Submitted to Statistical Science

Lasso Meets Horseshoe: A Survey

Anindya Bhadra
Purdue University

Jyotishka Datta
University of Arkansas

Nicholas G. Polson
University of Chicago
Booth School of Business

Brandon Willard
University of Chicago
Booth School of Business

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..

MSC 2010 subject classifications: Primary 62J07, 62J05, sparsity, regression, sparsity,, regression,, Lasso,, global-local priors,, horseshoe,, horseshoe+,, regularization,, hyper-parameter tuning.; secondary 62H15, 62F03.

1. INTRODUCTION

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

Sparsity and ultra-sparsity rely 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.

250 N. University St., West Lafayette, IN 47907. (e-mail: bhadra@purdue.edu)
1 University of Arkansas, Fayetteville, AR 72704. (e-mail: jd033@uark.edu)
5807 S. Woodlawn Ave., Chicago, IL 60637. (e-mail: ngp@chicagobooth.edu)
5807 S. Woodlawn Ave., Chicago, IL 60637. (e-mail: bwillard@uchicago.edu)

(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 ℓ_1 norm of the parameter vector. Lasso's widespread popularity is due to a multitude of factors, in particular its computational efficiency of Least Angle Regression method (Efron *et al.*, 2004) and 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 & van de Geer, 2011; James *et al.*, 2013; Hastie *et al.*, 2015). Table 1, adapted from Tibshirani (2014), gives a list of popular regularization methods based on Lasso.

TABLE 1
Lasso regularization methods

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

Bayes procedures, on the other hand, can be classified into two categories: discrete mixtures or two-groups model or spike-and-slab priors (Johnstone & Silverman, 2004; Efron, 2010, 2008; Bogdan *et al.*, 2011) and shrinkage priors (Bhadra *et al.*, 2017a; Armagan *et al.*, 2011, 2013a; Carvalho *et al.*, 2009, 2010; Griffin & Brown, 2010; Polson & Scott, 2010b; Castillo & van der Vaart, 2012). 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 shrink the entire coefficient towards zero. Table 2 provides a sampling of a few continuous shrinkage prior popular in the literature. Both these approaches have their own advantages and caveats, which we discuss in turn. A key duality being that a penalized approach can be interpreted as Bayesian mode of the posterior distribution under an appropriate shrinkage prior (Polson & Scott, 2016).

Both Lasso and horseshoe procedures come with strong theoretical guarantees for estimation, prediction and variable selection. Both procedures possess asymptotic Oracle properties, i.e. identify the true non-zero coefficients as well

TABLE 2
A catalog of Horseshoe and GL shrinkage priors

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

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). The horseshoe has been shown to achieve oracle properties in variable selection (Datta & Ghosh, 2013) and near-minimaxity in estimation (van der Pas *et al.*, 2017) and improved prediction performance in linear regression (Bhadra *et al.*, 2016c), although theoretical studies of the horseshoe 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 an optimization perspective 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 hyper-parameter selection and computational strategies. Section 7 provides two simulation experiments comparing 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 would be to simultaneously test if θ_i's are zero or not. We are interested in the sparse paradigm where a large proportion of the parameter vector contains zeros. The 'ultra-sparse' (Bhadra *et al.*, 2017a) or 'nearly black' (Donoho *et al.*, 1992) regime occurs when the parameter vector θ lies in the set $l_0[p_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq p_n\}$ with the upper bound on the number of non-zero parameter values $p_n = o(n)$ as $n \to \infty$.

A natural 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 θ_i's (*vide* Appendix A). The inference is then based on the posterior probabilities of non-zero θ_i's based on the discrete mixture model. The two-groups model

possesses a number of frequentist and Bayesian optimality properties. Johnstone & 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 ℓ_q norm for $q \in (0, 2]$ for $\boldsymbol{\theta}$ that are either nearly black or belong to an ℓ_p ball of 'small' radius. Castillo & van der Vaart (2012) treated a full Bayes version of the problem and again found an estimate that is minimax in ℓ_q norm for mean vectors that are either nearly black or have bounded weak ℓ_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 $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_p]$ is a $p \times n$ 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 θ_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:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{i,j} \right)^2 + \text{pen}_\lambda(\boldsymbol{\theta}), \tag{2.1}$$

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

Lasso uses an ℓ_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 ℓ_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 MCP (Zhang, 2010) or SCAD (Fan & Li, 2001), that act as a tool for variable selection and estimation. Penalization methods can be viewed in terms of the posterior modes they imply under an induced prior relating to the penalty in (2.1)–via $p(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta})$, where $\pi(\cdot)$ is a suitable prior for $\boldsymbol{\theta}$. 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 the convexity and sparsity entailed. The 'bet on sparsity' principle (Hastie *et al.*, 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 et al. (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 π_S on $\mathbb{R}^S$. The prior can be written as in Castillo *et al.* (2015):

$$(S, \boldsymbol{\theta}) \mapsto \binom{p}{|S|}^{-1} \pi_p(|S|)\pi_S(\boldsymbol{\theta}_S)\delta_0(\boldsymbol{\theta}_{S^c}) \tag{2.2}$$

Bayesian approaches for sparse linear regression include George (2000); George & Foster (2000); Mitchell & Beauchamp (1988); Ishwaran & Rao (2005) and more recently Ročková & George (2016), who introduce the spike-and-slab Lasso prior, where the hierarchical prior on the parameter and model spaces assumes the form:

$$\pi(\boldsymbol{\theta} \mid \gamma) = \prod_{i=1}^{p} [\gamma_i \pi_1(\theta_i) + (1 - \gamma_i)\pi_0(\theta_i)], \quad \gamma \sim p(\cdot), \tag{2.3}$$

where $\boldsymbol{\gamma}$ indexes the 2^p possible models, and π_0, π_1 model the null and non-null θ_i's respectively using two Laplace priors with different scales.

The continuous one-group shrinkage 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. The continuous shrinkage prior leads to faster computation relative to the spike-and-slab type priors 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. In comparison, global-local shrinkage priors typically lead to 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 & van der Vaart (2012). Ročková & George (2016) 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 & George, 2016) 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.

Carvalho *et al.* (2009); Polson & Scott (2010b); Carvalho *et al.* (2010); Polson & Scott (2012b) introduced the 'global-local' shrinkage priors. Global-local priors

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 a varying degree of theoretical and numerical performance. We compare these different priors and introduce a recently proposed family of horseshoe-like priors in §3.3.

The estimators resulting from the one-group shrinkage priors are very different from the shrinkage estimator due to James & 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 *et al.*, 2010). In reality, an ideal signal-recovery procedure should be robust to large signals.

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 $\phi(\theta)$. Probabilistically, $l(\theta)$ and $\text{pen}_\lambda(\boldsymbol{\theta})$ correspond to the negative logarithms of the likelihood and prior distribution, respectively.

