*Statistical Science*
2019, Vol. 34, No. 3, 405–427
https://doi.org/10.1214/19-STS700


# Lasso Meets Horseshoe: A Survey

**Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson and Brandon Willard**

*Abstract.* The goal of this paper is to contrast and survey the major advances in two of the most commonly used high-dimensional techniques, namely, the Lasso and horseshoe regularization. Lasso is a gold standard for predictor selection while horseshoe is a state-of-the-art Bayesian estimator for sparse signals. Lasso is fast and scalable and uses convex optimization whilst the horseshoe is non-convex. Our novel perspective focuses on three aspects: (i) theoretical optimality in high-dimensional inference for the Gaussian sparse model and beyond, (ii) efficiency and scalability of computation and (iii) methodological development and performance.

*Key words and phrases:* Global-local priors, horseshoe, horseshoe+, hyperparameter tuning, Lasso, regression, regularization, sparsity.

## 1. INTRODUCTION

High-dimensional predictor selection and sparse signal recovery are routine statistical and machine learning practices. There is a vast and growing literature focusing on computational aspects of large scale inference problems. Whilst this area is too large to review here, we revisit two popular sparse parameter estimation techniques, the Lasso (Tibshirani, 1996) and the horseshoe estimator (Carvalho, Polson and Scott, 2010). Specifically, we focus on three areas: performance in high-dimensional data, theoretical optimality and computational efficiency.

Sparsity relies on the property of a few large signals among many (nearly) zero noisy observations. A common goal in high-dimensional inference is to recover the low-dimensional signals observed in noisy observations. This problem encompasses four related areas:

(i) Estimation of the underlying sparse parameter vector.

(ii) Multiple testing where the # tests is much larger than the sample size, $n$.

(iii) Regression subset selection where # of covariates $p$ is far larger than $n$.

(iv) Out-of-sample prediction.

There are a rich variety of methodologies for high-dimensional regularization which implicitly or explicitly penalize model dimensionality. Lasso (Least Absolute Shrinkage and Selection Operator) produces a sparse estimate by constraining the $\ell_1$ norm of the parameter vector. Lasso's widespread popularity is due to a multitude of factors, in particular due to the computational efficiency of the least angle regression (LARS) (Efron et al., 2004) or the simple coordinate descent approaches of Friedman et al. (2007), and its ability to produce a sparse solution, with optimality (oracle) properties for both estimation and variable selection (*vide* Bühlmann and van de Geer, 2011, James et al., 2013, Hastie, Tibshirani and Wainwright, 2015). Table 1, adapted from Tibshirani (2014), gives a list of popular regularization methods based on Lasso.

Bayes procedures, on the other hand, can be classified into two categories: two-groups model or spike-and-slab priors (Johnstone and Silverman, 2004, Efron, 2008, 2010, Bogdan et al., 2011, Castillo and van der Vaart, 2012) and global–local shrinkage priors (Carvalho, Polson and Scott, 2009, 2010, Griffin and Brown, 2010, Armagan, Clyde and Dunson, 2011,

*Anindya Bhadra is Associate Professor, Department of Statistics, Purdue University, 250 N. University St., West Lafayette, Indiana 47907, USA (e-mail: bhadra@purdue.edu). Jyotishka Datta is Assistant Professor, Department of Mathematical Sciences, University of Arkansas, 1 University of Arkansas, Fayetteville, Arkansas 72704, USA (e-mail: jd033@uark.edu). Nicholas G. Polson is Professor, Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave., Chicago, Illinois 60637, USA (e-mail: ngp@chicagobooth.edu). Brandon Willard is Researcher, Booth School of Business, University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, Illinois 60637, USA (e-mail: bwillard@uchicago.edu).*

TABLE 1
*Lasso regularization methods*

| Method | Authors |
|---|---|
| Adaptive Lasso | Zou (2006) |
| Compressive sensing | Donoho (2006), Candès (2008) |
| Dantzig selector | Candes and Tao (2007) |
| Elastic net | Zou and Hastie (2005) |
| Fused Lasso | Tibshirani et al. (2005) |
| Generalized Lasso | Tibshirani and Taylor (2011) |
| Graphical Lasso | Friedman, Hastie and Tibshirani (2008) |
| Grouped Lasso | Yuan and Lin (2006) |
| Hierarchical interaction models | Bien, Taylor and Tibshirani (2013) |
| Matrix completion | Candès and Tao (2010), Mazumder, Hastie and Tibshirani (2010) |
| Multivariate methods | Jolliffe, Trendafilov and Uddin (2003), Witten, Tibshirani and Hastie (2009) |
| Near-isotonic regression | Tibshirani, Hoefling and Tibshirani (2011) |
| Square Root Lasso | Belloni, Chernozhukov and Wang (2011) |
| Scaled Lasso | Sun and Zhang (2012) |
| Minimum concave penalty | Zhang (2010) |
| SparseNet | Mazumder, Friedman and Hastie (2011) |

Armagan, Dunson and Lee, 2013, Polson and Scott, 2011), with the horseshoe prior (Carvalho, Polson and Scott, 2010) being one of the most popular methods. The first class, spike-and-slab prior, places a discrete mixture of a point mass at zero (the spike) and an absolutely continuous density (the slab) on each parameter. The second entails placing absolutely continuous shrinkage priors on the entire parameter vector that selectively shrinks the small signals. Table 2 provides a sampling of a few continuous shrinkage priors popular in the literature. Both these approaches have their own advantages and caveats, which we discuss in turn. A key duality is that the point estimate from a regularization approach can be interpreted as Bayesian mode of the posterior distribution under an appropriate shrinkage prior.

Both Lasso and horseshoe procedures come with strong theoretical guarantees for estimation, prediction and variable selection. Both procedures possess asymptotic oracle properties, that is, identify the true non-zero coefficients as well as achieve the optimal estimation rate. The behavior of the Lasso estimator in terms of the risk properties has been studied in depth and has resulted in many methods aiming to improve certain features (see Table 1). On the other hand, horseshoe and other global–local priors have been shown to achieve optimality in variable selection, estimation and prediction, that we review in Section 4, although theo-

TABLE 2
*A catalog of global–local shrinkage priors*

| Global-local shrinkage prior | Authors |
|---|---|
| Normal Exponential Gamma | Griffin and Brown (2010) |
| Horseshoe | Carvalho, Polson and Scott (2010, 2009) |
| Hypergeometric Inverted Beta | Polson and Scott (2010) |
| Generalized Double Pareto | Armagan, Clyde and Dunson (2011) |
| Generalized Beta | Armagan, Dunson and Lee (2013) |
| Dirichlet–Laplace | Bhattacharya et al. (2015) |
| Horseshoe+ | Bhadra et al. (2017b) |
| Horseshoe-like | Bhadra et al. (2017a) |
| Spike-and-Slab Lasso | Ročková and George (2018) |
| R2–D2 | Zhang, Reich and Bondell (2016) |
| Inverse-Gamma–Gamma | Bai and Ghosh (2017) |

retical studies of the continuous shrinkage priors is still an active area.

The rest of the paper is organized as follows: Section 2 provides historical background for the normal means (a.k.a the Gaussian compound decision problem) and the sparse regression problems. Section 3 provides the link between regularization and optimization perspectives viewed through a probabilistic Bayesian lens. Section 4 compares and contrasts the statistical risk properties of Lasso and the horseshoe prior. Sections 5 and 6 discuss the issues of hyperparameter selection and computational strategies. Section 7 provides two simulation experiments comparing the horseshoe prior with penalized regression methods for linear model and logistic regression with varying degree of dependence between predictors. We discuss applications of Lasso and the horseshoe in Section 8 and provide directions for future work in Section 9.

## 2. SPARSE NORMAL MEANS, REGRESSION AND VARIABLE SELECTION

### 2.1 Sparse Normal Means

Suppose that we observe data from the probability model $(y_i \mid \theta_i) \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1)$ for $i = 1, \ldots, n$. Our primary inferential goal is to estimate the vector of normal means $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and a secondary goal is to simultaneously test if $\theta_i$'s are coming from a null distribution. We are interested in the sparse paradigm where a large proportion of the parameter vector contains zeros. The 'nearly black' (Donoho et al., 1992) regime occurs when the parameter vector $\theta$ lies in the set $\ell_0[p_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq p_n\}$ with the upper bound on the number of non-zero parameter values $p_n = o(n)$ as $n \to \infty$.

A natural Bayesian solution for inference under sparsity is the two-groups model that puts a non-zero probability spike at zero and a suitable prior on the non-zero $\theta_i$'s (*vide* Appendix A). The inference is then based on the posterior probabilities of non-zero $\theta_i$'s based on the discrete mixture model. The two-groups model possesses a number of frequentist and Bayesian optimality properties. Johnstone and Silverman (2004) showed that a thresholding-based estimator for $\boldsymbol{\theta}$ under the two-groups model with an empirical Bayes estimate for the sparsity proportion attains the minimax rate in $\ell_q$ norm for $q \in (0, 2]$ for $\boldsymbol{\theta}$ that are either nearly black or belong to an $\ell_p$ ball of 'small' radius. Castillo and van der Vaart (2012) treated a full Bayes version of the problem and again found an estimate that is minimax in $\ell_q$ norm for mean vectors that are either nearly black or have bounded weak $\ell_p$ norm for $p \in (0, 2]$.

### 2.2 Sparse Linear Regression

A related inferential problem is high-dimensional linear regression with sparsity constraints on the parameter vector $\boldsymbol{\theta}$. We are interested in the linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ is a $n \times p$ matrix of predictors and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Our focus is on the sparse solution where $p \gg n$ and most of $\theta_i$'s are zero. Similar to the sparse normal means problem, our goal is to identify the non-zero entries of $\boldsymbol{\theta}$ as well as estimate it. There are a wide variety of methods based on the penalized likelihood approach that solves the following optimization problem:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left( y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{i,j} \right)^2 + \text{pen}_\lambda(\boldsymbol{\theta}),$$

(1) where

$$\text{pen}_\lambda(\boldsymbol{\theta}) = \sum_{j=1}^{p} p_\lambda(\theta_j) \quad \text{is a separable penalty.}$$

Lasso uses an $\ell_1$ penalty, $p_\lambda(\theta_j) = \lambda|\theta_j|$, and simultaneously performs variable selection while maintaining estimation accuracy. Another notable variant is the best subset selection procedure corresponding to the $\ell_0$ penalty $p_\lambda(\theta_j) = \lambda \mathbf{1}\{\theta_j \neq 0\}$. There has been a recent emphasis on non-concave separable penalties such as the minimax concave penalty or MCP (Zhang, 2010) or SCAD (Fan and Li, 2001), that act as a tool for variable selection and estimation. We discuss the penalization methods from a Bayesian viewpoint in the next section.

### 2.3 Variable Selection

Variable or predictor selection is intimately related to high-dimensional sparse linear regression. A sparse model provides interpretability, computational efficiency, and stability of inference. Lasso's success has inspired many estimation methods that rely on convexity and sparsity in a penalized estimation framework. The 'bet on sparsity' principle (Hastie, Tibshirani and Friedman, 2009) dictates the use of methods favoring sparsity, as no method uniformly dominates when the true model is dense.

REMARK 1. The LAVA method by Chernozhukov, Hansen and Liao (2017) strictly dominates both Lasso and ridge in a 'sparse + dense' model. In fact, the

LAVA estimator performs as well as Lasso in a sparse regime and as well as ridge (Tikhonov, 1963) in a dense regime. This questions the validity of the 'bet on sparsity' principle. Although there is no exact analogue of LAVA in the Bayesian world, the one-group shrinkage priors share a common philosophy. The horseshoe-type priors are also designed to work when true $\boldsymbol{\theta}$ has a few large entries and very many small non-zero entries and produces a 'non-sparse' estimator, but LAVA can recover both the dense and sparse components unlike horseshoe.

A parallel surge of Bayesian methodologies has emerged for sparse regression problems with an underlying variable selection procedure. Hierarchical Bayesian modeling proceeds by selecting a model dimension $s$, selecting a random subset $S$ of dimension $|S| = s$ and a prior $\pi_S$ on $\mathbb{R}^s$. The prior can be written as in Castillo, Schmidt-Hieber and van der Vaart (2015):

$$(S, \boldsymbol{\theta}) \mapsto \binom{p}{|S|}^{-1} \pi_p(|S|)\pi_S(\boldsymbol{\theta}_S)\delta_0(\boldsymbol{\theta}_{S^c}). \tag{2}$$

Bayesian approaches for sparse linear regression include George (2000), George and Foster (2000), Mitchell and Beauchamp (1988), Ishwaran and Rao (2005) and more recently Ročková and George (2018), who introduce the spike-and-slab Lasso prior, where the hierarchical prior on the parameter and model spaces assumes the form:

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \gamma) &= \prod_{i=1}^{p} [\gamma_i \pi_1(\theta_i) + (1 - \gamma_i)\pi_0(\theta_i)], \\ \gamma &\sim p(\cdot), \end{aligned} \tag{3}$$

where $\boldsymbol{\gamma}$ indexes the $2^p$ possible models, and $\pi_0$, $\pi_1$ model the null and non-null $\theta_i$'s respectively using two Laplace priors with different scales. However, the spike-and-slab type priors lead to substantial computational challenges as exploring the full posterior using point mass mixture priors is prohibitive due to a combinatorial complexity of updating the discrete indicators and infeasibility of block updating of model parameters.

