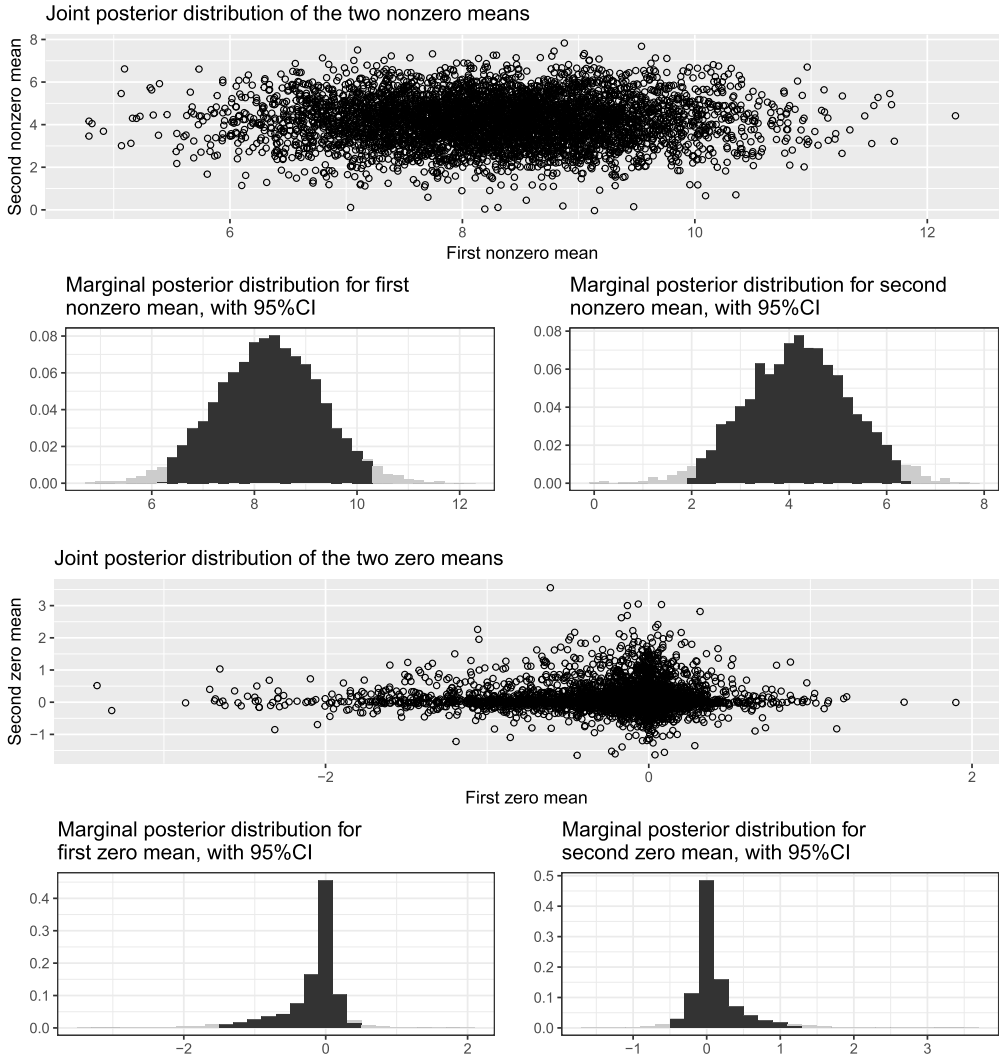


Figure 1: Scatterplots of bivariate marginal posterior distributions and histograms of the corresponding univariate marginal posterior distributions of a pair of parameters for which the true parameters are both above the detection boundary (top) or both zero (bottom). Markov Chain Monte Carlo (MCMC) samples computed in `R` with the horsehoe package.

a likelihood. That is true of course. By itself the horseshoe is even a prior for a single parameter, and it makes no logical sense to call one of its hyper parameters a proportion. Clearly the likelihood we have in mind is the one of the sequence model, and we give ample arguments for our claim, in the form of mathematical theorems backed up by simulations, referring both to recovery and uncertainty quantification. It is true that in

the paper under discussion we have set the variances of the observations equal to 1. In earlier work we used a flexible value, and this leads to the same findings as long as one scales the horseshoe prior with the error standard deviation, as is standard (and the threshold then scales with the error standard deviation as well). Piironen, Betancourt, Simpson and Vehtari have done simulations with the horseshoe prior on parameters in a regression model. This model is beyond our current paper, and similar results can only be expected if one makes strong assumptions on the design matrix, as is common in the literature ('restricted isometry', 'compatibility', etc.). This is studied in the Bayesian context for recovery in Castillo et al. (2015); we do not know of work on uncertainty quantification. The restrictions are necessary, because in high-dimensional regression models there must be collinearities between the columns of the design matrix, which the data cannot resolve, so that the posterior will load on multiple columns. We wonder if this is what happened in the figure of a bivariate posterior marginal they contributed in their discussion. We are certainly not against also considering bivariate marginals, or other projections, but in the context of our model they seem not so interesting, because the likelihood does not make the parameters interact. We confirmed this by making some pictures of bivariate posteriors in our model; examples are shown in Figure 1. Piironen, Betancourt, Simpson and Vehtari close with a warning against the marginal maximum likelihood estimator. They are not the first to do so. We can only say that we have not noted problems, not in the theory and not in the simulations. We also prefer full Bayes, but the greater efficiency may weigh in the other direction.

References

Belitser, E. (2017). "On coverage and local radial rates of credible sets." *Annals of Statistics*, 45(3): 1124–1151. URL http://dx.doi.org/10.1214/16-AOS1477 MR3662450. 1271

Belitser, E. and Nurushev, N. (2015). "Needles and straw in a haystack: empirical Bayes confidence for possibly sparse sequences." *ArXiv e-prints*. 1271

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43(5): 1986–2018. URL http://dx.doi.org/10.1214/15-AOS1334 MR3375874. 1270, 1271, 1274

Castillo, I. and Van der Vaart, A. W. (2012). "Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences." *Annals of Statistics*, 40(4): 2069–2101. MR3059077. 1271

Koenker, R. and Mizera, I. (2014). "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules." *Journal of the American Statistical Association*, 109(506): 674–685. URL http://dx.doi.org/10.1080/01621459.2013.869224 MR3223742. 1270

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). "Frequentist coverage of adaptive nonparametric Bayesian credible sets." *Annals of Statistics*, 43(4): 1391–1428. URL http://dx.doi.org/10.1214/14-AOS1270 1271