Regularization leads to an optimization problem of the form

$$\min_{\boldsymbol{\theta}\in\Re^d} \{l(y \mid \boldsymbol{\theta}) + \text{pen}_\lambda(\boldsymbol{\theta})\} \ , \tag{3.1}$$

and the probabilistic approach leads to a Bayesian hierarchical model

$$p(y \mid \boldsymbol{\theta}) \propto \exp\{-l(y \mid \boldsymbol{\theta})\} \ , \quad \pi_\lambda(\boldsymbol{\theta}) \propto \exp\{-\text{pen}_\lambda(\boldsymbol{\theta})\}. \tag{3.2}$$

For appropriate $l(y \mid \boldsymbol{\theta})$ and $\text{pen}_\lambda(\boldsymbol{\theta})$, the solution to (3.1) corresponds to the posterior mode of (3.2), $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta} \mid y)$, where $p(\boldsymbol{\theta} \mid y)$ denotes the posterior distribution. 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 ℓ_2 penalty or Ridge (Tikhonov, 1963; Hoerl & Kennard, 1970) corresponds to a Gaussian prior under the same observation distribution, and an ℓ_1 penalty (Lasso) (Tibshirani, 1996) corresponds to a double-exponential prior (Park & Casella, 2008).

One interpretation of Lasso and related ℓ_1 penalties are methods designed to perform selection, while ridge and related ℓ_2 based methods perform shrinkage. Selection-based methods such as the Lasso are unstable in many situations, e.g., in presence of multi-collinearity in the design (Hastie *et al.*, 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 shrinks 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. However, the continuous nature of prior on θ_i ensures a lack of exact zeros $P(\theta_i = 0 \mid y_i) = 0$, which is preferred over dichotomous models by some practitioners (Stephens & 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 estimate is 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 & van de Geer (2011, Eq. (2.8), Th. (6.1)) states that with a proper choice of λ 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 ϕ_0^2. The compatibility (or restricted eigenvalue) constant reflects the compatibility between the design matrix and the ℓ_1 norm of $\boldsymbol{\theta}$, and is defined as follows (Bühlmann & 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 $\boldsymbol{\theta}_{-S} = \theta_{S^c}$. Then the compatibility condition is satisfied for the design $\mathbf{X}$ for the true support set $S = supp(\boldsymbol{\theta})$, if letting $s_0 = |S|$,*

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

The constant ϕ_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 suitably λ, and optimality results on variable selection.

3.2 Bayesian Lasso and Elastic Net

As discussed before, the posterior mean under Bayesian Lasso, the Bayes estimate under squared error loss, will not satisfy the optimality properties of the posterior mode under the double-exponential prior. Along these lines, Castillo *et al.* (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 *et al.* (2015) provide theoretical result that the full Lasso posterior does not contract at the same speed as the posterior mode.

However, 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 & Scott (2010b); Datta & Ghosh (2013), and motivates the key properties of global-local priors. Figure 1a provides profile plots as a diagnostic of shrinkage behavior for different priors.

For correlated predictors, Zou & 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.*, 2016b). The Lasso penalty

arises as a Laplace global-local mixture (Andrews & 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 θ_i, given hyper-parameters λ_1 and λ_2, has a density function with the following form:

$$p(\theta_i \mid \lambda_1, \lambda_2) = \begin{cases} \phi(\theta_i \mid \frac{\lambda_1}{2\lambda_2}, \frac{\sigma^2}{\lambda_2}) / 2\Phi\left(-\frac{\lambda_1}{2\sigma\lambda_2^{1/2}}\right), & \theta_i < 0, \\ \phi(\theta_i \mid \frac{-\lambda_1}{2\lambda_2}, \frac{\sigma^2}{\lambda_2}) / 2\Phi\left(-\frac{\lambda_1}{2\sigma\lambda_2^{1/2}}\right), & \theta_i \geq 0. \end{cases} \tag{3.3}$$

3.3 Horseshoe Penalty and Prior

The horseshoe prior is a continuous shrinkage rule for sparse signal recovery. Specifically, the horseshoe prior for θ_i, given a global shrinkage parameter τ, is given by the hierarchical model

$$\begin{aligned} (y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2),\ (\theta_i \mid \lambda_i, \tau) \sim \mathcal{N}(0, \lambda_i^2\tau^2) \\ \lambda_i^2 \sim \mathrm{C}^+(0,1), \quad i = 1, \ldots, n. \end{aligned} \tag{3.4}$$

As previously noted, the horseshoe prior operates under a different philosophy: that of modeling the inclusion probability directly rather than using a discrete mixture to model sparsity. To see this, note that the posterior mean under the horseshoe prior can be written as a linear function of the observation:

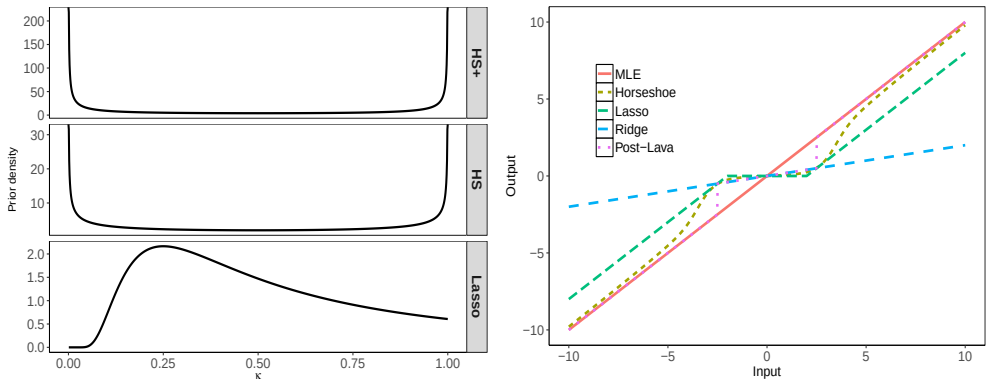
$$\mathbb{E}(\theta_i \mid y_i) = \{1 - \mathbb{E}(\kappa_i \mid y_i)\} y_i \text{ where } \kappa_i = 1/(1+\lambda_i^2\tau^2)$$

The name 'horseshoe' arises from the shape of the beta prior density of the shrinkage weights, κ_i. A comparison with the posterior mean obtained under the two-groups model reveals that the shrinkage weights perform the same job 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.'