The continuous shrinkage priors alleviate this by using efficient Gibbs sampling scheme based on block-updating the model parameters. We also note that while full posterior sampling remains a computational hurdle for the spike-and-slab prior, point estimates such as posterior mean and posterior quantiles can be obtained using a polynomial-time algorithm as shown by Castillo and van der Vaart (2012). Ročková and George (2018) discuss the inefficiency of stochastic search algorithms for exploring the posterior even for moderate dimensions and developed a deterministic alternative to quickly find the maximum a-posteriori model. Here (i) increasing the efficiency in computation in the spike-and-slab model remains an active area of research (see, e.g., Ročková and George, 2018) and (ii) 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.

Polson and Scott (2011, 2012b), Carvalho, Polson and Scott (2010) introduced the 'global–local' shrinkage priors that adjust to sparsity via global shrinkage, and identify signals by local shrinkage parameters. The global–local shrinkage idea has resulted in many different priors in the recent past, with varying degrees of success in theoretical and numerical performance. We compare these different priors and introduce a recently proposed family of horseshoe-like priors in Section 3.3.

The estimators resulting from the one-group shrinkage priors are very different from the shrinkage estimator due to James and Stein (1961), who showed that maximum likelihood estimators for multivariate normal means are inadmissible beyond two dimensions. The James–Stein estimator is primarily concerned about the total squared error loss, without much regard for the individual estimates. In problems involving observations lying far away on the tails this leads to 'over-shrinkage' (Carvalho, Polson and Scott, 2010). In reality, an ideal signal-recovery procedure should be robust to large signals. Connections between the global shrinkage of James–Stein and global–local shrinkage of the horseshoe are discussed in more details in Section 4.

## 3. LASSO AND HORSESHOE

Regularization requires the researcher to specify a measure of fit, denoted by $l(\boldsymbol{\theta})$ and a penalty function, denoted by $\text{pen}_\lambda(\boldsymbol{\theta})$. From a Bayesian perspective, $l(\theta)$ and $\text{pen}_\lambda(\boldsymbol{\theta})$ correspond to the negative logarithms of the likelihood and a suitable prior distribution, respectively. While regularization leads to an optimization problem of the form

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \{l(y \mid \boldsymbol{\theta}) + \text{pen}_\lambda(\boldsymbol{\theta})\}, \tag{4}$$

the probabilistic approach leads to a Bayesian hierarchical model

$$p(y \mid \boldsymbol{\theta}) \propto \exp\{-l(y \mid \boldsymbol{\theta})\}, \quad \pi_\lambda(\boldsymbol{\theta}) \propto \exp\{-\text{pen}_\lambda(\boldsymbol{\theta})\}. \tag{5}$$

For appropriate $l(y \mid \boldsymbol{\theta})$ and $\text{pen}_\lambda(\boldsymbol{\theta})$, the solution to (4) corresponds to the posterior mode of (5), $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta} \mid y)$, where $p(\boldsymbol{\theta} \mid y)$ denotes the posterior density. The properties of the penalty are then induced by those of the prior. For example, regression with a least squares log-likelihood subject to an $\ell_2$ penalty or ridge (Tikhonov, 1963, Hoerl and Kennard, 1970) corresponds to a Gaussian prior under the same observation distribution, and an $\ell_1$ penalty (Lasso) corresponds to a double-exponential prior (Tibshirani, 1996).

One interpretation of Lasso and related $\ell_1$ penalties is that these are methods designed to perform selection, while ridge and related $\ell_2$ based methods perform shrinkage. Selection-based methods such as the Lasso are unstable in many situations, for example, in presence of multi-collinearity in the design (Hastie, Tibshirani and Friedman, 2009, ch. 3).

Although 'shrinkage' and 'selection' are closely related, we tend to distinguish between them in the following sense. Shrinkage methods such as the horseshoe prior shrink towards 0 by thresholding the shrinkage weights that behave like posterior inclusion probabilities $P(\theta_i \neq 0 \mid y_i)$ to achieve variable selection. It should be noted that the continuous nature of prior on $\theta_i$ ensures a lack of exact zeros in the posterior, which is often preferred over dichotomous models by some practitioners (Stephens and Balding, 2009) as more realistic. This is unlike the Lasso that performs explicit selection by making some of estimates 0 and producing a true sparse solution. Ultimately, both selection and shrinkage have their advantages and disadvantages.

### 3.1 Lasso Penalty and Prior

As discussed before, the classical Lasso-based point estimate is the same as the posterior mode under component-wise Laplace prior, and the mode inherits the optimal properties of Lasso. For example, the oracle inequality in Bühlmann and van de Geer (2011), Eq. (2.8), Th. (6.1), states that with a proper choice of $\lambda$ of order $\sigma\sqrt{\log(p)/n}$, the mean squared prediction error of Lasso is of the same order as if one knew active set $S_0 = \{j : \theta_j^0 \neq 0\}$), up to $O(\log(p))$ and a compatibility constant $\phi_0^2$. The compatibility (or restricted eigenvalue) constant reflects the compatibility between the design matrix and the $\ell_1$ norm of $\boldsymbol{\theta}$, and is defined as follows (Bühlmann and van de Geer, 2011, Eq. (6.4)):

DEFINITION 2 (Compatibility condition). For $S \subset \{1, 2, \ldots, p\}$ and $\boldsymbol{\theta} \in \mathbb{R}^p$, let $\boldsymbol{\theta}_{j,S} \doteq \theta_j 1\{j \in S\} \in \mathbb{R}^p$ (with similar notation for $\boldsymbol{\theta}_{j\in S} \in \mathbb{R}^{|S|}$), and let $\boldsymbol{\theta}_{-S} = \boldsymbol{\theta}_{S^c}$. Then the compatibility condition is satisfied for the design $\mathbf{X}$ for the true support set $S = \text{supp}(\boldsymbol{\theta})$, if letting $s_0 = |S|$ one has,

$$\frac{1}{n}\|\mathbf{X}\boldsymbol{\theta}\|_2^2 \geq \frac{\phi_0^2}{s_0}\|\boldsymbol{\theta}_S\|_1^2, \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^p \text{ such that } \|\boldsymbol{\theta}_S\|_1 \leq 3\|\boldsymbol{\theta}_{-S}\|_1.$$

The constant $\phi_0^2$ is called the compatibility (or restricted eigenvalue) constant.

Lasso also exhibits other desirable properties such as computational tractability, consistency of point estimates of $\boldsymbol{\theta}$ for suitable $\lambda$, and optimality results on variable selection.

### 3.2 Bayesian Lasso and Elastic Net

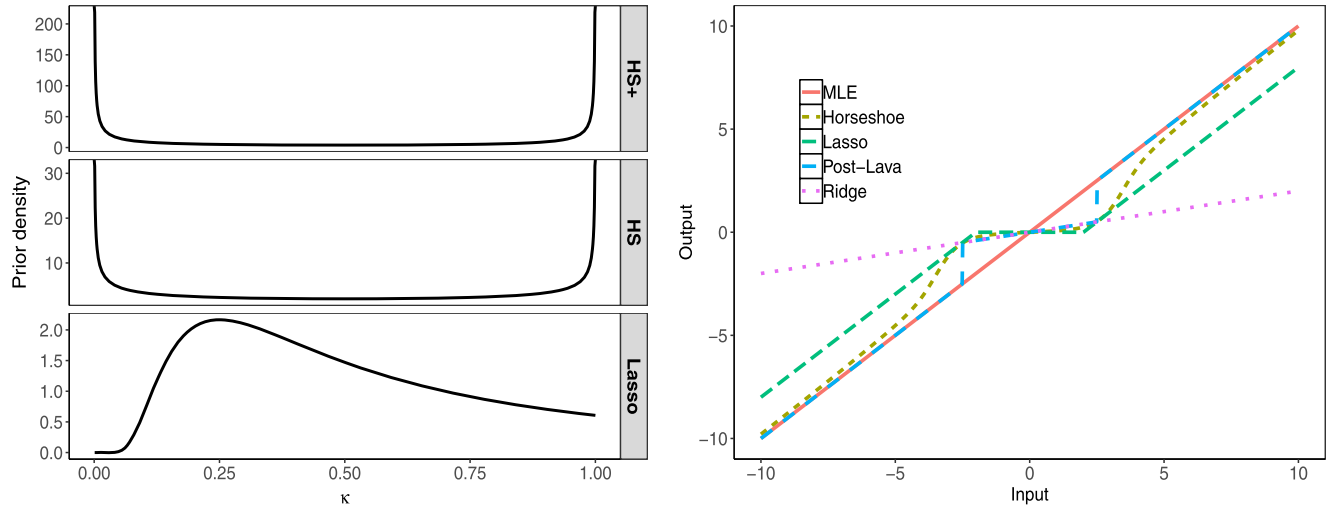
As discussed before, the posterior mean under the double-exponential prior, which is the Bayes estimate under squared error loss, does not satisfy the optimality properties of the posterior mode under the double-exponential prior (i.e., the Lasso). Along these lines, Castillo, Schmidt-Hieber and van der Vaart (2015) argue that the Lasso is essentially non-Bayesian, in that the "*full posterior distribution is useless for uncertainty quantification, the central idea of Bayesian inference*." Castillo, Schmidt-Hieber and van der Vaart (2015) provide theoretical result that the full Lasso posterior does not contract at the same speed as the posterior mode.

Thus, there are a number of caveats related to the use of a double-exponential prior for the general purposes of shrinkage. An important example is found in how it handles shrinkage for small observations and robustness to the large ones. This behavior is described by various authors, including Polson and Scott (2011), Datta and Ghosh (2013), and motivates the key properties of global–local priors. Figure 1(a) provides profile plots as a diagnostic of shrinkage behavior for different priors.

For correlated predictors, Zou and Hastie (2005) proposed a family of convex penalties called 'elastic net', which is a hybrid between Lasso and ridge. The penalty term is $\sum_{j=1}^p \lambda p_\alpha(\theta_j)$, where

$$p_\alpha(\theta_j) = \frac{1}{2}(1-\alpha)\theta_j^2 + \alpha|\theta_j|, \quad j = 1, \ldots, p.$$

Both Lasso and elastic net facilitate efficient Bayesian computation via a global–local scale mixture representation (Bhadra et al., 2016a). The Lasso




(a) Shrinkage factor $\kappa_i$ against $p(\kappa_i)$ (b) Score function

FIG. 1. *(a) Prior density of shrinkage weight $\kappa_i$ for the horseshoe, horseshoe+, and Laplace prior, where $\{1 - E(\kappa_i \mid y_i)\}$ can be interpreted as the pseudo posterior inclusion probability that mimics $P(\theta_i \neq 0 \mid y_i)$, and (b) shrinkage function for LAVA, Lasso, ridge and the horseshoe estimator. For LAVA shrinkage function, we have chosen $\lambda_1 = \lambda_l = 4$ and $\lambda_2 = \lambda_r = 4$, and for the horseshoe prior the value of global shrinkage parameter $\tau$ is fixed at* 0.1.

penalty arises as a Laplace global–local mixture (Andrews and Mallows, 1974), while the elastic-net regression can be recast as a global–local mixture with a mixing density belonging to the orthant-normal family of distributions (Hans, 2011). The orthant-normal prior on $\theta_i$, given hyper-parameters $\lambda_1$ and $\lambda_2$, has a density function with the following form:

$$
\begin{aligned}
&p(\theta_i \mid \lambda_1, \lambda_2) \\
&= \begin{cases} \phi\left(\theta_i \,\middle|\, \dfrac{\lambda_1}{2\lambda_2}, \dfrac{\sigma^2}{\lambda_2}\right) \Big/ 2\Phi\left(-\dfrac{\lambda_1}{2\sigma\lambda_2^{1/2}}\right), \\ \quad \theta_i < 0, \\ \phi\left(\theta_i \,\middle|\, \dfrac{-\lambda_1}{2\lambda_2}, \dfrac{\sigma^2}{\lambda_2}\right) \Big/ 2\Phi\left(-\dfrac{\lambda_1}{2\sigma\lambda_2^{1/2}}\right), \\ \quad \theta_i \geq 0. \end{cases}
\end{aligned} \tag{6}
$$

### 3.3 Horseshoe Penalty and Prior

The horseshoe prior is a continuous shrinkage rule for sparse signal recovery. Here we discuss the motivation behind the horseshoe prior for the Gaussian sequence model as it was developed in Carvalho, Polson and Scott (2010), but note that it is applicable to sparse signal recovery in regression models and beyond, as we discuss in Section 4.4. Consider the normal means model: $y_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1)$, $\theta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \lambda_i^2\tau^2)$, $i = 1, 2, \ldots, n$. The horseshoe prior for $\theta_i$, given a global shrinkage parameter $\tau$, is given by the hierarchical model

$$
\begin{aligned}
(y_i \mid \theta_i) &\sim \mathcal{N}(\theta_i, \sigma^2), \\
(\theta_i \mid \lambda_i, \tau) &\sim \mathcal{N}(0, \lambda_i^2\tau^2), \\
\lambda_i^2 &\sim \mathrm{C}^+(0, 1), \quad i = 1, \ldots, n.
\end{aligned} \tag{7}
$$

As discussed before, the spike-and-slab prior or the two-groups model (*vide* Appendix A) with two dedicated components for separating noise and signal is a natural Bayesian solution but it leads to substantial computational burden. The horseshoe prior takes a different approach: instead of placing a prior on the model space to yield a sparse estimator, it models the posterior inclusion probabilities $P(\theta_i \neq 0 \mid y_i)$ directly. To see this, note that the posterior mean under the horseshoe prior can be written as a linear function of the observation:

$$
\begin{aligned}
\mathbb{E}(\theta_i \mid y_i) &= \{1 - \mathbb{E}(\kappa_i \mid y_i)\} y_i \\
&\text{where } \kappa_i = 1/(1 + \lambda_i^2\tau^2).
\end{aligned} \tag{8}
$$

The name 'horseshoe' arises from the shape of the beta prior density of the shrinkage weights $\kappa_i$. A comparison with the posterior mean obtained under the two-groups model reveals that the shrinkage weights perform the same function as the posterior inclusion probability $P(\theta_i \neq 0 \mid y_i)$ for recovering a sparse signal. Since the shrinkage coefficients are not formal Bayesian posterior quantities, we refer to them as 'pseudo posterior inclusion probabilities.'

