Lasso Meets Horseshoe

Anindya Bhadra
Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907-2066
bhadra@purdue.edu

Jyotishka Datta
Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701.
jd033@uark.edu

Nicholas G. Polson and Brandon Willard
The University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637
ngp@chicagobooth.edu, brandonwillard@gmail.com

Abstract

We compare the horseshoe prior, which has become the current state-of-the-art formal Bayesian procedure for sparse signal recovery problem with the popular regularisation method LASSO and some of its variants, that are often used as a gold standard for selecting the best subset of predictors. We consider two different perspectives for this problem: optimization or penalization that corresponds to finding the posterior mode under a suitable prior or the probabilistic mean that entails minimizing Bayes risk under quadratic loss under a suitable prior. We survey the major advances for both these approaches in terms of all three aspects, theoretical optimality, efficiency and scalability of computation and methodological development and performance in high-dimensional inference for the Gaussian sparse model and beyond.

Keywords: global-local priors; horseshoe; horseshoe+; lasso; regularization; sparsity.

1 Introduction

Issues

- Minimaxity
- Sparse signal recovery: nearly black objects.
- Concentration rates.
- Admissibility (Minimax not necessarily admissible : James–Stein / Ridge).
- Super-efficiency? (Hodges-Lehmann).
- Prediction

- Hyper-parameter / regularisation path (sensitivity analysis)
- Model selection

1.1 Sparsity

Define typical sparse problem. Similarity with model subset selection (AIC, BIC). Sparsity/ Model Selection / ill-posed regularization problem.

High-dimensional data has become a routine in many application areas, thanks to recent development of high-throughput devices across all major disciplines in science. This has led to an exponentially growing literature on both frequentist and Bayesian methodology and computation for large scale inference problems. While the general area is too large to cover in a single review article, and entire books have been written on the subject (see, e.g., Hastie et al., 2009), the goal of this review article is more modest, if not more specific. We revisit at least two distinct approaches to sparse parameter estimation problems, primarily from a Bayesian point of view.

It is rather obvious that the classical methods such as the maximum likelihood estimator were not designed to confront the challenges posed by multiplicity. One needs additional assumptions like sparsity of the unknown coefficients to make high dimensional statistical inference possible. Sparsity is construed as presence of a few large signals among many zero or nearly zero noisy observations. A common goal in all of the high dimensional inferential problems is to filter the low-dimensional signals observed in white noise, i.e. to identify subset of hypotheses that show significant deviation from the null hypotheses. This encompasses both estimation of the underlying sparse parameter as well as multiple testing where the number of tests is much larger than the sample size or a linear model where the number of covariates p is much larger than the sample size n.

There is a rich variety of methodologies for high-dimensional inference based on regularization that works by implicitly or explicitly penalizing models based on their dimensions. One of the most popular methods, Lasso (acronym for Least Absolute Shrinkage and Selection Operator) produces a point estimate of the coefficient vector relying on the ℓ_1 penalization for a negative log-likelihood for Gaussian observations. Lasso's widespread popularity is due to its computational efficiency based on the Least Angle Regression method due to Efron as well as its ability to produce a sparse solution, with optimality properties for both estimation and variable selection. Bühlmann and van de Geer (2011) provide a excellent reference for the theory based on Lasso and its various modifications.

The Bayesian alternatives to the sparse signal-recovery problem for high dimensional data can be broadly classified into two categories: discrete mixtures or "two-groups" model, or "spike-and-slab" priors (vide Bogdan et al. (2011), Efron (2008, 2010), Johnstone and Silverman (2004)) and shrinkage priors (Armagan et al. (2011, 2013), Carvalho et al. (2009, 2010), Castillo and van der Vaart (2012), Polson and Scott (2010), **?**). The first approach is based on putting a point mass at zero and an absolutely continuous prior on the non-zero elements of the parameter vector. The second approach entails putting absolutely continuous shrinkage priors on the entire parameter vector, that shr,castillo2012needlesink the entire coefficient towards zero. Both these approaches have their own advantages, and we discuss the tradeoffs associated with choosing one over the other a little later. It should also be noted that most of the penalization approaches can be interpreted in a Bayesian sense, by considering the mode of the posterior distribution under an appropriate shrinkage prior.