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 marginal likelihood after reparametrizing $\kappa_i = (1+\lambda_i^2\tau^2)^{-1}$ is, $p(y_i \mid \kappa_i, \tau) = \kappa_i^{1/2} \exp\left(-\kappa_i y_i^2/2\right)$. The posterior density of κ_i identifies signals and noises by letting $\kappa_i \to 0$ and $\kappa_i \to 1$ respectively. Since the marginal likelihood puts no probability density on $\kappa_i = 0$, 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 does exactly that: it cancels the $\kappa_i^{1/2}$ term and replaces it with a $(1-\kappa_i)^{-1/2}$ to enable $\kappa_i \to 1$ in the posterior. The horseshoe+ prior (Bhadra *et al.*, 2017a) takes this philosophy one step further, by creating a U-shaped Jacobian for transformation from λ_i to κ_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.

Figure 1a plots the posterior density $p(\kappa_i \mid y_i)$ for the horseshoe, horseshoe+, and the Laplace priors. Figure 1b 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 *et al.*, 2017) (*vide* Appendix C) which works well on dense+sparse signals, and has the robustness property lacking in Lasso/Laplace prior.



(a) Shrinkage factor κ_i,$(p(\kappa_i \mid y_i))$ (b) Score function

Fig 1: Posterior density of shrinkage weight for the horseshoe, horseshoe+, and Laplace prior, where $(1-\kappa_i)$ can be interpreted as the pseudo posterior inclusion probability or $P(\theta_i \neq 0 \mid y_i)$, 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 τ is fixed at 0.1.

Carvalho *et al.* (2010) provided strong numerical evidence that this one-group shrinkage rule approximately behaves like the answers from a two-groups model under sparsity and attains super-efficiency in reconstructing the true density. Although, the main goal of a shrinkage prior is estimation, this interpretation of shrinkage weights as inclusion probabilities led Carvalho *et al.* (2010) to propose a multiple testing rule by using a threshold on $1-\hat{\kappa}_i$ values. Datta & Ghosh (2013) investigated the theoretical optimality of such a decision rule under a 0-1 additive loss and showed that the horseshoe multiple testing rule attains the Bayes oracle up to a multiplicative constant.

There are a number of closed-form results for the posterior distribution under a horseshoe prior. Although the prior density under the horseshoe prior doesn't 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:

$$\mathbb{E}(\theta_i \mid y_i, \tau) = y_i \left(1 - \frac{2\Phi_1(\frac{1}{2}, 1, \frac{5}{2}, \frac{y_i^2}{2\sigma^2}, 1-\frac{1}{\tau^2})}{3\Phi_1(\frac{1}{2}, 1, \frac{3}{2}, \frac{y_i^2}{2\sigma^2}, 1-\frac{1}{\tau^2})}\right), \tag{3.5}$$

where Φ_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'

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 ($DL_{\frac{1}{n}}$)	$\lvert\theta\rvert^{a-1}$ ($\lvert\theta\rvert^{\frac{1}{n}-1}$)	$\exp(-b\lvert\theta\rvert)$

TABLE 3
Different Priors: Behavior near origin and tails

approach with estimated values of the hyper-parameter τ. In a series of papers, van der Pas *et al.* (2014, 2016a,c, 2017) showed that the empirical Bayes posterior mean estimator enjoys a 'near-minimax' rate of estimation if the global shrinkage parameter τ is chosen suitably. We discuss the statistical properties of horseshoe posterior mean estimator and the induced decision rule in more details in § 4.

The horseshoe prior is a member of a wider class of global-local scale mixtures of normals that admit following hierarchical form (Polson & Scott, 2010b):

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

These priors are collectively called global-local shrinkage priors in Polson & Scott (2010b), since they recover signals by a local shrinkage parameter and adapt to sparsity by a global shrinkage parameter. Some of the popular shrinkage priors include the generalized double Pareto (GDP) (Armagan *et al.*, 2013a), the three-parameter beta (Armagan *et al.*, 2011), and the more recent horseshoe+ (Bhadra *et al.*, 2017a) and the Dirichlet–Laplace (Bhattacharya *et al.*, 2015) priors. A natural question is *how do we compare these priors?* It is known due to several authors (Polson & Scott, 2010b; Bhadra *et al.*, 2016a; van der Pas *et al.*, 2016a, e.g.) 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 global-local shrinkage priors along with their behavior near origin and the tails on Table 2. A detailed list of shrinkage priors proposed in the recent past is deferred to § 8.

From a regularization view-point, one way to judge a prior is by the penalty it imposes on a likelihood (3.1), although in a strict Bayesian spirit, a prior should be evaluated based on the whole posterior as shown by several authors including (Castillo *et al.*, 2015; van der Pas *et al.*, 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 *et al.* (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.* (2017b) 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 θ, yet with a spike at origin and heavy tails and

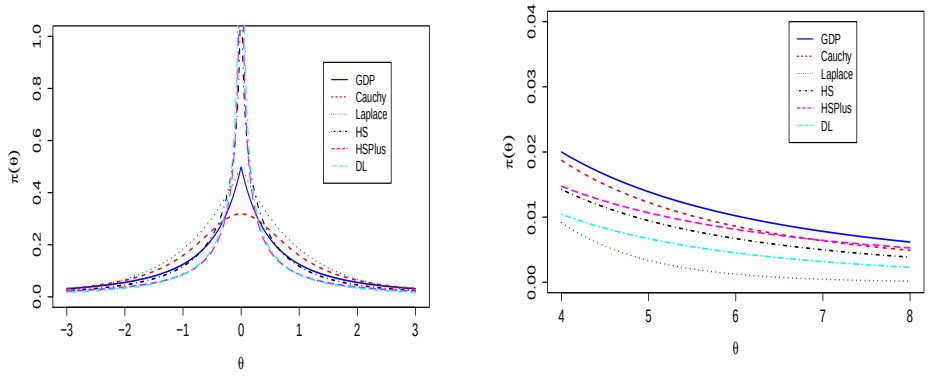


Fig 2: Marginal prior densities near the origin (left) and in the tail regions (right). The legends denote the horseshoe+ (HSPlus), horseshoe (HS), Dirichlet-Laplace (DL), generalized double Pareto (GDP), Cauchy and Laplace priors.

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 & Swirles, 1972, pages 406-407), and to save substantial additional space we refer the readers to the proof in §5, Lemma 5.1 and Proposition 5.1 of Bhadra *et al.* (2017b).