Carvalho, Polson and Scott (2010) provided strong numerical evidence that the shrinkage weights from a one-group prior accurately approximates the inclusion probabilities under a two-groups model, and used this property to construct a multiple testing rule. The thresholding rule rejects the $i$th null hypothesis $H_{0i}$ : $\theta_i = 0$ if the shrinkage weight $1 - \hat{\kappa}_i$ exceed 0.5. Datta and Ghosh (2013) validated this theoretically by proving that the horseshoe multiple testing rule attains the Bayes oracle up to a multiplicative constant under a 0–1 additive loss.

The marginal likelihood after reparametrizing $\kappa_i = (1 + \lambda_i^2\tau^2)^{-1}$ is: $p(y_i \mid \kappa_i, \tau) = \kappa_i^{1/2} \exp(-\kappa_i y_i^2/2)$. The posterior density of $\kappa_i$ identifies signals and noises by letting $\kappa_i \to 0$ and $\kappa_i \to 1$ respectively. Since the marginal likelihood is zero when $\kappa_i = 0$, it does not help identify the signals. Intuitively, any prior that drives the probability to either extremities should be a good candidate for sparse signal reconstruction. The horseshoe prior, with an induced prior density on $\kappa_i$ proportional to $\kappa_i^{-1/2}(1-\kappa_i)^{-1/2}$ does exactly that: it cancels the $\kappa_i^{1/2}$ term in the marginal likelihood and replaces it with $(1-\kappa_i)^{-1/2}$ to enable $\kappa_i \to 1$ in the posterior. The horseshoe+ prior (Bhadra et al., 2017b) takes this philosophy one step further, by creating a $U$-shaped Jacobian for transformation from $\lambda_i$ to $\kappa_i$-scale. The double-exponential on the other hand, yields a prior that decays at both ends with a mode near $\kappa_i = 1/4$, thus leading to a posterior that is neither good at adjusting to sparsity, nor at recovering large signals (see Table 3).

Figure 1(a) plots the prior density $p(\kappa_i)$ for the horseshoe, horseshoe+, and the Laplace priors. Figure 1(b) shows the resulting shrinkage function by plotting the input observations against the output estimates for horseshoe, horseshoe+, and Laplace priors, along with the maximum likelihood estimator ($\hat{\boldsymbol{\theta}} = \mathbf{y}$). Both Lasso and horseshoe shrink the small observations, but while horseshoe and horseshoe+ leave the large inputs unshrunk, Lasso shrinks them by a non-vanishing amount, resulting in a non-zero bias. We also plot the shrinkage function for the post-lava estimator (Chernozhukov, Hansen and Liao, 2017) (*vide* Appendix C) which works well on dense + sparse signals, and has the robustness property lacking in Bayesian Lasso or the Laplace prior.

TABLE 3
*Priors for $\lambda_i$ and $\kappa_i$ for a few popular shrinkage rules*

| Prior for $\theta_i$ | Prior for $\lambda_i$ | Prior for $\kappa_i$ |
|---|---|---|
| Horseshoe | $2/\{\pi\tau(1+(\lambda_i/\tau)^2)\}$ | $\frac{\tau}{\sqrt{\kappa_i(1-\kappa_i)}}\frac{1}{(1+\kappa_i(\tau^2-1))}$ |
| Horseshoe+ | $\frac{4\log\lambda_i/\tau}{\{\pi^2\tau(\lambda_i/\tau)^2-1)\}}$ | $\frac{\tau}{\sqrt{\kappa_i(1-\kappa_i)}}\frac{\log\{(1-\kappa_i)/\kappa_i\tau^2\}}{(1-\kappa_i(\tau^2+1))}$ |
| Double Exponential | $\lambda_i \exp(-\lambda_i^2/2)$ | $\kappa_i^{-2}\exp-\frac{1}{2\kappa_i}$ |

There are a number of closed-form results for the posterior distribution under a horseshoe prior. Although the prior density under the horseshoe prior does not admit a closed form, we can write the horseshoe posterior mean using the Tweedies' formula $\mathbb{E}(\theta \mid y) = y + \frac{\partial \ln m(y)}{\partial y}\sigma^2$, which is also the Bayes adjustment that provides an optimal bias-variance trade-off. For the horseshoe prior, Tweedies' formula yields:

$$\mathbb{E}(\theta_i \mid y_i, \tau) = y_i\left(1 - \frac{2\Phi_1(\frac{1}{2}, 1, \frac{5}{2}, \frac{y_i^2}{2\sigma^2}, 1-\frac{1}{\tau^2})}{3\Phi_1(\frac{1}{2}, 1, \frac{3}{2}, \frac{y_i^2}{2\sigma^2}, 1-\frac{1}{\tau^2})}\right), \tag{9}$$

where $\Phi_1$ is the bivariate confluent hypergeometric function (Gordy, 1998). A similar formula is available for the posterior variance. This enables one to rapidly calculate the posterior mean estimator under the horseshoe prior via a 'plug-in' approach with estimated values of the hyper-parameter $\tau$. We discuss the different approaches for handling $\tau$ in Section 5 and statistical properties of horseshoe posterior mean estimator and the induced decision rule in more details in Section 4.

The horseshoe prior is a member of a wider class of global–local scale mixtures of normals that admit following hierarchical form (Polson and Scott, 2011):

$$(\mathbf{y} \mid \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}); \quad \theta_i \sim \mathcal{N}(0, \lambda_i^2\tau^2),$$
$$\lambda_i^2 \sim \pi(\lambda_i^2); \quad (\tau, \sigma^2) \sim \pi(\tau^2, \sigma^2), \quad i = 1, \ldots, n.$$

These priors are collectively called global–local shrinkage priors in Polson and Scott (2011), since they recover signals by a local shrinkage parameter and adapt to sparsity by a global shrinkage parameter. Table 2 provides a list of the popular and recent global–local shrinkage priors. A natural question is *how do we compare these priors?* It is known due to several authors (e.g., Polson and Scott, 2011, Bhadra et al., 2016c) that the key features of a global–local shrinkage prior is a peak at origin and heavy tails. An early example of such a prior was proposed by Cutillo et al. (2008) in the context of wavelet thresholding where a heavier tail was attained by modeling $\theta \sim \mathcal{N}(0, \tau^2)$ and $\tau^2 \sim (\tau^2)^{-k}$ where $k > 1/2$. We list a few popular

TABLE 4
*Origin and tail behaviors of different priors*

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

global–local shrinkage priors along with their behavior near origin and the tails on Table 4 and plot the density functions on Fig. 2, and a discussion of the recent extensions of global–local priors beyond the Gaussian model is deferred to Section 8.

From a regularization view-point, one way to judge a prior is by the penalty it imposes on a likelihood (4), although in a strict Bayesian spirit, a prior should be evaluated based on the whole posterior, as shown by several authors including Castillo, Schmidt-Hieber and van der Vaart (2015) and van der Pas, Szabó and van der Vaart (2017). Although the horseshoe prior leads to optimal performance as a shrinkage prior, the induced penalty $\log \pi(\theta)$ does not admit a closed form as the marginal prior $\pi(\theta)$ is not analytically tractable. This poses a hindrance in learning via Expectation-Maximization or other similar algorithms. The generalized double Pareto prior of Armagan, Clyde and Dunson (2011) admits a closed form solution, but it does not have an infinite spike near zero needed for sparse recovery. Motivated by this fact, Bhadra et al. (2017a) recently proposed the 'horseshoe-like' prior by normalizing the tight bounds for the horseshoe prior. Thus, the horseshoe-like prior attains a unique status within its class: it has a closed form marginal prior for $\theta$, yet with a spike at origin and heavy tails and more importantly, admits a global–local scale mixture representation. The scale mixture representation supports both a traditional MCMC sampling for uncertainty quantification in full Bayes inference and EM/MM or proximal learning when computational efficiency is the primary concern.

Since the aim of designing a sparsity prior is achieving higher spike near zero while maintaining regularly varying tails, a useful strategy is to split the range of the prior into disjoint intervals: $[0, 1)$ and $[1, \infty)$, and aim for higher spike in one and heavier tail in the other. This leads to a class of 'horseshoe-like' priors with more flexibility in shape than any single shrinkage prior. We provide the general form of horseshoe-like priors and a key representation theorem. The proof that horseshoe-like prior is a scale mixture with Slash normal mixing density involves Frullani's probabilistic identity (*vide* Jeffreys and Swirles, 1972, pages 406–407), and to save substantial additional space we refer the readers to the proof in Section 5, Lemma 5.1 and Proposition 5.1 of Bhadra et al. (2017a).

*Horseshoe-like priors*: Bhadra et al. (2017a) have the following marginal prior density for $\theta_i$:

$$\tilde{p}_{\tilde{\text{HS}}}(\theta_i \mid \tau^2) = \frac{1}{2\pi\tau}\log\left(1+\frac{\tau^2}{\theta_i^2}\right), \quad \theta_i \in \mathbb{R}, \tau > 0. \tag{10}$$

The general family of horseshoe-like priors can be constructed as a density split into disjoint intervals as fol-

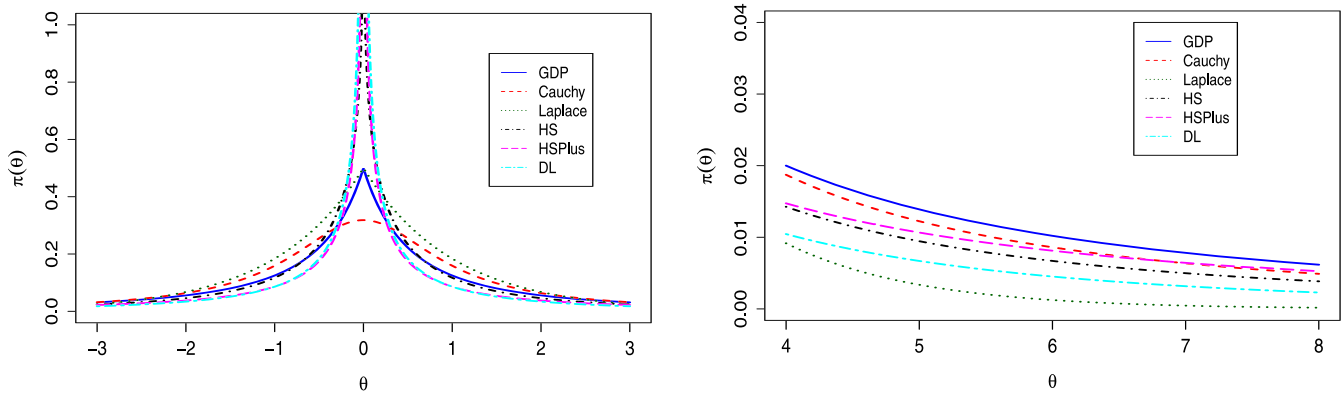



FIG. 2. *Marginal prior densities near the origin (left) and in the tail regions (right). The legends denote the horseshoe+ (HSPlus), horseshoe (HS), Dirichlet-Laplace (DL), generalized double Pareto (GDP), Cauchy and Laplace priors.*

lows:

$$
\begin{aligned}
&p_{\mathrm{hs}}(\theta_i \mid \tau^2) \\
&\quad \propto \begin{cases} \dfrac{1}{\theta_i^{1-\epsilon}} \log\left(1+\dfrac{\tau^2}{\theta_i^2}\right) & \text{if } |\theta_i| < 1, \\ \theta_i^{1-\epsilon} \log\left(1+\dfrac{\tau^2}{\theta_i^2}\right) & \text{if } |\theta_i| \geq 1, \end{cases} \\
&\quad \epsilon \geq 0, \tau > 0.
\end{aligned} \tag{11}
$$

*Normal scale mixture*: The horseshoe-like prior (10) is a Gaussian scale mixture with a Slash Normal mixing density, which is in turn another Gaussian scale mixture of Pareto(1/2) density, yielding the following representation theorem:

THEOREM 3 (Bhadra et al., 2017a). *The horseshoe-like prior in* (10) *has the following global–local scale mixture representation*:

$$
\begin{aligned}
&(\theta_i \mid t_i, \tau) \sim \mathcal{N}\left(0, \frac{\tau^2}{t_i^2}\right), \quad (t_i \mid s_i) \sim \mathcal{N}(0, s_i), \\
&s_i \sim \text{Pareto}\left(\frac{1}{2}\right), \quad t_i \in \mathbb{R}, \tau \geq 0.
\end{aligned} \tag{12}
$$

## 4. STATISTICAL RISK PROPERTIES

### 4.1 Inadmissibility of MLE

We briefly discuss the Stein shrinkage phenomenon as it provides an useful insight into the development of global–local shrinkage estimators in high-dimensional problems. The James–Stein (JS) estimator is $\hat{\boldsymbol{\theta}}_{\mathrm{JS}}(\mathbf{y}) = \{1 - (n-2)/\|\mathbf{y}\|^2\}\mathbf{y}$ which is equivalent to the empirical Bayes estimate $\hat{\theta}_{\mathrm{Bayes}} = \hat{\tau}^2/(\hat{\tau}^2+1)\mathbf{y}$, under i.i.d. $\mathcal{N}(0, \tau^2)$ priors on $\theta_i$ and $\hat{\tau}$ being the empirical Bayes estimate of $\tau$ from the data (Efron, 2010). Thus, the James–Stein estimator corresponds to the Bayes risk of $n\tau^2/(\tau^2+1) + 2/(1+\tau^2)$. We argue below that a global shrinkage rule such as the James–Stein estimator or $\ell_2$ regularization does not work in the sparse regime as it lacks local parameters for handling sparsity.