As a starting point, consider the classical normal means inference problem. We observe data from the probability model $(y_i|\theta_i) \sim \mathcal{N}(\theta_i, 1)$ for $i = 1, \ldots, n$. We wish to provide an estimator for the vector of normal means $\theta = (\theta_1, \ldots, \theta_n)$. Sparsity occurs when a large portion of the parameter vector contains zeros. The "ultra-sparse" or "nearly black" vector case occurs when the parameter vector θ lies in the set $l_0[p_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq p_n\}$ with the upper bound on the number of non-zero parameter values $p_n = o(n)$ as $n \to \infty$.

History of Shrinkage Estimation The story of shrinkage estimation harks back to Charles Stein's proof in 1955 that the maximum likelihood estimators for normal data are inadmissible beyond $\mathcal{R}^2$. The James-Stein estimator is $\hat{\theta}^{JS} = (1 - \frac{m-2}{||\mathbf{Y}||^2})Y$ with posterior mean $\hat{\theta}_{\text{Bayes}} = \frac{\tau^2}{\tau^2+1}Y$, which corresponds to the Bayes risk of $\frac{\tau^2}{\tau^2+1}.m$. Without any prior guess about τ^2, one could use an Empirical Bayes estimate and the resulting James-Stein estimator is $\hat{\theta}^{JS} = (1 - \frac{m-2}{||\mathbf{Y}||^2})Y$. James and Stein proved that this estimator dominates the MLE in terms of the expected total squarred error for every choice of $\boldsymbol{\theta}$, i.e. it outperforms the MLE no matter what the true $\boldsymbol{\theta}$ is. To motivate the need for developing new prior distributions, consider the classic James–Stein "global" shrinkage rule, $\hat{\theta}_{JS}(y)$. This estimator uniformly dominates the traditional sample mean estimator, $\hat{\theta}$. For all values of the true parameter θ and for $n > 2$, we have the classical mean squared error (MSE) risk bound:

$$R(\hat{\theta}_{JS}, \boldsymbol{\theta}) := \mathbb{E}_{y|\boldsymbol{\theta}} ||\hat{\theta}_{JS}(\mathbf{y}) - \boldsymbol{\theta}||^2 < n = \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}} ||\mathbf{y} - \boldsymbol{\theta}||^2, \quad \forall \boldsymbol{\theta}.$$

However, for the sparse signal problem, the standard James-Stein shrinkage rule $\hat{\theta}_{JS}$ performs poorly in a sparse setting is the case of the r-spike parameter value θ_r with r coordinates at $\sqrt{n/r}$ which has $\|\theta\| = n$. Johnstone and Silverman (2004) show that $E\|\hat{\theta}^{JS} - \theta\| \leq n$ with risk 2 at the origin. Moreover,

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

Hence for the r-spike parameter value $R\left(\hat{\theta}^{JS}, \theta_r\right) \geq (n/2)$. The thresholding rule $\hat{\theta}^{TS} = \sqrt{2\log n}$ has better risk $\sqrt{\log n}$.

The asymptotically minimax risk rate in ℓ_2 for nearly black objects is given by Donoho et al. (1992) to be $p_n \log(n/p_n)$. Here $a_n \asymp b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

1.2 Regularisation and Bayes

Regularization requires the researcher to specify a measure of fit, denoted by $l(\theta)$ and a penalty function, denoted by $\phi(\theta)$. Probabilistically, $l(\theta)$ and $\phi(\theta)$ correspond to the negative logarithms of the likelihood and prior distribution, respectively. Regularization leads to an optimization problem of the form

$$\underset{\theta \in \Re^d}{\text{minimise}} \quad l(\theta) + \phi(\theta) \,. \tag{1.1}$$

The probabilistic approach leads to a Bayesian hierarchical model

$$p(y \mid \theta) \propto \exp\{-l(\theta)\} \,, \quad p(\theta) \propto \exp\{-\phi(\theta)\} \,.$$

The solution to the minimisation problem estimated by regularisation (1.1) corresponds to the posterior mode, $\hat{\theta} = \arg\max_\theta \, p(\theta|y)$, where $p(\theta|y)$ denotes the posterior distribution. For example, regression with a least squares log-likelihood subject to a penalty such as an L^2-norm (ridge) (Hoerl and Kennard, 1970) Gaussian probability model or L^1-norm (lasso) (Tibshirani, 1996) double exponential probability model.