Horseshoe-like priors Bhadra *et al.* (2017b) have the following marginal prior density for θ_i:

$$\tilde{p}_{\tilde{HS}}(\theta_i \mid \tau^2) = \frac{1}{2\pi\tau} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right), \quad \theta_i \in \mathbb{R},\ \tau > 0. \tag{3.6}$$

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

$$p_{hs}(\theta_i \mid \tau^2) \propto \begin{cases} \frac{1}{\theta_i^{1-\epsilon}} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right) & \text{if } |\theta_i| < 1 \\ \theta_i^{1-\epsilon} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right) & \text{if } |\theta_i| \geq 1, \end{cases} \quad \epsilon \geq 0, \tau > 0. \tag{3.7}$$

Normal scale mixture The horseshoe-like prior (3.6) 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.* (2017b))**.** *The horseshoe-like prior in* (3.6) *has the following global-local scale mixture representation:*

$$
\begin{aligned}
(\theta_i \mid t_i, \tau) &\sim \mathcal{N}\left(0, \frac{\tau^2}{t_i^2}\right), \quad (t_i \mid s_i) \sim \mathcal{N}(0, s_i), \\
s_i &\sim \text{Pareto}\left(\frac{1}{2}\right), \quad t_i \in \mathbb{R},\ \tau \geq 0.
\end{aligned} \tag{3.8}
$$

TABLE 4
Priors for λ_i and κ_i for a few popular shrinkage rules

Prior for θ_i	Prior for λ_i	Prior for κ_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	$\lambda_i \exp(-\lambda_i^2/2)$	$\kappa_i^{-2}\exp-\frac{1}{2\kappa_i}$

4. STATISTICAL RISK PROPERTIES

4.1 Inadmissibility of MLE

The story of shrinkage estimation goes back to the proof in Stein (1956) that the maximum likelihood estimators for normal data are inadmissible beyond $\mathbb{R}^2$. The James-Stein (JS) estimator is $\hat{\theta}^{JS} = \{1-(n-2)/\,||\mathbf{y}||^2\}\mathbf{y}$ which is equivalent to the posterior mean $\hat{\theta}_{\text{Bayes}} = \tau^2/(\tau^2+1)\mathbf{y}$, under i.i.d. $\mathcal{N}(0,\tau^2)$ priors on θ_i. Thus, the James–Stein estimator corresponds to the Bayes risk of $n\tau^2/(\tau^2+1)$. We argue below that a global shrinkage rule such as the James–Stein estimator or ℓ_2 regularization does not work in the sparse regime as it lacks local parameters for handling sparsity.

James & Stein (1961) proved that this estimator dominates the MLE in terms of the expected total squared error for every choice of $\boldsymbol{\theta}$, i.e. it outperforms the MLE no matter what the true $\boldsymbol{\theta}$ is. To motivate the need for developing a local shrinkage rule, consider the classic James–Stein (JS) 'global' shrinkage rule, $\hat{\boldsymbol{\theta}}_{JS}(\mathbf{y})$. 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:

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

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

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

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

$$\tilde{c}(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||^2 / (n + ||\boldsymbol{\theta}||^2). \tag{4.1}$$

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

Lemma 4. *Consider the 'ideal' estimator* $\tilde{\boldsymbol{\theta}}_{IS}(\mathbf{y}) = \tilde{c}(\boldsymbol{\theta})(\mathbf{y})$ *in* (4.1). *For all* $p \geq 2$ *and for all* $\boldsymbol{\theta} \in \mathbb{R}^p$,

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

Here, $\hat{\boldsymbol{\theta}}_{JS}(\mathbf{y})$ for the r-spike parameter value has risk at least $R\left(\hat{\theta}^{JS}, \theta_r\right) \geq (n/2)$. This is nowhere near optimal. As Donoho & 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 $\hat{\boldsymbol{\theta}}^S(\mathbf{y}, \lambda) = sgn(\mathbf{y})(|\mathbf{y}| - \lambda)_+$ 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.

4.2 Near minimax risk

The asymptotically minimax risk rate in ℓ_2 for nearly black objects is given by Donoho *et al.* (1992) to be $p_n \log(n/p_n)$. Here $a_n \asymp b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$. Specifically, for any estimator $\delta(\mathbf{y})$, we have a lower bound ($\sigma^2 = 1$):

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \|\delta(Y) - \theta_0\|^2 \geq 2 p_n \log(n/p_n)(1 + o(1)). \tag{4.2}$$

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 *et al.* (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 *et al.*, 2014, 2016a,c, 2017). Specifically, for $\sigma^2 = 1$, the horseshoe estimator achieves

$$\sup_{\boldsymbol{\theta} \in \ell_0[p_n]} \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}} \|\hat{\theta}_{HS}(\mathbf{y}) - \boldsymbol{\theta}\|^2 \asymp p_n \log(n/p_n), \tag{4.3}$$

van der Pas *et al.* (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, τ is unknown and must either be estimated from the data or handled via a fully

Bayesian approach by putting a suitable prior on τ. van der Pas *et al.* (2017) show that the theoretical optimality properties for the popular horseshoe prior holds true if the global shrinkage parameter τ is learned via the maximum marginal likelihood estimator (MMLE) or a full Bayes approach. Independently, van der Pas *et al.* (2016a) and Ghosh & 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 *et al.* (2016a) apply to a wider class of priors, including the horseshoe+ prior (Bhadra *et al.*, 2017a) and spike-and-slab Lasso (Ročková & George, 2016), it is worth pointing out the difference between van der Pas *et al.* (2016a) and Ghosh & Chakrabarti (2017). van der Pas *et al.* (2016a) 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 & Chakrabarti (2017) attain 'exact' minimaxity for 'horseshoe-type' priors under suitable conditions on the global parameter τ, 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 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 & 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 τ.

The posterior distribution under the horseshoe prior leads to a natural model selection strategy under the two-groups model. Carvalho *et al.* (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:

$$\text{Reject the } i^{th} \text{ null hypothesis } H_{0i} : \theta_i = 0 \text{ if } 1-\hat{\kappa}_i > \frac{1}{2}\,. \tag{4.4}$$

Under the two-groups model (A.2), 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 values of the sparsity and the parameters of the non-null distribution, we can derive a decision rule that is impossible to beat in practice, this 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

$$p_n \to 0,\ u_n = \psi_n^2 \to \infty,\ \text{and}\ \log(v_n)/u_n \to C \in (0,\infty) \tag{4.5}$$

where $v_n = \psi_n^2(\frac{1-p_n}{p_n})^2$. The Bayes risk for the Bayes oracle under the above framework (4.5) is given by:

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

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, e.g. Benjamini–Hochberg FDR controlling rule to be ABOS. Datta & Ghosh (2013) first showed that the horseshoe decision rule (4.4) is also ABOS up to a multiplicative constant if τ is chosen suitably to reflect the sparsity, namely $\tau = O(\pi)$. The proof in Datta & 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 & 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 τ is treated as a tuning parameter that mimics π in the theoretical treatment, in practice, π is an unknown parameter. Several authors Datta & Ghosh (2013); Ghosh & Chakrabarti (2017); Ghosh *et al.* (2016); van der Pas *et al.* (2016c) have shown that usual estimates of τ 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. They have also shown that the testing rule under a horseshoe-type prior is *exactly* ABOS, when $\lim_{n\to\infty} \tau/p \in (0,\infty)$.