The story of shrinkage estimation goes back to the proof in Stein (1956) that the maximum likelihood estimators for normal data are inadmissible beyond $\mathbb{R}^2$. James and Stein (1961) proved that this estimator dominates the MLE in terms of the expected total squared error for every choice of $\boldsymbol{\theta}$, that is, it outperforms the MLE no matter what the true $\boldsymbol{\theta}$ is. To motivate the need for developing a local shrinkage rule, consider the classic James–Stein (JS) 'global' shrinkage rule, $\hat{\boldsymbol{\theta}}_{\mathrm{JS}}(\mathbf{y})$. The JS estimator uniformly dominates the traditional sample mean estimator, $\hat{\boldsymbol{\theta}}$. For all values of the true parameter $\boldsymbol{\theta}$ and for $n > 2$, we have the classical mean squared error (MSE) risk bound:

$$
\begin{aligned}
R(\hat{\boldsymbol{\theta}}_{\mathrm{JS}}, \boldsymbol{\theta}) &:= \mathbb{E}_{y|\boldsymbol{\theta}} \left\|\hat{\boldsymbol{\theta}}_{\mathrm{JS}}(\mathbf{y}) - \boldsymbol{\theta}\right\|^2 < n \\
&= \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}} \|\mathbf{y} - \boldsymbol{\theta}\|^2, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^n, n \geq 3.
\end{aligned}
$$

For sparse signal problem the standard James–Stein shrinkage rule, $\hat{\boldsymbol{\theta}}_{\mathrm{JS}}$, performs poorly. This is best seen in the sparse setting for a $r$-spike parameter value $\theta_r$ with $r$ coordinates at $\sqrt{n/r}$ which has $\|\theta\|^2 = n$. Johnstone and Silverman (2004) show that $E\|\hat{\boldsymbol{\theta}}_{\mathrm{JS}} - \boldsymbol{\theta}\| \leq n$ with risk 2 at the origin. This leads to a bound (for $\sigma^2 = 1$)

$$
\frac{n\|\boldsymbol{\theta}\|^2}{n + \|\boldsymbol{\theta}\|^2} \leq R(\hat{\boldsymbol{\theta}}_{\mathrm{JS}}, \theta_r) \leq 2 + \frac{n\|\boldsymbol{\theta}\|^2}{n + \|\boldsymbol{\theta}\|^2}.
$$

The lower bound is the risk of an 'ideal' linear estimator $\hat{\boldsymbol{\theta}}_c(\mathbf{y}) = c\mathbf{y}$. For an 'ideal' estimator, $\|\boldsymbol{\theta}\|$ is known and $c$ is chosen to minimize the MSE, which gives

$$
\tilde{c}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2/(n + \|\boldsymbol{\theta}\|^2). \tag{13}
$$

Theorem 5 of Donoho and Johnstone (1995) states the following result, an *oracle inequality* for the James–Stein estimator:

LEMMA 4. *Consider the 'ideal' estimator* $\tilde{\boldsymbol{\theta}}_{\mathrm{IS}}(\mathbf{y}) = \tilde{c}(\boldsymbol{\theta})(\mathbf{y})$ *in* (13). *For all* $p \geq 2$ *and for all* $\boldsymbol{\theta} \in \mathbb{R}^p$,

$$
\begin{aligned}
R\big(\hat{\boldsymbol{\theta}}_{\mathrm{JS}}(\mathbf{y}), \boldsymbol{\theta}_r\big) &\leq 2 + \inf_c R\big(\hat{\boldsymbol{\theta}}_c(\mathbf{y}), \boldsymbol{\theta}_r\big) \\
&= 2 + R\big(\tilde{\boldsymbol{\theta}}_{\mathrm{IS}}(\mathbf{y}), \boldsymbol{\theta}_r\big).
\end{aligned}
$$

Here, $\hat{\boldsymbol{\theta}}_{\mathrm{JS}}(\mathbf{y})$ for the $r$-spike parameter value has risk at least $R(\hat{\theta}^{\mathrm{JS}}, \theta_r) \geq (n/2)$. This is nowhere near optimal. As Donoho and Johnstone (1994) showed, simpler rules such as the hard-thresholding and soft-thresholding estimates given by $\hat{\boldsymbol{\theta}}^H(\mathbf{y}, \lambda) = \mathbf{y}I\{|\mathbf{y}| \geq \lambda\}$ and $\hat{\boldsymbol{\theta}}^S(\mathbf{y}, \lambda) = \mathrm{sgn}(\mathbf{y})(|\mathbf{y}| - \lambda)_+$ satisfy an oracle inequality. In particular, when the thresholding sequence is close to $\sqrt{2\log n}$ (universal threshold), these estimators attain the 'oracle risk' up to a factor of $2\log(n)$. Intuitively, this is not surprising as the high-dimensional normal prior places most of its mass on circular regions—and does not support sparse, spiky vectors. The James–Stein estimator was not built for sparse estimation and it is ambivalent to sparsity assumptions, but the shrinkage phenomenon in lower dimensional regime paves the way for building shrinkage rules for sparse regime, where one needs an additional 'local' shrinkage term to recover the signals.

### 4.2 Near Minimax Risk

The asymptotically minimax risk rate in $\ell_2$ 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 $\lim_{n\to\infty} a_n/b_n = 1$. Specifically, for any estimator $\delta(\mathbf{y})$, we have a lower bound ($\sigma^2 = 1$):

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \|\delta(Y) - \theta_0\|^2 \geq 2p_n \log(n/p_n)\big(1 + o(1)\big). \tag{14}$$

The minimax rate, which is a frequentist criteria for evaluating the convergence of point estimators to the underlying true parameter, is a validation criteria for posterior contraction as well. This result, due to Ghosal, Ghosh and van der Vaart (2000), showed that the minimax rate is the fastest that the posterior distribution can contract.

A key advantage of the horseshoe estimators is that they enjoy near-minimax rates in both an empirical Bayes and full Bayes approach, provided that the hyper-parameters or the priors are suitably chosen–as proved in a series of papers (van der Pas, Kleijn and van der Vaart, 2014, van der Pas, Salomond and Schmidt-Hieber, 2016, van der Pas, Szabó and van der Vaart, 2016, 2017). Specifically, for $\sigma^2 = 1$, the horseshoe estimator achieves

$$\sup_{\boldsymbol{\theta} \in \ell_0[p_n]} \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}} \|\hat{\theta}_{\mathrm{HS}}(\mathbf{y}) - \boldsymbol{\theta}\|^2 \asymp p_n \log(n/p_n), \tag{15}$$

van der Pas, Kleijn and van der Vaart (2014) showed that the near-minimax rate can be achieved by setting the global shrinkage parameter $\tau = (p_n/n)\log(n/p_n)$. In practice, $\tau$ is unknown and must either be estimated from the data or handled via a fully Bayesian approach by putting a suitable prior on $\tau$. van der Pas, Szabó and van der Vaart (2017) show that the theoretical optimality properties for the popular horseshoe prior holds true if the global shrinkage parameter $\tau$ is learned via the maximum marginal likelihood estimator (MMLE) or a full Bayes approach. Independently, van der Pas, Salomond and Schmidt-Hieber (2016) and Ghosh and Chakrabarti (2017) showed that these optimality properties are not unique features of the horseshoe prior and they hold for a general class of global–local shrinkage priors. While the results of van der Pas, Salomond and Schmidt-Hieber (2016) apply to a wider class of priors, including the horseshoe+ prior (Bhadra et al., 2017b) and spike-and-slab Lasso (Ročková and George, 2018), it is worth pointing out the difference between van der Pas, Salomond and Schmidt-Hieber (2016) and Ghosh and Chakrabarti (2017). van der Pas, Salomond and Schmidt-Hieber (2016) prove 'near-minimaxity' under 'uniform regular variation' conditions on the prior on local shrinkage parameters for a general class of global–local priors that allow exponential tails. On the other hand, Ghosh and Chakrabarti (2017) attain 'exact' minimaxity for 'horseshoe-type' priors under suitable conditions on the global parameter $\tau$, but they allow only polynomial tails, leading to a narrower class.

### 4.3 Variable Selection: Frequentist and Bayes Optimality

Here we compare the relative performance of horseshoe and Lasso for multiple testing under the two-groups model and a 0–1 additive loss framework. One of the main reasons behind the widespread popularity of Lasso is the in-built mechanism for performing simultaneous shrinkage and selection. The horseshoe estimator, on the other hand, is a shrinkage rule that induces a selection rule through thresholding the pseudo posterior inclusion probabilities. Datta and Ghosh (2013) proved that for large scale testing problems the horseshoe prior attains the oracle property while double-exponential tails prove to be insufficiently heavy, leading to a higher misclassification rate compared to the horseshoe prior. The main reasons behind the horseshoe prior's optimality are the posterior density of shrinkage weights that concentrates near 0 and 1 and the adaptability of the global shrinkage parameter $\tau$.

The posterior distribution under the horseshoe prior leads to a natural model selection strategy under the two-groups model. Carvalho, Polson and Scott (2010) argued that the shrinkage coefficient $1 - \hat{\kappa}_i$ can be viewed as a pseudo-inclusion probability $P(\theta_i \neq 0 \mid y_i)$ and induces a multiple testing rule:

$$\text{Reject the } i\text{th null hypothesis } H_{0i} : \theta_i = 0 \text{ if } 1 - \hat{\kappa}_i > \frac{1}{2}. \tag{16}$$

Under the two-groups model (24), and a 0–1 loss, the Bayes risk is

$$R = \sum_{i=1}^{n} \{(1-\pi)t_{1i} + \pi t_{2i}\},$$

where $t_{1i}$ and $t_{2i}$ denote the probabilities of type 1 and type 2 error corresponding to the $i$th hypothesis respectively.

If we know the true proportion of sparsity and the parameters of the non-null distribution, we can derive

a decision rule that is impossible to outperform in theory, which is called the Bayes oracle for multiple testing (Bogdan et al., 2011). The oracle risk serves as the lower bound for any multiple testing rule under the two-groups model and thus provides an asymptotic optimality criteria when the number of tests go to infinity. The framework of Bogdan et al. (2011) is:

$$\pi_n \to 0, \quad u_n = \psi_n^2 \to \infty, \quad \text{and} \quad \log(v_n)/u_n \to C \in (0, \infty), \tag{17}$$

where $v_n = \psi_n^2(\frac{1-\pi_n}{\pi_n})^2$. Dropping the subscript $n$ from parameters for notational simplicity, the Bayes risk for the Bayes oracle under the above framework (17) is:

$$R_{\text{oracle}} = n\pi\big(2\Phi(\sqrt{C}) - 1\big)\big(1 + o(1)\big).$$

A multiple testing rule is said to possess asymptotic Bayes optimality under sparsity (ABOS) if it attains the oracle risk as $n \to \infty$. Bogdan et al. (2011) provided conditions for a few popular testing rules, for example, Benjamini–Hochberg FDR controlling rule to be ABOS. Datta and Ghosh (2013) first showed that the horseshoe decision rule (16) is also ABOS up to a multiplicative constant if $\tau$ is chosen suitably to reflect the sparsity, namely $\tau = O(\pi)$. The proof in Datta and Ghosh (2013) hinges on the concentration of the posterior distribution near 0 and 1, depending on the trade-off between signal strength and sparsity. In numerical experiments, Datta and Ghosh (2013) also confirmed that the horseshoe decision rule outperforms the shrinkage rule induced by the double-exponential prior under various levels of sparsity. Although $\tau$ is treated as a tuning parameter that mimics $\pi$ in the theoretical treatment, in practice, $\pi$ is an unknown parameter. Several authors (Datta and Ghosh, 2013, Ghosh and Chakrabarti, 2017, Ghosh et al., 2016, van der Pas, Szabó and van der Vaart, 2016) have shown that usual estimates of $\tau$ adapts to sparsity, a condition that also guarantees near-minimaxity in estimation. Ghosh et al. (2016) extended the ABOS property to a wider class of global–local shrinkage priors, with conditions on the slowly varying tails of the local shrinkage prior. Ghosh et al. (2016) prove a stronger result, namely, the testing rule under a global–local prior attains the ABOS property *exactly*, when the global shrinkage parameter $\tau$ is of the same asymptotic order as the sparsity proportion $\pi$.