1.3 Spike-and-Slab: The Gold-Standard

The two-groups model is a natural hierarchical Bayesian solution to the sparse signal-recovery problem, that is "almost as simple to describe as the problem". The two-groups solution to the signal detection problem is as follows:

1. Assume each θ_i is non-zero with some common prior probability $\pi_1 = (1 - \pi_0)$, and that the nonzero θ_i come from a commonn density $\mathcal{N}(0, \psi^2)$.

2. Calculate the posterior probabilities (using Bayes' rule) that each y_i comes from $\mathcal{N}(0, \psi^2)$.

The most important aspect of this model is that it automatically adjusts for multiplicity without any ad-hoc regularization, i.e. it lets the data choose π_0 and then carry out the tests on the basis of the posterior inclusion probabilities $\omega_i = P(\theta_i \neq 0|y_i)$. Formally, in a two-groups model θ_i's are modeled as

$$\theta_i|\mu = (1 - \pi_0)\delta_{\{0\}} + \pi_0 \mathcal{N}(0, \psi^2), \tag{1.2}$$

where $\delta_{\{0\}}$ denotes a point mass at zero and the parameter $\psi^2 > 0$ is the non-centrality parameter that determines the separation between the two groups. Under this setting, the marginal distribution of $(y_i \mid \pi_0)$ is given by

$$y_i \mid \pi_0 \sim (1 - \pi_0)\mathcal{N}(0, 1) + \pi_0 \mathcal{N}(0, 1 + \psi^2). \tag{1.3}$$

As can be seen from Equation (1.3), the two-groups model leads to a sparse estimate, i.e., it puts exact zeros in the model.

1. Johnstone and Silverman (2004) showed that a thresholding-based estimator for θ under the two-groups model with an empirical Bayes estimate for μ is minimax in ℓ_2 sense.

2. Castillo and van der Vaart (2012) treated a full Bayes version of the problem and again found an estimate that is minimax in ℓ_2.

3. Bogdan et al. (2011) found that the estimator under the two-groups model provides asymptotically optimal performance in testing, in the sense that its performance matches the Bayes oracle up to a constant.

1.4 Towards the One-group Model

Despite the attractive theoretical properties outlined in the previous section, the discrete indicators in spike-and-slab models give rise to a combinatorial problem. While some posterior point estimates such as the posterior mean or quantiles might be easily computable for spike-and-slab (Castillo et al., 2015, Castillo and van der Vaart, 2012), exploring the full posterior using Markov chain Mote Carlo (MCMC) is typically more challenging using point mass mixture priors. A primary reason for this is exploring the posterior results in a combinatorial problem due to the discrete indicators and block updates of model parameters is also typically not feasible. Thus, one has to resort to some variation of a stochastic search algorithm in MCMC for model fitting, such as the stochastic search variable selection of George and McCulloch (1993, 1997) or shotgun stochastic search of Hans et al. (2007). Rovcková and George (2014) commented on the inefficiency of the stochastic search algorithms for exploring the posterior even for moderate dimensions and developed a deterministic alternative to quickly find the maximum a-posteriori model. We note that (a) increasing the efficiency in computation in the spike-and-slab model remains an active area of research (see, e.g., Rovcková and George, 2014) and (b) some complicating factors in the spike-and-slab model, such as a lack of suitable block updates, have fairly easy solutions for their continuous global-local shrinkage counterparts, facilitating posterior exploration.

We would also like to point out that the estimators resulting from these one-group shrinkage priors are very different from the shrinkage estimator due to James-Stein, who proved that the maximum likelihood estimators for normal data are inadmissible beyond $\mathcal{R}^2$. James-Stein estimator only worries about the total squared error loss, without much concern for the individual estimates. In problems involving observations lying far away on the tails, this could lead to 'overshrinkage'. In reality, an ideal signal-recovery procedure should be robust to large signals.