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 & van de Geer, 2011, e.g.), whereas similar results for Bayesian methods using shrinkage prior are relatively sparser. 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 Castillo *et al.* (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, e.g. 'irrepresentability' and 'mutual coherence' conditions (Bühlmann & van de Geer, 2011, *vide* Ch. 6) and (Zhao & Yu, 2006). Similarly, for recovery under point-mass mixture priors, Castillo *et al.* (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 ℓ_1 norm, ℓ_2 norm and ℓ_∞ 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:

$$\left|\left|\hat{\Sigma}_{p-s_0,s_0}\hat{\Sigma}^{-1}_{s_0,s_0} sign(\boldsymbol{\theta}_S)\right|\right|_\infty \leq 1-\eta \text{ for positive constant vector } \eta. \tag{4.6}$$

Zhao & 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:

$$\eta_\infty = 1 - \left|\left|\hat{\Sigma}_{p-s_0,s_0}\hat{\Sigma}^{-1}_{s_0,s_0} sign(\boldsymbol{\theta}_S)\right|\right|_\infty \tag{4.7}$$

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

$$\text{mc}(\mathbf{X}) = \max_{1\leq i\neq j\leq p} \frac{|\langle X_{.,i} X_{.,j}\rangle|}{||X_{.,i}||_2\, ||X_{.,j}||_2}. \tag{4.8}$$

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

Continuous shrinkage priors: Polson & Scott (2010b) 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 & Scott, 2010b, 2012a; Datta & 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.* (2013b) proved posterior consistency in $p \leq n$ situation for commonly used shrinkage prior including generalized double Pareto and horseshoe-type priors under simple sufficient conditions, e.g. boundedness of the eigenvalues of $\mathbf{X}^T\mathbf{X}/n$ and $\pi_n = o(n/\log n)$. 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 ultra-high dimensional setting is still an active area.

There is some recent developments on theoretical properties for predictive risk and variable selection properties of the horseshoe posterior under a orthogonal design matrix in $p \leq n$ 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 $p \leq n$ under the model:

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

Tang *et al.* (2016) proposed the half-thresholding estimator,

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

where $\hat{\theta}_i^{PM}$ and $\hat{\theta}_i^{OLS}$ 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.* (2016c) 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 $\mathbf{X}$):

$$\begin{aligned} \mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon &\overset{\text{reparametrize}}{\Rightarrow} \mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \epsilon \\ \text{where } \mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{W}^T,\ \mathbf{Z} = \mathbf{U}\boldsymbol{\alpha},\ \boldsymbol{\alpha} &= \mathbf{W}^T\boldsymbol{\theta}.(\text{Rank}(\mathbf{D}) = n). \end{aligned} \tag{4.9}$$

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

Theorem 5. *Prediction risk for the purely local horseshoe regression (Bhadra et al., 2016c). Let* $\mathbf{D} = \mathbf{I}$ *in* (4.9) *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 ℓ_2 or ℓ_1 norm carry over to horseshoe prior in the ultra high-dimensional set-up under compatibility or coherence conditions on the design matrix as used by Bühlmann & van de Geer (2011); Castillo *et al.* (2015).

4.5 Uncertainty quantification

Reliable uncertainty quantification is a key challenge in high-dimensional inference. While some authors (Chatterjee & Lahiri, 2011, e.g.) observed that the Lasso-based estimates do not yield meaningful standard errors for the parameter estimates, Castillo *et al.* (2015) showed poor posterior contraction for Bayesian Lasso. These results motivate Bayesian approach with appropriately heavy-tailed priors that produces automatic and reliable uncertainty quantification. Chatterjee & Lahiri (2011) also proposed a Bootstrap-based estimator for the limiting distribution under Lasso that attains consistency. Several authors including Liu & Yu (2013); Zhang & Zhang (2014); Van de Geer *et al.* (2014); Javanmard &

Montanari (2014) have proposed constructing intervals based on de-biasing or de-sparsifying the standard 'Lasso' estimator to achieve an asymptotic Gaussian limiting distribution for single coordinates θ_i or low-dimensional parameters of interest. The general form of the 'de-sparsified' estimators given in Javanmard & Montanari (2014) is:

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

As Javanmard & Montanari (2014) explain, $\mathbf{M} \in \mathbb{R}^{p\times p}$ plays a crucial role in 'decorrelating' the columns of $\mathbf{X}$. The algorithmic construction of $\mathbf{M}$ optimizes two quantities: the entry-wise ℓ_∞ norm $|\mathbf{M}\hat{\Sigma} - \mathbf{I}|_\infty$ which protects against non-normality and bias and $[M\hat{\Sigma}M]_{i,i}$ that controls the variance of the de-biased estimator.

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 ℓ_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 θ_0, be it Bayesian or non-Bayesian. In particular, sparsity-adaptive credible sets can not be 'honest' (Nickl *et al.*, 2013; Li, 1989) in the sense that it is impossible to construct credible sets that have both their diameter adapting to the minimax rate for the unknown sparsity π as well as provide nominal coverage probability over the full parameter space.

In the context of sequence models, as van der Pas *et al.* (2017) point out that since the horseshoe prior achieves adaptive posterior contraction at the near-minimax rate $p_n \log(n/p_n)$ in (4.3) for nearly-black objects, one needs additional condition, e.g. excessive bias-restriction (Belitser & 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 π cross the detectability threshold, $\sqrt{2\log(n/p_n)}$. We refer the readers to Theorem 5 of van der Pas *et al.* (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 like horseshoe attains 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 τ is critical for success of the horseshoe estimator in a sparse regime as it captures the level of sparsity in the data (Carvalho *et al.*, 2010; Datta & Ghosh, 2013; van der Pas *et al.*, 2016a). However, in nearly black situation a naive estimate of τ could collapse to zero, and care must be taken to prevent possible degeneracy in inference. There are two main approaches regarding choice of τ: first, an empirical Bayesian approach that estimates τ 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 τ.

5.1 Marginal Likelihood

We first take a closer look at how τ affects the marginal likelihood under the horseshoe prior and the maximum marginal likelihood approach of van der Pas *et al.* (2017). We can write the marginal likelihood under the horseshoe prior after marginalizing out θ_i in (3.4) for $\sigma^2 = 1$ from the model as:

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

Tiao & Tan (1966) observe that the marginal likelihood is positive at $\tau = 0$, hence the impropriety of the prior of τ^{-2} at the origin translates to the posterior. As a result, a maximum likelihood estimator of τ has a potential danger of collapsing to zero in very sparse problems (Polson & Scott, 2010b; Datta & Ghosh, 2013). In van der Pas *et al.* (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 τ using the approach of van der Pas *et al.* (2017), we first calculate the marginal prior of θ_i after integrating out λ_i^2 in Equation (3.4):

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

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

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

The lower bound of the maximization interval prevents against a degenerate solution of τ in sparse case.