### 4.4 Sparse Linear Regression

One of the major advantages of Lasso and other frequentist penalized methods is their theoretical optimality properties in the regression setting $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$ (Bühlmann and van de Geer, 2011), whereas similar results for Bayesian methods using shrinkage priors are relatively less common. We review extant theoretical results for Bayesian sparse regression covering both point-mass mixture and continuous shrinkage priors.

*Point mass mixture priors*: Arguably the most notable contribution is due to for example, Castillo, Schmidt-Hieber and van der Vaart (2015), who showed that the posterior under a point-mass mixture prior contracts at the optimal rate for sparse parameter recovery and prediction, given a suitable 'compatibility' condition on the design matrix $\mathbf{X}$ is satisfied. Such compatibility conditions also govern oracle properties for Lasso-type methods, for example, 'irrepresentability' and 'mutual coherence' conditions (Bühlmann and van de Geer, 2011, *vide* Ch. 6) and (Zhao and Yu, 2006). Similarly, for recovery under point-mass mixture priors, Castillo, Schmidt-Hieber and van der Vaart (2015) define three local invertibility conditions on the regression matrix: $\bar{\phi}(s)$ (uniform compatibility in sparse vectors), $\tilde{\phi}(s)$ (smallest scaled sparse singular value), and $\text{mc}(\mathbf{X})$ (mutual coherence), for recovery with respect to $\ell_1$ norm, $\ell_2$ norm and $\ell_\infty$ norm respectively. We define the irrepresentability and mutual coherence condition below.

First, suppose the sample covariance matrix is denoted by $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$ and the active set $S = \{j : \theta_j \neq 0\}$ consists of first $s_0$ elements of $\boldsymbol{\theta}$ as in Definition 2. One can partition the $\hat{\Sigma}$ matrix as

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{s_0,s_0} & \hat{\Sigma}_{s_0,p-s_0} \\ \hat{\Sigma}_{p-s_0,s_0} & \hat{\Sigma}_{p-s_0,p-s_0} \end{bmatrix},$$

where $\hat{\Sigma}_{s_0,s_0}$ is the $s_0 \times s_0$ sub-matrix corresponding to the active variables. The strong irrepresentable condition for the variable selection consistency of Lasso is:

$$\big\|\hat{\Sigma}_{p-s_0,s_0}\hat{\Sigma}_{s_0,s_0}^{-1}\operatorname{sign}(\boldsymbol{\theta}_S)\big\|_\infty \leq 1 - \eta \quad \text{for positive constant vector } \eta. \tag{18}$$

Zhao and Yu (2006) illustrated the importance of strong irrepresentable condition on Lasso's model selection performance by showing that the probability of selecting the true sparse model is an increasing function of the irrepresentability condition number, defined as:

$$\eta_\infty = 1 - \big\|\hat{\Sigma}_{p-s_0,s_0}\hat{\Sigma}_{s_0,s_0}^{-1}\operatorname{sign}(\boldsymbol{\theta}_S)\big\|_\infty. \tag{19}$$

The strongest of these conditions, mutual coherence ($\mathrm{mc}(\mathbf{X})$), is defined as:

$$\mathrm{mc}(\mathbf{X}) = \max_{1\le i\ne j\le p} \frac{|\langle X_{\cdot,i} X_{\cdot,j}\rangle|}{\|X_{\cdot,i}\|_2 \|X_{\cdot,j}\|_2}. \tag{20}$$

Bühlmann and van de Geer (2011) establishes the relationship between the different conditions (*vide* Figure 6.1). Clearly, these optimality results carry over to the sparse normal means problem ('sequence model') where the design matrix is identity or to regression models with an orthogonal design matrix.

*Continuous shrinkage priors*: Polson and Scott (2011) point out that the one-group priors mimic Bayesian model averaging, where one achieves better predictive performance by averaging over models supported by data, without the computational burden. Several authors (Polson and Scott, 2011, 2012a, Datta and Ghosh, 2015) have shown empirically horseshoe outperforms Lasso (as well as Bayesian model averaging) in terms of out-of-sample predictive sum-of-squares error.

Armagan et al. (2013) proved posterior consistency in $p \le n$ situation for commonly used shrinkage prior including generalized double Pareto and horseshoe-type priors under simple sufficient conditions, for example, boundedness of the eigenvalues of $\mathbf{X}^T\mathbf{X}/n$ and the number of non-zero elements $p_n = o(n/\log n)$. Under similar conditions, minimax posterior contraction rates for the Dirichlet–Laplace prior (Bhattacharya et al., 2015) can be extended to the regression coefficients $\boldsymbol{\theta}$. Non-trivial extension to the high dimensional setting is still an active area.

There are some recent developments on theoretical properties for predictive risk and variable selection properties of the horseshoe posterior under a orthogonal design matrix in $p \le n$ situation. It is worth noting that there are two slightly different approaches for specifying the horseshoe prior. First, suppose a horseshoe prior is placed directly on the regression coefficient $\boldsymbol{\theta}$ where $p \le n$ under the model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\theta} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n), \\ \theta_j \mid \lambda_j, \tau, \sigma &\sim \mathcal{N}\big(0, \lambda_j^2 \tau^2 \sigma^2\big), \\ \lambda_j &\sim f(\cdot), \quad \tau \sim g(\cdot), \quad \sigma \sim h(\cdot). \end{aligned}$$

Tang et al. (2018) proposed the half-thresholding estimator,

$$\hat{\theta}_i^{\mathrm{HT}} = \hat{\theta}_i^{\mathrm{PM}} I\left(|\hat{\theta}_i^{\mathrm{PM}}/\hat{\theta}_i^{\mathrm{OLS}}| > \frac{1}{2}\right),$$

where $\hat{\theta}_i^{\mathrm{PM}}$ and $\hat{\theta}_i^{\mathrm{OLS}}$ are the posterior mean and the OLS solution, respectively, and showed this estimator achieves oracle property (variable selection consistency and optimal estimation) if local shrinkage priors have polynomial tails. On the other hand, Bhadra et al. (2016b) specifies the prior on a reparametrized $\boldsymbol{\alpha}$ follows (noting that $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are one-to-one functions for a fixed design $\mathbf{X}$):

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon \overset{\text{reparameterize}}{\Longrightarrow} \mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \epsilon, \\ &\text{where} \\ &\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{W}^T, \quad \text{(singular value decomposition)}, \\ &\mathbf{Z} = \mathbf{U}\mathbf{D}, \quad \boldsymbol{\alpha} = \mathbf{W}^T\boldsymbol{\theta}. \quad \big(\mathrm{Rank}(\mathbf{D}) = n\big). \end{aligned} \tag{21}$$

Under assumption of an orthogonal design, Bhadra et al. (2016b) investigated Stein's unbiased risk estimate for prediction, defined as $\mathrm{SURE} = \|\mathbf{y} - \tilde{\mathbf{y}}\|^2 + 2\sigma^2 \sum_{i=1}^n \frac{\partial \tilde{y}_i}{\partial y_i}$ for the horseshoe prior and proved that it leads to improved finite sample prediction risk, over ridge regression risk of $2n\sigma^2$.

THEOREM 5. *Prediction risk for the purely local horseshoe regression* (*Bhadra et al.*, 2016b). *Let* $\mathbf{D} = \mathbf{I}$ *in* (21) *and let the global shrinkage parameter in the horseshoe regression be* $\tau^2 = 1$. *When true* $\alpha_i = 0$, *an upper bound of the component-wise risk of the purely local horseshoe regression is* $1.75\sigma^2 < 2\sigma^2$.

As pointed out before, it remains to be settled whether stronger theoretical results hold for the horseshoe or other GL priors, e.g. whether oracle properties or minimaxity results under $\ell_2$ or $\ell_1$ norm carry over to horseshoe prior in the high-dimensional set-up under compatibility or coherence conditions on the design matrix as used by Bühlmann and van de Geer (2011) and Castillo, Schmidt-Hieber and van der Vaart (2015).

### 4.5 Uncertainty Quantification

Reliable uncertainty quantification is a key challenge in high-dimensional inference. While some authors (e.g., Chatterjee and Lahiri, 2011) observed that the Lasso-based estimates do not yield meaningful standard errors for the parameter estimates, Castillo, Schmidt-Hieber and van der Vaart (2015) showed poor posterior contraction for the Bayesian Lasso. These results motivate Bayesian approaches with appropriately heavy-tailed priors that produce automatic and reliable uncertainty quantification.

Chatterjee and Lahiri (2011) also proposed a bootstrap-based estimator for the limiting distribution

of the Lasso that attains some non-uniform consistency. Similarly, Liu and Yu (2013) argue that the bootstrap could be used. But these attempts are exposed to severe super-efficiency phenomena. In contrast, Zhang and Zhang (2014) pioneered the idea to de-bias the Lasso for obtaining an asymptotic Gaussian limiting distribution for single coordinates $\theta_i$ or other low-dimensional parameters of interest. The de-biased Lasso is of the form:

$$\hat{\boldsymbol{\theta}}^d = \hat{\boldsymbol{\theta}}^{\text{Lasso}} + \frac{1}{n}\mathbf{M}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}^{\text{Lasso}}),$$

The matrix $M$ is constructed from the node-wise Lasso, as also advocated by van de Geer et al. (2014) or using a convex program as proposed by Javanmard and Montanari (2014). In both approaches, exploiting KKT conditions for the node-wise Lasso or by construction, the $\ell_\infty$ norm $|\mathbf{M}\hat{\boldsymbol{\Sigma}} - \mathbf{I}|_\infty$ is controlling the bias and $[\mathbf{M}\hat{\boldsymbol{\Sigma}}\mathbf{M}]_{i,i}$ is governing the variance of the de-biased or de-sparsified Lasso.

Although one can always get confidence sets for a fixed coefficient, arguably a more specific question here is whether these credible sets (marginal credible intervals or credible $\ell_2$ balls) have both the minimax radius and the correct coverage. At the heart of these results is the impossibility theorem by Li (1989), that says one can not construct confidence sets to be both 'honest' and 'adaptive' uniformly for all $\theta_0$, be it Bayesian or non-Bayesian. In particular, sparsity-adaptive credible sets can not be 'honest' (Nickl and van de Geer, 2013, Li, 1989) in the sense that it is impossible to construct credible sets that have both their diameters adapt to the minimax rate for the unknown sparsity $\pi$ as well as provide nominal coverage probability over the full parameter space.

In the context of sequence models, as van der Pas, Szabó and van der Vaart (2017) point out that since the horseshoe prior achieves adaptive posterior contraction at the near-minimax rate $p_n \log(n/p_n)$ in (15) for nearly-black objects, one needs additional conditions, e.g., excessive bias-restriction (Belitser and Nurushev, 2015) or self-similarity to ensure good coverage. In particular, they prove that credible balls provide uncertainty quantification up to a correct multiplicative factor, provided the sparsity proportion $\pi$ crosses the detectability threshold, $\sqrt{2\log(n/p_n)}$. We refer the readers to Theorem 5 of van der Pas, Szabó and van der Vaart (2017) for a precise statement concerning the coverage and size of the horseshoe credible sets. It appears that there is a trade-off between honesty and adaptation, and Bayesian procedures such as the horseshoe attain adaptation over honesty and de-biased methods offer honesty, often by sacrificing the optimal diameter criterion.

## 5. HYPER-PARAMETERS

Careful handling of the global shrinkage parameter $\tau$ is critical for success of the horseshoe estimator in a sparse regime as it captures the level of sparsity in the data (Carvalho, Polson and Scott, 2010, Datta and Ghosh, 2013, van der Pas, Salomond and Schmidt-Hieber, 2016). However, in nearly black situations a naïve estimate of $\tau$ could collapse to zero, and care must be taken to prevent possible degeneracy in inference. There are two main approaches regarding choice of $\tau$: first, an empirical Bayesian approach that estimates $\tau$ from the data using a simple thresholding or maximum marginal likelihood approach (MMLE) and second, a fully Bayesian approach that specifies a hyper-prior on $\tau$.