1.5 Global-local Horseshoe and Horseshoe+, $\phi(\theta)$

In a series of remarkable papers, Carvalho, Polson, and Scott (2009, 2010), Polson and Scott (2010, 2012) introduced a continuous "one-group" shrinkage rule based on what they call the horseshoe prior for multiple testing and model selection. The name 'Horseshoe' is attributed to the shape of the density of the shrinkage weight, κ, for each observation. Carvalho et al. (2010) also provide strong numerical evidence that the "one-group" shrinkage rule approximately behaves like the answers from a two-groups model. There have been another set of priors collectively called the "global-local" shrinkage priors after **?**. These new, wide class of shrinkage priors include the Generalized Double Pareto (GDP) (Armagan et al., 2013), the Three-Parameter Beta (**?**), the Hypergeometric Inverted Beta (**?**) and the more recent horseshoe+(**?**) and the Dirichlet-Laplace (Bhattacharya et al., 2015) prior, among others. The most recent member of this group is the horseshoe-like prior proposed by Bhadra et al. (2017).

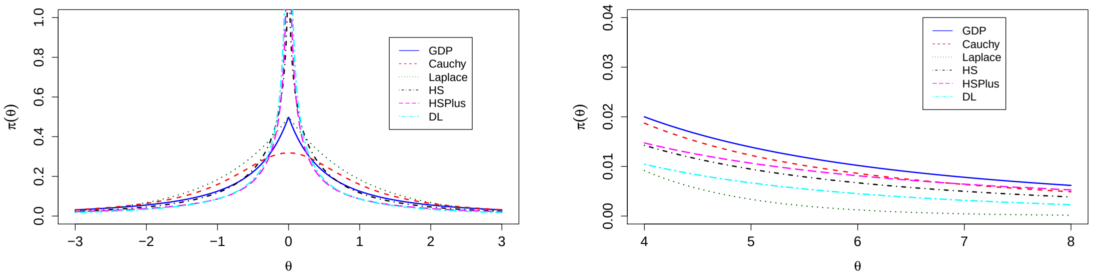
The global-local scale mixtures of normals can be written in the following hierarchical form (**?**):

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I});\ \beta_i \sim \mathcal{N}(0, \lambda_i^2\tau^2)$$
$$\lambda_i^2 \sim \pi(\lambda_i^2);\ (\tau, \sigma^2) \sim \pi(\tau^2, \sigma^2)$$

We list below the popular global-local shrinkage priors along with their behaviour near origin and the tails.

Prior	Origin Behavior	Tails
Horseshoe	$-\log(\|\theta\|)$	$\|\theta\|^{-2}$
Horseshoe+	$-\log(\|\theta\|)$	$\|\theta\|^{-1}$
Horseshoe-like	$-\|\theta\|^{1-\epsilon}\log(\|\theta\|)$	$\|\theta\|^{1-\epsilon}\ \epsilon \geq 0$
GDP	Bounded at origin	$\|\theta\|^{-(\alpha+1)}, \alpha \geq 0$
DL_a $(DL_{\frac{1}{n}})$	$\|\theta\|^{a-1}$ $(\|\theta\|^{\frac{1}{n}-1})$	$\exp(-b\|\theta\|)$

Table 1: Different Priors: Behaviour near origin and tails



(a) Marginal prior densities near the origin. The legends denote the horseshoe+ (HSPlus), horseshoe (HS), Dirichlet-Laplace (DL), generalized double Pareto (GDP), Cauchy and Laplace priors.

(b) Marginal prior densities in the tail regions. The legends denote the horseshoe+ (HSPlus), horseshoe (HS), Dirichlet-Laplace (DL), generalized double Pareto (GDP), Cauchy and Laplace priors.

2 Lasso and Horseshoe

Define estimators / compare shrinkage profiles. Asymptotic of finding zeroes. Explain $\phi(\theta)$ for horseshoe etc.

2.1 Shrinkage Profiles

κ-scale. Figures. $p(\kappa \mid \tau)$.

3 Risk Calculation

Normal means model $(y|\theta) \sim \mathcal{N}(0, \sigma^2 I_n)$. Problems with estimating θ when it is nearly black.

3.1 Minimax ℓ_2 risk achieved by global-local

1. van der Pas et al. (2014) showed it for horseshoe.

2. van der Pas et al. (2016) showed it for horseshoe+ and several other "global-local" models.

3. Ghosh and Chakrabarti (2014) is similar to van der Pas et al. (2016).

4. van der Pas et al. (2016) is a new paper that we need to read and possibly cite.