Handling τ is still an area of research: some papers (Carvalho *et al.*, 2010; Datta & Ghosh, 2013; Piironen & Vehtari, 2017, e.g.) advocate using a full Bayes approach instead of a 'plug-in' maximum likelihood approach to avoid potential issues such as $\hat{\tau}$ collapsing to 0. On the other hand, van der Pas *et al.* (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 *et al.*, 2017, *vide* Rejoinder p. 1274).

In practice, the MMLE approach of van der Pas *et al.* (2017) achieves both theoretical optimality as well as good numerical performance and it is computed over the interval $[1/n, 1]$, which connects to the interpretation of τ as sparsity as well as prevent any computational issues.

5.2 Optimization and Cross-validation

In a recent paper, van der Pas *et al.* (2017) have investigated the empirical Bayes and full Bayes approach for τ, and have shown that the full Bayes and the MMLE estimator 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 *et al.* (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 non-zero signals, even those below the threshold $\sqrt{2\log(n)}$, such as $\theta_i = 1$ when $n = 100$.

A third approach could be treating τ as a tuning parameter and using a k-fold cross-validation to select τ. 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 τ was proposed by Piironen & Vehtari (2016), who have investigated the choice of τ for a linear regression model and have suggested choosing a prior for τ 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.* (2016c) suggest selecting τ by minimizing SURE, for which they provide an explicit form under the model in (4.9).

6. COMPUTATION

Over the last few years, several different implementations of the horseshoe prior for normal means and regression model have been proposed. The MCMC based implementations usually proceed via block-updating $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ and τ using either a Gibbs or 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 *et al.* (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 τ. Furthermore, Scott (2010) pointed out that the solution to this lies in marginalizing over the local shrinkage parameter λ_j's. On a somewhat similar route, Makalic & Schmidt (2016) uses a inverse-gamma scale mixture identity to construct a Gibbs sampling scheme for horseshoe and horseshoe+ prior for linear regression as well as logistic and negative binomial regression.

The `horseshoe` package implements the MMLE and truncated prior approaches for handling τ proposed in van der Pas *et al.* (2017). Hahn *et al.* (2016) proposed an elliptical slice sampler and argues that it wins over 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 horseshoe prior in linear regression is Bhattacharya *et al.* (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 & Orenstein (2017) claims to improve this even further by implementing a block update strategy but using a random walk Metropolis–Hastings algorithm on $\log(1/\tau^2)$ for block-updating $\tau \mid \lambda$. We provide a list of all the implementations known to us on Table 5.

Bayesian methods using an MCMC is sequential in nature and the extra time comes with better uncertainty quantification. However, sparse Bayesian methods including 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 *et al.* (2016) 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.*, 2017b).

TABLE 5
Implementations of Horseshoe and Other Shrinkage Priors

Implementation (Package/URL)	Authors
R package: `monomvn`	Gramacy *et al.* (2010)
R code in paper	Scott (2010)
R package: `horseshoe`	van der Pas *et al.* (2016b)
R package: `fastHorseshoe`	Hahn *et al.* (2016)
MATLAB code	Bhattacharya *et al.* (2016)
GPU accelerated Gibbs sampling	Terenin *et al.* (2016)
`bayesreg` + MATLAB code in paper	Makalic & Schmidt (2016)
MATLAB code	Johndrow & Orenstein (2017)
R package:`bayeslm`	Hahn *et al.* (2017)

7. SIMULATION EXPERIMENTS

7.1 Effect of Correlated Predictors

As we discussed in §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 *et al.* (2017) point out, such conditions are expected to be necessary for optimal recovery as in the context of spike-and-slab prior (Castillo *et al.*, 2015).

For this simulation study, we follow the set-up in Zhao & 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 σ^2 was set to 0.1 to obey the asymptotic properties of the Lasso.

We first draw the covariance matrix Σ from Wishart(p, I_p) and then generate design matrix $\mathbf{X}$ from $\mathcal{N}(0, \Sigma)$. Zhao & Yu (2006) showed that the Strong Irrepresentability Condition (4.6) may not hold for such a design matrix. We generate 100 such design matrices to obtain a range of different η_∞ values. In our simulation studies the η_∞ values in (4.7) 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 & 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 & 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 (4.8) number for the

same design matrices to see the effect on these two methods. The mc($\mathbf{X}$) numbers were between $[0.21, 0.54]$.

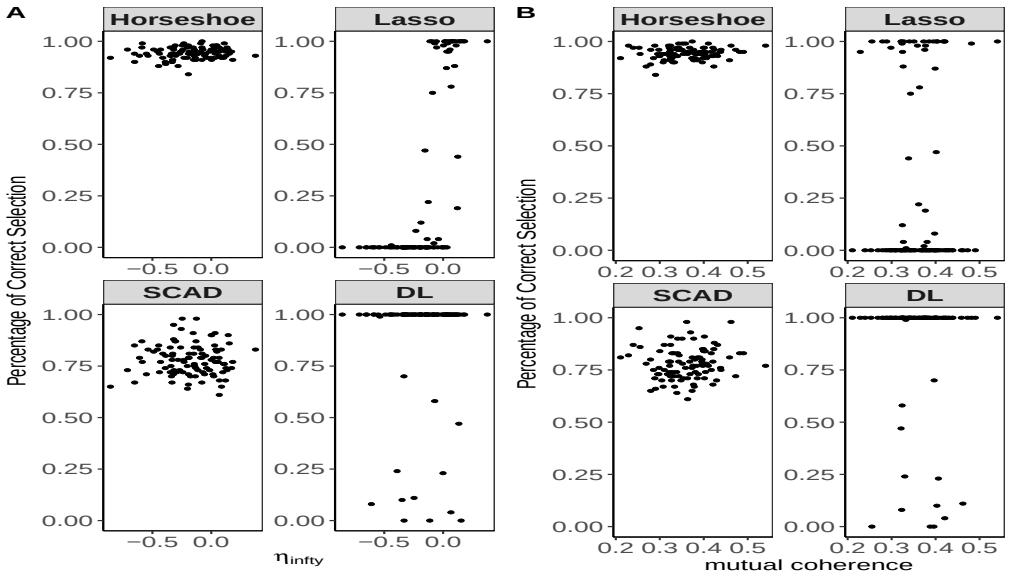


Fig 3: Effect of Strong Irrepresentability Condition η_∞ and Mutual Coherence (maximum column correlation) on the percentage of selecting the correct model by Lasso and SCAD penalty as well as horseshoe and Dirichlet–Laplace (DL) prior.