### 5.1 Marginal Likelihood

We first take a closer look at how $\tau$ affects the marginal likelihood under the horseshoe prior and the maximum marginal likelihood approach of van der Pas, Szabó and van der Vaart (2017). We can write the marginal likelihood under the horseshoe prior after marginalizing out $\theta_i$ in (7) for $\sigma^2 = 1$ from the model as:

$$m(y \mid \tau) = \prod_{i=1}^{n} (1+\lambda_i^2\tau^2)^{-\frac{1}{2}} \exp\left\{-\frac{y_i^2}{2(1+\lambda_i^2\tau^2)}\right\} \times (1+\lambda_i^2)^{-1}\, d\lambda_i.$$

Tiao and Tan (1966) observe that the marginal likelihood is positive at $\tau = 0$, hence the impropriety of the prior of $\tau^{-2}$ at the origin translates to the posterior. As a result, the maximum likelihood estimator of $\tau$ could potentially collapse to zero in very sparse problems (Polson and Scott, 2011, Datta and Ghosh, 2013). In van der Pas, Szabó and van der Vaart (2017), both the empirical Bayes MMLE and the full Bayes solution are restricted in the interval $[1/n, 1]$ to preempt this behavior. To get the MMLE of $\tau$ using the approach of van der Pas, Szabó and van der Vaart (2017), we first calculate the marginal prior of $\theta_i$ after integrating out $\lambda_i^2$ in Equation (7):

$$p_\tau(\theta_i) = \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\theta_i^2}{2\lambda^2\tau^2}\right\} \frac{1}{\lambda\tau} \frac{2}{\pi(1+\lambda^2)}\, d\lambda.$$

The MMLE is then obtained as the maximizer of the marginal likelihood restricted to the interval $[1/n, 1]$:

$$\hat{\tau}_M = \underset{\tau \in [1/n,1]}{\operatorname{argmax}} \prod_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_i-\theta_i)^2}{2}\right\} \times p_\tau(\theta_i)\, d\theta_i.$$

The lower bound of the maximization interval prevents a degenerate solution of $\tau$ in the sparse case.

Handling $\tau$ is still an area of research: some papers (e.g., Carvalho, Polson and Scott, 2010, Datta and Ghosh, 2013, Piironen and Vehtari, 2017b) advocate using a full Bayes approach instead of a 'plug-in' maximum likelihood approach to avoid potential issues such as $\hat{\tau}$ collapsing to zero. On the other hand, van der Pas, Szabó and van der Vaart (2017) note the following:

"*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.*" (van der Pas, Szabó and van der Vaart, 2017, *vide* Rejoinder p. 1274).

In practice, the MMLE approach of van der Pas, Szabó and van der Vaart (2017) achieves both theoretical optimality and good numerical performance. It is computed over the interval $[1/n, 1]$, which connects to the interpretation of $\tau$ as sparsity and prevents any computational issues.

### 5.2 Optimization and Cross-Validation

In a recent paper, van der Pas, Szabó and van der Vaart (2017) have investigated the empirical Bayes and full Bayes approach for $\tau$, and have shown that the full Bayes and the MMLE estimators achieve the near minimax rate, namely $p_n \log(n)$, under similar conditions. For the full Bayes estimator, these conditions are easily seen to satisfied by a half-Cauchy prior truncated to the interval $[1/n, 1]$, which also does well in numerical experiments, both in 'sparse' and 'less-sparse' situations.

The MMLE estimator of van der Pas, Szabó and van der Vaart (2017) outperforms the simple thresholding estimator given by:

$$\hat{\tau}_s(c_1, c_2) = \max\left\{\frac{\sum_{i=1}^{n} \mathbf{1}\{|y_i| \geq \sqrt{c_1 \log(n)}\}}{c_2 n}, \frac{1}{n}\right\}.$$

Rather, the MMLE estimator can detect smaller nonzero signals, even those below the threshold $\sqrt{2\log(n)}$, such as $\theta_i = 1$ when $n = 100$.

A third approach could be treating $\tau$ as a tuning parameter and using a $k$-fold cross-validation to select $\tau$. As in the full Bayes and empirical Bayes approach, the cross-validated choice of $\hat{\tau}$ can also converge to zero and care should be taken to avoid such situations. Yet another approach for handling $\tau$ was proposed by Piironen and Vehtari (2017a), who have investigated the choice of $\tau$ for a linear regression model and have suggested choosing a prior for $\tau$ by studying the prior for $m_{\text{eff}} = \sum_{i=1}^{n}(1-\kappa_i)$, the effective number of non-zero parameters. When better prediction is desired, Bhadra et al. (2016b) suggest selecting $\tau$ by minimizing SURE, for which they provide an explicit form under the model in (21).

## 6. COMPUTATION

Over the last few years, several different implementations of the horseshoe prior for the normal means and regression models have been proposed. The MCMC based implementations usually proceed via block-updating $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ and $\tau$ using either a Gibbs, parameter expansion or slice sampling strategy. The first R package to offer horseshoe prior for regression along with Lasso, Bayesian Lasso and ridge was the `monomvn` package by Gramacy and Pantaleo (2010). In an unpublished technical report, Scott (2010) proposed a parameter expansion strategy for the horseshoe prior and studied its effect on the autocorrelation of $\tau$. Furthermore, Scott (2010) pointed out that the solution to this lies in marginalizing over the local shrinkage parameter $\lambda_j$'s. On a somewhat similar route, Makalic and Schmidt (2016) uses a inverse-gamma scale mixture identity to construct a Gibbs sampling scheme for the horseshoe and horseshoe+ priors for linear regression as well as logistic and negative binomial regressions.

The `horseshoe` package implements the MMLE and truncated prior approaches for handling $\tau$ proposed in van der Pas, Szabó and van der Vaart (2017). Hahn, He and Lopes (2016) proposed an elliptical slice sampler and argue that it outperforms Gibbs strategies for higher dimensional problems both in per-sample speed and quality of samples (i.e. effective sample size). The state-of-the-art implementation for the horseshoe prior in linear regression is by Bhattacharya, Chakraborty and Mallick (2016) who used a Gaussian sampling alternative to the naïve Cholesky decomposition to reduce the computational burden from $O(p^3)$ to $O(n^2 p)$. A very recent paper by Johndrow and Orenstein (2017) claims to improve this even further by implementing a block update strategy but using a random

TABLE 5
*Implementations of the horseshoe and other shrinkage priors*

| Implementation (Package/URL) | Authors |
|---|---|
| R package: `monomvn` | Gramacy and Pantaleo (2010) |
| R code in paper | Scott (2010) |
| R package: `horseshoe` | van der Pas et al. (2016) |
| R package: `fastHorseshoe` | Hahn, He and Lopes (2016) |
| MATLAB code | Bhattacharya, Chakraborty and Mallick (2016) |
| GPU accelerated Gibbs sampling | Terenin, Dong and Draper (2019) |
| `bayesreg` + MATLAB code in paper | Makalic and Schmidt (2016) |
| MATLAB code | Johndrow and Orenstein (2017) |
| R package:`bayeslm` | Hahn, He and Lopes (2019) |

walk Metropolis–Hastings algorithm on $\log(1/\tau^2)$ for block-updating $\tau \mid \lambda$. We provide a list of all implementations known to us on Table 5.

Bayesian methods using MCMC are sequential in nature. The methods are typically computation intensive, but one is able to perform probabilistic uncertainty quantification. However, sparse Bayesian methods including the horseshoe regression can be computed for $p \approx 10^6$, using parallel architecture of the latent variable representation to be able to retain the fully Bayesian nature via MCMC sampling. Terenin, Dong and Draper (2019) implement a horseshoe-probit regression using GPU that takes $\approx 2$ minutes for calculations involving a design matrix $\mathbf{X}$ of dimensions $10^6 \times 10^3$. If only point estimates are desired, of course Bayesian posterior modes can be computed as fast as penalized likelihood estimates (Bhadra et al., 2017a).

## 7. SIMULATION EXPERIMENTS

### 7.1 Effect of Correlated Predictors

As we discussed in Section 4.4, Lasso as well as Bayesian spike-and-slab priors can recover regression parameters under strong assumptions on the design matrix such as 'irrepresentability' or 'mutual coherence'. As van der Pas, Szabó and van der Vaart (2017) point out, such conditions are expected to be necessary for optimal recovery as in the context of spike-and-slab prior (Castillo, Schmidt-Hieber and van der Vaart, 2015).

For this simulation study, we follow the set-up in Zhao and Yu (2006) closely. Let $S = \{j : \theta_{j0} \neq 0\}$ be the active set of predictors, and let $s_0 = |S|$. We simulate data with $n = 100$, $p = 60$ and $s_0 = 7$ with the sparse coefficient vector $\boldsymbol{\theta}_S^* = (7, 5, 5, 4, 4, 3, 3)^T$. The error variance $\sigma^2$ was set to 0.1 to obey the asymptotic properties of the Lasso.

We first draw the covariance matrix $\Sigma$ from Wishart$(p, I_p)$ and then generate design matrix $\mathbf{X}$ from $\mathcal{N}(0, \Sigma)$. Zhao and Yu (2006) showed that the Strong Irrepresentability Condition (18) may not hold for such a design matrix. We generate 100 such design matrices to obtain a range of different $\eta_\infty$ values. In our simulation studies the $\eta_\infty$ values in (19) for the 100 simulated designs were between $[-0.86, 0.38]$. To see how the irrepresentability condition affects probability of selecting the correct model, 100 simulations were conducted for each design matrix. We compare four different methods: two penalized likelihood methods: Lasso, SCAD (Smoothly Clipped Absolute Deviation) (Fan and Li, 2001), and two Bayesian methods: horseshoe and Dirichlet–Laplace (Bhattacharya et al., 2015) in terms of percentage of these methods selecting the correct model. For model selection, we use the credible intervals for the horseshoe prior and $k$-means clustering for the Dirichlet–Laplace prior, following the simulation study in Bhattacharya et al. (2015).

Like Zhao and Yu (2006), we expect the Lasso to select the true model with a high probability when $\eta_\infty > 0$ and poorly when $\eta_\infty < 0$, with the sharpest ascent around the origin. We also calculated the mutual coherence (20) number for the same design matrices to see the effect on these two methods. The mc$(\mathbf{X})$ numbers were between $[0.21, 0.54]$.

Figure 3 shows the percentage of correctly selected model as a function of the irrepresentable condition number, $\eta_\infty$ and mutual coherence for the four candidates: Lasso, SCAD, horseshoe and Dirichlet–Laplace. For this simulation experiment, Lasso's model selection performance is dependent on the irrepresentability condition, deteriorating with decreasing $\eta_\infty$. Surprisingly, the effect is weaker for SCAD as well as both the horseshoe and the Dirichlet–Laplace priors.

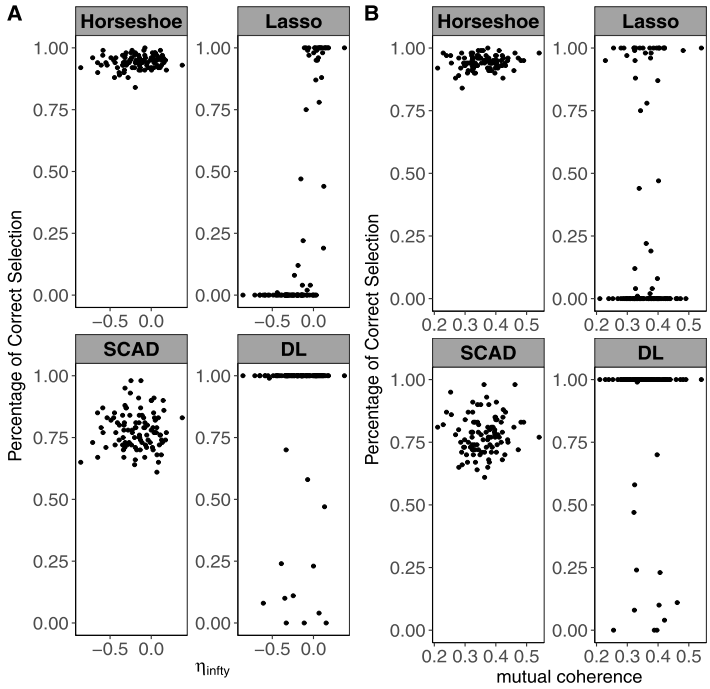



FIG. 3. *Effect of Strong Irrepresentability Condition $\eta_\infty$ (Panel A) and Mutual Coherence or maximum column correlation (Panel B) on the percentage of selecting the correct model by Lasso and SCAD penalties as well as the horseshoe and Dirichlet–Laplace (DL) priors.*

While the horseshoe almost always recovers the true sparse $\boldsymbol{\theta}$ vector irrespective of $\eta_\infty$, SCAD exhibits a high percentage (mean $= 0.75$, range $= [0.61, 0.98]$). Since we have calculated mutual coherence for the same design matrices, in this set-up it does not affect the horseshoe prior's variable selection, and its effect shows no clear pattern on any other candidates, apart from the Lasso.