3.2 Posterior Concentration and Optimal Bayes Risk

1. For horseshoe: Datta and Ghosh (2013).

2. For horseshoe+: Bhadra et al. (2016c).

3.3 Prediction using global-local priors

1. Carvalho et al. (2010): K-L superefficiency for predictive density for horseshoe.

4 Regularisation and Hyper-parameter Selection

Path (computational speed), Sensitivity Analysis. Cross-validation approach (has caveats), marginal MLE, $\text{argmax}_\tau\, p(y \mid \tau)$. Plug-in estimators can be inadmissible, SURE: procedure.

Tiao and Tan (1965) show that the marginal likelihood, taking $\sigma^2 = 1$, is

$$p(y|\tau) \equiv \prod_{i=1}^{p}(1+\tau^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{(y_i-\mu)^2}{1+\tau^2}\right)$$

This is positive at $\tau = 0$. Hence the impropriety of the prior τ^{-2} at the origin translates to the posterior. The marginal likelihood is decreasing at zero when the S_i's are small enough to make the exponential term nearly constant (Tiao and Tan, 1965). This is precisely the sparse coefficient case.

A number of default choices have been proposed to overcome this issues. Morris and Tang (2011) propose a flat prior $p(\tau) \equiv 1$. The tails of the likelihood are sufficient so as to lead to a proper posterior. This is also related to Stein's harmonic prior $||\theta||^{-(k-2)}$ for $k \geq 3$.

Methods for choosing τ will involve minimizing some criteria:

5 Simulations

1. Mention R package.

2. Compare with various methods.

6 Applications and Extensions

Literature review of applications and extensions and recent development. List papers. Fused and group lasso. Extension to logistic (Polya-Gamma for logistic : MCMC still convex).

1. Bhadra et al. (2016a) use global-local priors in default Bayes Efron problems.

2. Bhadra et al. (2016) show how to use global-local priors for prediction. Performs better than a variety of competitors including lasso, ridge, PCR and sparse PLS.

3. Datta and Dunson (2016) use global-local priors to model the rate parameter for Poisson count data.

7 Discussion

What's left to do? Horseshoe subset selection. Lasso computationally quick / scalable. Horseshoe needs MCMC etc.

1. So many global-local priors; what is the unifying theme?

2. How close to minimax constant can we get in estimation?

3. How close to oracle risk can we get in testing?

4. Is prediction performance optimal?

5. Yet to rigorously show global-local priors have any information theoretic properties in default Bayes problems that reference priors (Bernardo, 1979) enjoy.

6. Bhadra et al. (2016b) demonstrate how global-local mixtures can be generated using two integral identities. This might prove useful in EM and MCMC.

8 Appendix: Algorithms

Literature review of algorithms and R packages. LASSO: glmnet, genlasso. Horseshoe: horseshoe, fastHorseshoe, monomvn, our own package.

Other refs

The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators - Stephen M. Stigler

References

Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems*, pages 523–531.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23,** 119–143.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **41,** 113–147.

Bhadra, A., Datta, J., Li, Y., Polson, N. G., and Willard, B. (2016). Prediction risk for global-local shrinkage regression. *arXiv preprint arXiv:1605.04796* .

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016a). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* **to appear,**.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016b). Global-local mixtures. *arXiv preprint arXiv:1604.07487* .

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016c). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **to appear,**.

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

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics* **39,** 1551–1579.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer-Verlag Berlin Heidelberg.

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.

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

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43,** 1986–2018.

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

Datta, J. and Dunson, D. B. (2016). Inference on high-dimensional sparse count data. *Biometrika* **to appear,**.

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

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **54,** 41–81.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23,** 1–22.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica* pages 339–373.

Ghosh, P. and Chakrabarti, A. (2014). Posterior Concentration Properties of a General Class of Shrinkage Estimators around Nearly Black Vectors. *ArXiv: 1412.8161* .

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* **102,** 507–516.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, New York, 2nd edition.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

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

Morris, C. and Tang, R. (2011). Estimating random effects via adjustment for density maximization. *Statist. Sci.* **26,** 271–287.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics* **9,** 501–538.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* **7,** 887–902.

Rovcková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association* **109,** 828–846.

Tiao, G. C. and Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. i. posterior distribution of variance-components. *Biometrika* **52,** 37–53.

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

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,** 2585–2618.

van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics* **10,** 976–1000.

van der Pas, S., Szabó, B., and van der Vaart, A. (2016). How many needles in the haystack? adaptive inference and uncertainty quantification for the horseshoe. *arXiv:1607.01892* .