Figure 3 shows the percentage of correctly selected model as a function of the irrepresentable condition number, η_∞ 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 increasing η_∞. Surprisingly, the effect is weaker for SCAD as well as both the horseshoe and the Dirichlet–Laplace prior.

While horseshoe almost always recovers the true sparse $\boldsymbol{\theta}$ vector irrespective of η_∞, 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.

7.2 Binary Response: Logistic Regression

We compare the performance of horseshoe prior and Lasso for logistic regression for varying degree of dependence between the columns of a design matrix. We generated $n = 100$ binary observation 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 σ^2 was set to 0.1. We set the covariance matrix same as the last example, i.e. $\Sigma_{ij} = \rho^{|i-j|}$ 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 & Vehtari (2017), where they use a second-order Taylor expansion for the log posterior distribution. Piironen & Vehtari (2017) also propose the regularized horseshoe prior where one introduces an additional slab width

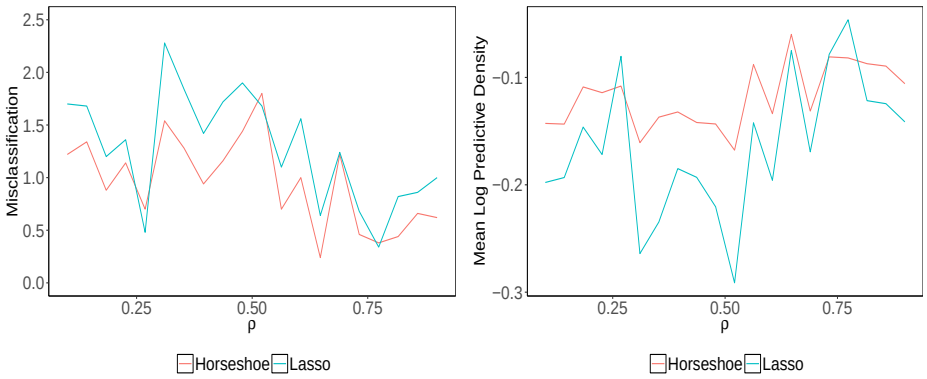


Fig 4: (a) Number of misclassified test data points and (b) mean log predictive density in (7.1) by the horseshoe and Lasso across different values of correlation ρ: a higher value of ρ represents higher dependence between the columns of $\mathbf{X}$.

c to allow for shrinkage even on the extreme tails. Following the recommendations of Piironen & Vehtari (2017), we use the regularized horseshoe prior with a hyper-prior $c \sim \text{Inv-Gamma}(2, 8)$ that corresponds to a Student-t$(0, 2^2)$ slab. We use $1,000$ 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 *et al.* (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:

$$\text{MLPD} = \frac{1}{m} \sum_{j=1}^{m} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(\mathbf{y}_j \mid \boldsymbol{\theta}^s) \right) \tag{7.1}$$

Figure ?? shows the average number of misclassified observations by horseshoe is a little lower than Lasso for all but two values of ρ. For the same values of ρ, Fig. ?? shows that the predictive accuracy under the horseshoe prior is a little better than the Lasso. We direct the readers to Piironen & Vehtari (2017) for a thorough comparison between the different variants of 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 of 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, 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 ℓ_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 of Horseshoe

As discussed in Section 3.3, the horseshoe prior belongs to a wider class of global-local shrinkage priors (Polson & Scott, 2010b) 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. Several authors including van der Pas *et al.* (2016a); Ghosh & Chakrabarti (2017); Ghosh *et al.* (2016) have provided conditions for optimality of one-group continuous priors for estimation of sparse normal means and multiple testing.

Although the original horseshoe prior was developed for signal recovery with sparse Gaussian means, the idea of directly modeling the posterior inclusion probability and use of normal-scale mixture to facilitate sparsity is a flexible idea and can be easily generalized to a wider class of problems. Bhadra *et al.* (2016a) 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.* (2016c) 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 & 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 & Vehtari (2016) used a Gaussian approximation using a second-order Taylor expansion for the log-likelihood to apply the horseshoe prior for the generalized linear model. Wang & Pillai (2013) proposed a shrinkage prior based on a scale mixture of uniform for covariance matrix estimation. Peltola *et al.* (2014) applies the horseshoe prior for Bayesian linear survival regression for selecting covariates with highest predictive values. A sample of the many applications of horseshoe prior is given in Table 6. Given the explosive growth of the methods in this area, we conjecture that the horseshoe prior would be regarded as a key tool sparse signal recovery and as a default prior for objective Bayesian inference for 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 reason for better performance by the horseshoe prior

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 & Marks (2016)
Linear regression for Causal inference and Instrumental variable models	Hahn & Lopes (2014); Hahn *et al.* (2016)
Multiclass prediction using DOLDA (Diagonally orthant Latent Dirichlet Allocation)	Magnusson *et al.* (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 & Minin (2015)
Quasi-Sparse Count Data	Datta & Dunson (2016)
Variable Selection under the projection predictive framework	Piironen & Vehtari (2015)
Dynamic shrinkage Process (dynamic linear model and trend filtering)	Kowal *et al.* (2017)
Logistic regression with horseshoe prior	Piironen & Vehtari (2017); Wei (2017)
Tree ensembles with rule structured horseshoe regularization	Nalenz & Villani (2017)
Bayesian compression for deep learning	Louizos *et al.* (2017)
Precision matrix estimation	Li *et al.* (2017)

is its heavy tails and probability spike at zero, which makes 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 §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 §4 of Bhadra *et al.* (2017b) is orders of magnitude than the non-convex SCAD or MCP (Bhadra *et al.*, 2017b, *vide* Table 1). Another fruitful strategy is to employ proximal algorithms similar to expectation-maximization (Polson *et al.*, 2015). These algorithms can be specifically designed to outperform staples such as Lasso (Bhadra *et al.*, 2017b) 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, e.g. using GPU-accelerated computing (Terenin *et al.*, 2016).

We have discussed the theoretical optimality properties for both Lasso and horseshoe estimator. The optimality properties of Lasso in regression are well-known and they depend on 'neighborhood stability' or 'irrepresentability' condition (4.6) and 'beta-min' condition. Similarly, adaptive posterior concentration for horseshoe depends on 'excessive bias restriction', a condition analogous

to 'beta-min'. 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 *et al.*, 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.* (2016a), but a thorough study of horseshoe prior for default Bayes problems is still an unexplored area.

Global-local shrinkage is still a fruitful area for future research. We list a few possible directions here.