### 7.2 Binary Response: Logistic Regression

We compare the performance of the horseshoe prior and Lasso for logistic regression for varying degree of dependence between the columns of a design matrix. We generated $n = 100$ binary observations for the standard logistic regression. The true parameter $\boldsymbol{\theta}^* \in \mathbb{R}^p$ where $p = 32$, $\boldsymbol{\theta}^*$ is sparse and has 5 non-zero elements $(7, 4, 2, 1, 1)$, and $\sigma^2$ was set to 0.1. We set the covariance matrix as $\Sigma_{ij} = \rho^{|i-j|}$ and then generate design matrix $\mathbf{X}$ from $\mathcal{N}(0, \Sigma)$ for 20 different values of $\rho \in [0.1, 0.9]$. Since the original horseshoe prior was not designed to handle the logistic likelihood, we use the Gaussian approximation method by Piironen and Vehtari (2017b), where they use a second-order Taylor expansion for the log posterior distribution. Piironen and Vehtari (2017b) also propose the regularized horseshoe prior where one introduces an additional slab width $c$ to allow for shrinkage even on the extreme tails. Following the recommendations of Piironen and Vehtari (2017b), we use the regularized horseshoe prior with a hyper-prior $c \sim$ Inv-Gamma$(2, 8)$ that corresponds to a Student-t$(0, 2^2)$ slab. We use 1000 posterior draws per chain with the NUTS algorithm in Stan. For Lasso, we use the `glmnet` package in R with a 10-folds cross-validation.

To compare the two methods for classification and predictive accuracy, we train the models on 80% of the data, with the remaining as test set and average the results over 50 random splits. We measure classification accuracy by the number of misclassified response $y_i$'s in test data. For predictive accuracy, we compare the mean log predictive density (MLPD) proposed in Gelman, Hwang and Vehtari (2014) as the mean of the computed log pointwise predictive density, defined as follows.

Let $\boldsymbol{\theta}^s$; $s = 1, \ldots, S$ be the posterior draws from $p(\boldsymbol{\theta} \mid \mathbf{y})$, and $\mathbf{y}_j$, $j = 1, \ldots, m$ be the $j$th test data, then MLPD is:

$$\text{MLPD} = \frac{1}{m} \sum_{j=1}^{m} \log\left( \frac{1}{S} \sum_{s=1}^{S} p(\mathbf{y}_j \mid \boldsymbol{\theta}^s) \right). \tag{22}$$

Figure 4(a) shows the average number of misclassified observations by horseshoe is a little lower than

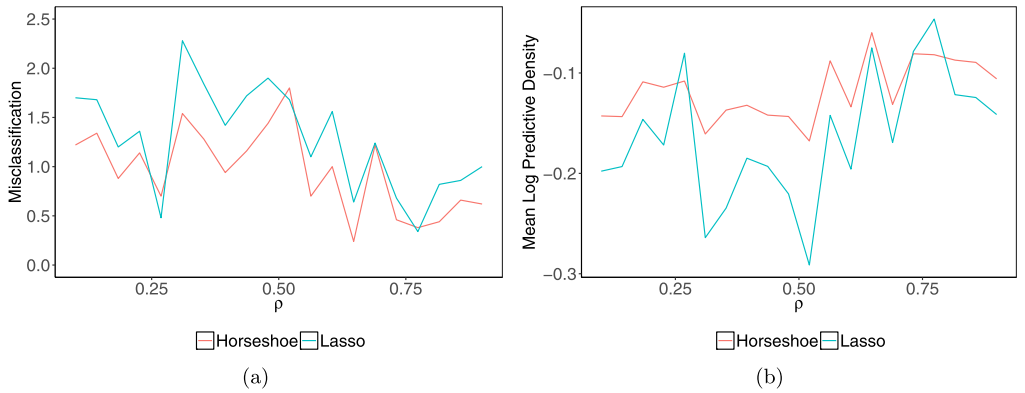



FIG. 4. (*a*) *Number of misclassified test data points and* (*b*) *mean log predictive density in* (22) *by the horseshoe and Lasso across different values of correlation* $\rho$, *where a higher value of* $\rho$ *represents higher dependence between the columns of* $\mathbf{X}$.

Lasso for all but two values of $\rho$. For the same values of $\rho$, Figure 4(b) shows that the predictive accuracy under the horseshoe prior is a little better than the Lasso. We direct the readers to Piironen and Vehtari (2017b) for a thorough comparison between the different variants of the horseshoe prior with Lasso for a few real data set as well as a synthetic data-set with a separable predictor.

## 8. FURTHER DEVELOPMENTS

### 8.1 Further Developments on Lasso

Since the inception of Lasso as a regularization method for linear regression in 1996, a great deal of extensions and applications have been proposed in the literature. The combined effect of convex penalty and sparsity of the final solution lead to huge computational gains by using powerful convex optimization methods on problems of massive dimensions. The coordinate descent approach (Friedman et al., 2007, Friedman, Hastie and Tibshirani, 2010) is one particularly promising approach, that works by applying soft-threshold to the least-squares solution obtained on partial residuals, one at a time. The coordinate descent approach is flexible and easy and can be proved to converge to the solution as long as the log-likelihood and penalty are convex (Tseng, 2001), paving the way for wide applicability of $\ell_1$ penalty in generalized linear models (GLM). The popular R package `glmnet` provides a nice and easy interface for applying Lasso and elastic-net penalty for a general sparse GLM.

### 8.2 Further Developments on Horseshoe

As discussed in Section 3.3, the horseshoe prior belongs to a wider class of global–local shrinkage priors (Polson and Scott, 2011) that are characterized by a local shrinkage parameter for recovering large signals and a global shrinkage parameter for adapting to overall sparsity. The class of global–local priors, although differing in their specific goals and design, exhibit some common features: heavy tails for tail-robustness and appreciable mass near zero for sparsity, leading to shared optimality properties.

Although the original horseshoe prior was developed for signal recovery with sparse Gaussian means, the idea of directly modeling the posterior inclusion probabilities and the use of normal-scale mixtures to facilitate sparsity is very flexible and can be easily generalized to a wider class of problems. Bhadra et al. (2016c) show that the horseshoe prior is a good candidate as a default prior for low-dimensional, possibly non-linear functionals of high-dimensional parameter and can resolve long-standing marginalization paradoxes for such problems. Bhadra et al. (2016b) show how to use global–local priors for prediction and provide theoretical and numerical evidence that it performs better than a variety of competitors including Lasso, ridge, PCR and sparse PLS.

Moving beyond Gaussianity, Datta and Dunson (2016) re-discovered the Gauss hypergeometric prior for flexible shrinkage needed for quasi-sparse count data, with a tighter control on false discoveries. Piironen and Vehtari (2017a) used a Gaussian approximation using a second-order Taylor expansion for the log-likelihood to apply the horseshoe prior in generalized linear models. Wang and Pillai (2013) proposed a shrinkage prior based on a scale mixture of uniform for covariance matrix estimation. Peltola et al. (2014) applied the horseshoe prior for Bayesian linear survival regression for selecting covariates with highest predictive values. A sample of the many applications of the

TABLE 6
*Applications of the horseshoe prior*

| Application | Authors |
|---|---|
| *Fadeout* method for mean-field variational inference under non-centered parameterizations and stochastic variational inference for undirected graphical model. | Ingraham and Marks (2016) |
| Linear regression for Causal inference and Instrumental variable models | Hahn, He and Lopes (2018, 2016) |
| Multiclass prediction using DOLDA (Diagonally orthant Latent Dirichlet Allocation) | Magnusson, Jonsson and Villani (2016) |
| Mendelian Randomization to detect causal effects of interest | Berzuini et al. (2016) |
| Locally adaptive nonparametric curve fitting with shrinkage prior Markov random field (SPMRF) | Faulkner and Minin (2015) |
| Quasi-Sparse Count Data | Datta and Dunson (2016) |
| Variable Selection under the projection predictive framework | Piironen and Vehtari (2015) |
| Dynamic shrinkage Process (dynamic linear model and trend filtering) | Kowal, Matteson and Ruppert (2017) |
| Logistic regression with horseshoe prior | Piironen and Vehtari (2017b), Wei (2017) |
| Tree ensembles with rule structured horseshoe regularization | Nalenz and Villani (2018) |
| Bayesian compression for deep learning | Louizos, Ullrich and Welling (2017) |
| Precision matrix estimation | Li, Craig and Bhadra (2017) |

horseshoe prior is given in Table 6. Given the explosive growth of methodology in this area, we conjecture that the horseshoe prior would be regarded as a key tool for sparse signal recovery and as a default prior for objective Bayesian inference in many important problems.

## 9. DISCUSSION

Sparsity can be achieved with Lasso and horseshoe regularization, a member of the class of global–local shrinkage priors. The horseshoe prior offers better computational efficiency than the Bayesian two-group priors, while still mimicking the inference and it outperforms the estimator based on Laplace prior, the Bayesian dual of Lasso. The intuitive reasons for better performance by the horseshoe prior are its heavy tails and probability spike at zero, which make it adaptive to sparsity and robust to large signals. A number of computing strategies have been proposed for both the Lasso and the horseshoe prior, based on variants of coordinate descent and MCMC respectively. We have outlined the distinct algorithmic implementations in Section 6 and Table 5. Since the goal of Lasso-based estimator is to produce a point estimate, rather than samples from the full posterior distribution of the underlying parameter, Lasso-based methods are typically faster than the horseshoe and related shrinkage priors.

The lack of speed can be overcome easily by employing a strategy based on expectation-maximization or proximal algorithm, which is often faster than the Lasso or other penalty based methods, for example the EM algorithm proposed in Section 4 of Bhadra et al. (2017a) is orders of magnitude faster than the non-convex SCAD or MCP (Bhadra et al., 2017a, *vide* Table 1). Another fruitful strategy is to employ proximal algorithms similar to expectation-maximization (Polson, Scott and Willard, 2015). These algorithms can be specifically designed to achieve better estimation and prediction error compared to Lasso (Bhadra et al., 2017a) by using clever decompositions of the objective function and some convenient properties (e.g., strong convexity) of the resulting parts. As discussed before, an active area of research is designing algorithms to handle Bayesian shrinkage in big data problems, for example, using GPU-accelerated computing (Terenin, Dong and Draper, 2019).

We have discussed the theoretical optimality properties for both Lasso and horseshoe estimators. The optimality properties of Lasso in regression are well-known and they depend on the 'neighborhood stability' or 'irrepresentability' condition (18) and the 'beta-min' condition. Similarly, adaptive posterior concentration for horseshoe depends on 'excessive bias restriction', a condition analogous to the 'beta-min' condition. Although horseshoe regression has not been studied to the same depth as penalized regression, it is expected that optimality will depend on conditions that guarantee against ill-posed design matrix and separability of signal and noise parameters. For the sequence model, the horseshoe posterior mean enjoys near-minimaxity in estimation, and the induced decision rule achieves asymptotic Bayes optimality for multiple testing as discussed in Section 4.

The horseshoe estimator of the sampling density converges to the the true sampling density $p(y \mid \theta_0)$ at a super-efficient rate at $\theta_0 = 0$, compared to any Bayes estimator with a bounded prior density at the

origin (Carvalho, Polson and Scott, 2010, *vide* Theorem 4). The rate of convergence of the Cesáro-average Bayes risk at $\theta_0 = 0$ for horseshoe is $O(n^{-1}(\log n - b \log\log n))$. This is called the 'Kullback–Leibler super-efficiency' in true density recovery for the horseshoe estimator. The horseshoe priors are also good default priors for many-to-one functionals as shown in Bhadra et al. (2016c), but a thorough study of horseshoe prior for default Bayes problems is still an unexplored area. We end the current article with a few other possible directions for future investigations.

(i) The square-root Lasso (Belloni, Chernozhukov and Wang, 2011) or scaled Lasso (Sun and Zhang, 2012) improves over the Lasso by making the inference ambivalent towards $\sigma$, while making the estimator scale-invariant. It might be interesting to study the effect of marginalizing the global parameters such as $\tau$ and $\sigma$ on inference from shrinkage priors. Our preliminary investigation suggests that scaling the prior on $\tau$ by $\sigma$ or marginalizing out $\sigma$ improves the robustness of the shrinkage priors.

(ii) One promising area is to extend the inferential capacity for the exponential family, and whether or not the optimality properties carry over to the non-Gaussian cases. Some early research on this is Datta and Dunson (2016) and Wei (2017).

(iii) Another interesting direction could include structured sparsity under the horseshoe prior, such as grouped variable selection and Gaussian graphical models, as explored in Li, Craig and Bhadra (2017).

## APPENDIX A: TWO-GROUPS MODEL

The two-groups model is a natural hierarchical Bayesian model for the sparse signal-recovery problem. The two-groups solution to the signal detection problem is as follows:

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

(ii) Calculate the posterior probabilities 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 $\pi$ and then carry out the tests on the basis of the posterior inclusion probabilities $\omega_i = P(\theta_i \neq 0 \mid y_i)$. Formally, in a two-groups model $\theta_i$'s are modeled as

$$\theta_i \mid \pi, \psi = (1-\pi)\delta_0 + \pi\mathcal{N}\big(0, \psi^2\big), \tag{23}$$

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 these assumptions, the marginal distribution of $(y_i \mid \pi, \psi)$ is given by:

$$y_i \mid \pi, \psi \sim (1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}\big(0, 1+\psi^2\big). \tag{24}$$

From (24), we see that the two-groups model leads to a sparse estimate, that is, it puts exact zeros in the model.

## APPENDIX B: PROOF OF EQUATION (3.5)

Assume $\sigma^2 = 1$ without loss of generality. The hierarchical model for horseshoe prior is $y_i \sim \mathcal{N}(\theta_i, 1)$ and $\theta_i \sim \mathcal{N}(0, \lambda_i^2\tau^2)$. Using Bayes' rule, posterior density of $\theta_i$ is Gaussian with mean $(1-\kappa_i)y_i$ where $\kappa_i = 1/(1+\lambda_i^2\tau^2)$. It follows from Fubini's theorem:

$$\begin{aligned} E(\theta_i \mid y_i) &= \int_0^1 (1-\kappa_i) y_i \, p(\kappa_i \mid y_i)\, d\kappa_i \\ &= \big\{1 - E(\kappa_i \mid y_i)\big\} y_i. \end{aligned}$$

## APPENDIX C: SHRINKAGE PROFILES

We compare the shrinkage functions for Lasso, ridge, and the horseshoe estimator with that of the post-lava estimator (Chernozhukov, Hansen and Liao, 2017). The shrinkage functions for these methods are given below:

$$\begin{aligned} d_{\text{lasso}}(z) &= \operatorname*{argmin}_{\theta\in\mathbb{R}} \big\{(z-\theta)^2 + \lambda_l|\theta|\big\} \\ &= \big(|z| - \lambda_l/2\big)_+ \operatorname{sgn}(z), \end{aligned} \tag{25}$$

$$\begin{aligned} d_{\text{ridge}}(z) &= \operatorname*{argmin}_{\theta\in\mathbb{R}} \big\{(z-\theta)^2 + \lambda_r\theta^2\big\} \\ &= (1+\lambda_r)^{-1} z, \end{aligned} \tag{26}$$

$$d_{\text{post-lava}}(z) = \begin{cases} z & |z| > \lambda_1/2k, \\ (1-k)z & |z| \le \lambda_1/2k, \end{cases} \tag{27}$$

where $k = \lambda_2/(1+\lambda_2)$,

$$d_{\text{horseshoe}}(z) = z\bigg(1 - \frac{2\Phi_1(1/2, 1, 5/2, z^2/2, 1-1/\tau^2)}{3\Phi_1(1/2, 1, 3/2, z^2/2, 1-1/\tau^2)}\bigg). \tag{28}$$

Figure 1(b) shows the post-lava and the horseshoe shrinkage function along with Lasso and ridge shrinkage functions for $z > 0$. Although a theoretical analysis is beyond the scope of the current article, we can see the similarities between the lava and horseshoe shrinkage. They both shrink aggressively for small values

of $z$ and provide robustness for large signals $z$, as the shrinkage function becomes closer to the 45° line.

## ACKNOWLEDGEMENTS

We thank the AE and two anonymous referees for constructive suggestions. Bhadra and Polson are partially supported by Grant No. DMS-1613063 by the US National Science Foundation.

## REFERENCES

ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36** 99–102. MR0359122

ARMAGAN, A., CLYDE, M. and DUNSON, D. B. (2011). Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems* 523–531.

ARMAGAN, A., DUNSON, D. B. and LEE, J. (2013). Generalized double Pareto shrinkage. *Statist. Sinica* **23** 119–143. MR3076161

ARMAGAN, A., DUNSON, D. B., LEE, J., BAJWA, W. U. and STRAWN, N. (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika* **100** 1011–1018. MR3142348

BAI, R. and GHOSH, M. (2017). The inverse gamma–gamma prior for optimal posterior contraction and multiple hypothesis testing. arXiv preprint, arXiv:1710.04369.

BELITSER, E. and NURUSHEV, N. (2015). Needles and straw in a haystack: Robust confidence for possibly sparse sequences. arXiv preprint, arXiv:1511.01803.

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

BERZUINI, C., GUO, H., BURGESS, S. and BERNARDINELLI, L. (2016). Mendelian randomization with poor instruments: A Bayesian approach. arXiv:1608.02990 [math, stat], Aug. arXiv:1608.02990.

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2016a). Global–local mixtures. arXiv preprint arXiv:1604.07487.

BHADRA, A., DATTA, J., LI, Y., POLSON, N. G. and WILLARD, B. (2016b). Prediction risk for the horseshoe regression. arXiv preprint arXiv:1605.04796.

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2016c). Default Bayesian analysis with global–local shrinkage priors. *Biometrika* **103** 955–969. MR3620450

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2017a). Horseshoe regularization for feature subset selection. arXiv preprint arXiv:1702.07400.

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2017b). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* **12** 1105–1131. MR3724980

BHATTACHARYA, A., CHAKRABORTY, A. and MALLICK, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103** 985–991. MR3620452

BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805

BOGDAN, M., CHAKRABARTI, A., FROMMLET, F. and GHOSH, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.* **39** 1551–1579. MR2850212

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761

CANDÈS, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris* **346** 589–592. MR2412803

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$. *Ann. Statist.* **35** 2313–2351. MR2382644

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. *J. Mach. Learn. Res.* **5** 73–80.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874

CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. MR2847974

CHERNOZHUKOV, V., HANSEN, C. and LIAO, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *Ann. Statist.* **45** 39–76. MR3611486

CUTILLO, L., JUNG, Y. Y., RUGGERI, F. and VIDAKOVIC, B. (2008). Larger posterior mode wavelet thresholding and applications. *J. Statist. Plann. Inference* **138** 3758–3773. MR2455964

DATTA, J. and DUNSON, D. B. (2016). Bayesian inference on quasi-sparse count data. *Biometrika* **103** 971–983. MR3620451

DATTA, J. and GHOSH, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Anal.* **8** 111–131. MR3036256

DATTA, J. and GHOSH, J. K. (2015). In search of optimal objective priors for model selection and estimation. In *Current Trends in Bayesian Methodology with Applications* 225–243. CRC Press, Boca Raton, FL. MR3644673

DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. With discussion and a reply by the authors. MR1157714

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

EFRON, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics (IMS) Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. MR2060166

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAULKNER, J. R. and MININ, V. N. (2015). Bayesian trend filtering: Adaptive temporal smoothing with shrinkage priors.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. MR3253850

GEORGE, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.* **95** 1304–1308. MR1825282

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

GHOSH, P. and CHAKRABARTI, A. (2017). Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Anal.* **12** 1133–1161. MR3724981

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 Anal.* **11** 753–796. MR3498045

GORDY, M. B. (1998). Computationally convenient distributional assumptions for common-value auctions. *Comput. Econ.* **12** 61–78.

GRAMACY, R. B. and PANTALEO, E. (2010). Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. *Bayesian Anal.* **5** 237–262. MR2719652

GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. MR2596440

HAHN, P. R., HE, J. and LOPES, H. (2016). Elliptical slice sampling for Bayesian shrinkage regression with applications to causal inference. Tech. rept.

HAHN, P. R., HE, J. and LOPES, H. (2018). Bayesian factor model shrinkage for linear IV regression with many instruments. *J. Bus. Econom. Statist.* **36** 278–287. MR3790214

HAHN, P. R., HE, J. and LOPES, H. F. (2019). Efficient sampling for Gaussian linear regression with arbitrary priors. *J. Comput. Graph. Statist.* **28** 142–154. MR3939378

HANS, C. (2011). Elastic net regression modeling with the orthant normal prior. *J. Amer. Statist. Assoc.* **106** 1383–1393. MR2896843

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations. Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

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

INGRAHAM, J. B. and MARKS, D. S. (2016). Bayesian sparsity for intractable distributions. arXiv preprint, arXiv:1602.03807.

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics* **103**. Springer, New York. MR3100153

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

JEFFREYS, H. and SWIRLES, B. (1972). *Methods of Mathematical Physics*, 3rd ed. Cambridge Univ. Press, Cambridge. MR1744997

JOHNDROW, J. E. and ORENSTEIN, P. (2017). Scalable MCMC for Bayes shrinkage priors. arXiv preprint, arXiv:1705.00841.

JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

JOLLIFFE, I. T., TRENDAFILOV, N. T. and UDDIN, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12** 531–547. MR2002634

KOWAL, D. R., MATTESON, D. S. and RUPPERT, D. (2017). Dynamic shrinkage processes. arXiv preprint, arXiv:1707.00763.

LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135

LI, Y., CRAIG, B. A. and BHADRA, A. (2017). The graphical horseshoe estimator for inverse covariance matrices. arXiv preprint, arXiv:1707.06661.

LIU, H. and YU, B. (2013). Asymptotic properties of Lasso + mLS and Lasso + Ridge in sparse high-dimensional linear regression. *Electron. J. Stat.* **7** 3124–3169. MR3151764

LOUIZOS, C., ULLRICH, K. and WELLING, M. (2017). Bayesian compression for deep learning. In *Advances in Neural Information Processing Systems* 3290–3300.

MAGNUSSON, M., JONSSON, L. and VILLANI, M. (2016). DOLDA—A regularized supervised topic model for high-dimensional multi-class regression. arXiv preprint arXiv:1602.00260, Jan.

MAKALIC, E. and SCHMIDT, D. F. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. arXiv preprint, arXiv:1611.06649.

MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769

MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors. MR0997578

NALENZ, M. and VILLANI, M. (2018). Tree ensembles with rule structured horseshoe regularization. *Ann. Appl. Stat.* **12** 2379–2408. MR3875705

NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

PELTOLA, T., HAVULINNA, A. S., SALOMAA, V. and VEHTARI, A. (2014). Hierarchical Bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop*, *Vol.* 1218 79–88 CEUR-WS.org.

PIIRONEN, J. and VEHTARI, A. (2015). Projection predictive variable selection using Stan + R. arXiv preprint arXiv:1508.02502, Aug.

PIIRONEN, J. and VEHTARI, A. (2017a). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Artificial Intelligence and Statistics* 905–913.

PIIRONEN, J. and VEHTARI, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11** 5018–5051. MR3738204

POLSON, N. G. and SCOTT, J. G. (2010). Large-scale simultaneous testing with hypergeometric inverted-beta priors. arXiv preprint, arXiv:1010.5223.

POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* 9 501–538. Oxford Univ. Press, Oxford. With discussions by Bertrand Clark, C. Severinski, Merlise A. Clyde, Robert L. Wolpert, Jim e. Griffin, Philiip J. Brown, Chris Hans, Luis R. Pericchi, Christian P. Robert and Julyan Arbel. MR3204017

POLSON, N. G. and SCOTT, J. G. (2012a). Local shrinkage rules, Lévy processes and regularized regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 287–311. MR2899864

POLSON, N. G. and SCOTT, J. G. (2012b). On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.* **7** 887–902. MR3000018

POLSON, N. G., SCOTT, J. G. and WILLARD, B. T. (2015). Proximal algorithms in statistics and machine learning. *Statist. Sci.* **30** 559–581. MR3432841

ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476

SCOTT, J. G. (2010). Parameter expansion in local-shrinkage models. arXiv preprint, arXiv:1010.5265.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley and Los Angeles. MR0084922

STEPHENS, M. and BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10** 681–690.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

TANG, X., GHOSH, M., XU, X. and GHOSH, P. (2018). Bayesian variable selection and estimation based on global–local shrinkage priors. *Sankhya A* **80** 215–246. MR3850065

TERENIN, A., DONG, S. and DRAPER, D. (2019). GPU-accelerated Gibbs sampling: A case study of the horseshoe probit model. *Stat. Comput.* **29** 301–310. MR3914622

TIAO, G. C. and TAN, W. Y. (1966). Bayesian analysis of random-effect models in the analysis of variance. II. Effect of autocorrelated errors. *Biometrika* **53** 477–495. MR0214218

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. (2014). In praise of sparsity and convexity. In *Past, Present, and Future of Statistical Science* 497–505.

TIBSHIRANI, R. J., HOEFLING, H. and TIBSHIRANI, R. (2011). Nearly-isotonic regression. *Technometrics* **53** 54–61. MR2791946

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

TIKHONOV, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Sov. Math., Dokl.* **4** 1035–1038.

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

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. *Electron. J. Stat.* **8** 2585–2618. MR3285877

VAN DER PAS, S. L., SALOMOND, J.-B. and SCHMIDT-HIEBER, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.* **10** 976–1000. MR3486423

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.

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11** 3196–3225. MR3705450

VAN DER PAS, S., SCOTT, J., CHAKRABORTY, A. and BHATTACHARYA, A. (2016). horseshoe: Implementation of the horseshoe prior. R package version 0.1.0.

WANG, H. and PILLAI, N. S. (2013). On a class of shrinkage priors for covariance matrix estimation. *J. Comput. Graph. Statist.* **22** 689–707. MR3173737

WEI, R. (2017). *Bayesian Variable Selection Using Continuous Shrinkage Priors for Nonparametric Models and Non-Gaussian Data*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.), North Carolina State Univ. MR3797433

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, Y., REICH, B. J. and BONDELL, H. D. (2016). High dimensional linear regression via the R2–D2 shrinkage prior. arXiv preprint arXiv:1609.00046.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327