(i) The square-root Lasso (Belloni *et al.*, 2011) or scaled Lasso (Sun & Zhang, 2012) improves over the Lasso by making the inference ambivalent towards σ, while making the estimator scale-invariant. It might be interesting to study the effect of marginalizing the global parameters such as τ and σ on inference from shrinkage priors. Our preliminary investigation suggests that scaling the prior on τ by σ or marginalizing out σ 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 & 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 *et al.* (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 θ_i is non-zero with some common prior probability $(1 - \pi)$, and that the nonzero θ_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 π 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 θ_i's are modeled as

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

where δ_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:

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

From (A.2), we see that the two-groups model leads to a sparse estimate, i.e., it puts exact zeros in the model.

APPENDIX B: PROOF OF EQUATION (3.6)

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 θ_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:

$$E(\theta_i \mid y_i) = \int_0^1 (1-\kappa_i) y_i p(\kappa_i \mid y_i) d\kappa_i = \{1 - E(\kappa_i \mid y_i)\} y_i$$

APPENDIX C: SHRINKAGE PROFILES

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

$$d_{\text{lasso}}(z) = \underset{\theta\in\mathbb{R}}{\operatorname{argmin}}\{(z-\theta)^2 + \lambda_l|\theta|\} = (|z| - \lambda_l/2)_+ sgn(z) \tag{C.1}$$

$$d_{\text{ridge}}(z) = \underset{\theta\in\mathbb{R}}{\operatorname{argmin}}\{(z-\theta)^2 + \lambda_r\theta^2\} = (1+\lambda_r)^{-1} z \tag{C.2}$$

$$d_{\text{post-lava}}(z) = \begin{cases} z & |z| > \lambda_1/2k \\ (1-k)z & |z| \le \lambda_1/2k, \end{cases} \text{ where } k = \lambda_2/(1+\lambda_2) \tag{C.3}$$

$$d_{\text{horseshoe}}(z) = z\left(1 - \frac{2\Phi_1(1/2, 1, 5/2, z^2/2, 1-1/\tau^2)}{3\Phi_1(1/2, 1, 3/2, z^2/2, 1-1/\tau^2)}\right) \tag{C.4}$$

Figure 1b 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., & Mallows, C. L. 1974. Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B*, **36**, 99–102.

Armagan, Artin, Clyde, Merlise, & Dunson, David B. 2011. Generalized Beta Mixtures of Gaussians. *Pages 523–531 of: Advances in Neural Information Processing Systems*.

Armagan, Artin, Dunson, David B, & Lee, Jaeyong. 2013a. Generalized Double Pareto Shrinkage. *Statistica Sinica*, **23**(1), 119–143.

Armagan, Artin, Dunson, David B, Lee, Jaeyong, Bajwa, Waheed U, & Strawn, Nate. 2013b. Posterior consistency in linear models under shrinkage priors. *Biometrika*, **100**(4), 1011–1018.

Bai, Ray, & Ghosh, Malay. 2017. The Inverse Gamma-Gamma Prior for Optimal Posterior Contraction and Multiple Hypothesis Testing. *arXiv preprint arXiv:1710.04369*.

Belitser, Eduard, & Nurushev, Nurzhan. 2015. Needles and straw in a haystack: robust confidence for possibly sparse sequences. *arXiv preprint arXiv:1511.01803*.

Belloni, Alexandre, Chernozhukov, Victor, & Wang, Lie. 2011. Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika*, **98**(4), 791–806.

Berzuini, Carlo, Guo, Hui, Burgess, Stephen, & Bernardinelli, Luisa. 2016. Mendelian Randomization with Poor Instruments: a Bayesian Approach. *arXiv:1608.02990 [math, stat]*, Aug. arXiv: 1608.02990.

Bhadra, Anindya, Datta, Jyotishka, Polson, Nicholas G, & Willard, Brandon. 2016a. Default Bayesian analysis with global-local shrinkage priors. *Biometrika*, **103**(4), 955–969.

Bhadra, Anindya, Datta, Jyotishka, Polson, Nicholas G, & Willard, Brandon. 2016b. Global-Local Mixtures. *arXiv preprint arXiv:1604.07487*.

Bhadra, Anindya, Datta, Jyotishka, Li, Yunfan, Polson, Nicholas G, & Willard, Brandon. 2016c. Prediction Risk for the Horseshoe Regression. *arXiv preprint arXiv:1605.04796*.

Bhadra, Anindya, Datta, Jyotishka, Polson, Nicholas G, & Willard, Brandon. 2017a. The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Analysis*, **12**(4), 1105–1131.

Bhadra, Anindya, Datta, Jyotishka, Polson, Nicholas G, & Willard, Brandon. 2017b. Horseshoe Regularization for Feature Subset Selection. *arXiv preprint arXiv:1702.07400*.

Bhattacharya, Anirban, Pati, Debdeep, Pillai, Natesh S., & Dunson, David B. 2015. Dirichlet-Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.

Bhattacharya, Anirban, Chakraborty, Antik, & Mallick, Bani K. 2016. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, asw042.

Bien, Jacob, Taylor, Jonathan, & Tibshirani, Robert. 2013. A Lasso for Hierarchical Interactions. *Annals of Statistics*, **41**(3), 1111–1141.

Bogdan, Małgorzata, Chakrabarti, Arijit, Frommlet, Florian, & Ghosh, Jayanta K. 2011. Asymptotic Bayes-Optimality under Sparsity of Some Multiple Testing Procedures. *The Annals of Statistics*, **39**(3), 1551–1579.

Bühlmann, Peter, & van de Geer, Sara. 2011. *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg.

Candes, Emmanuel, & Tao, Terence. 2007. The Dantzig Selector: Statistical Estimation When p Is Much Larger than N. *The Annals of Statistics*, 2313–2351.

Candes, Emmanuel J. 2008. The Restricted Isometry Property and Its Implications for Compressed Sensing. *Comptes Rendus Mathematique*, **346**(9-10), 589–592.

Candès, Emmanuel J, & Tao, Terence. 2010. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, **56**(5), 2053–2080.

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

Carvalho, Carlos M, Polson, Nicholas G, & Scott, James G. 2010. The Horseshoe Estimator for Sparse Signals. *Biometrika*, **97**, 465–480.

Castillo, Ismaël, & van der Vaart, Aad. 2012. Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences. *The Annals of Statistics*, **40**(4), 2069–2101.

Castillo, Ismail, Schmidt-Hieber, Johannes, & van der Vaart, Aad. 2015. Bayesian Linear Regression with Sparse Priors. *The Annals of Statistics*, **43**(5), 1986–2018.

Chatterjee, Arindam, & Lahiri, Soumendra Nath. 2011. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, **106**(494), 608–625.

Chernozhukov, Victor, Hansen, Christian, Liao, Yuan, *et al.* 2017. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, **45**(1), 39–76.

Cutillo, Luisa, Jung, Yoon Young, Ruggeri, Fabrizio, & Vidakovic, Brani. 2008. Larger posterior

mode wavelet thresholding and applications. *Journal of Statistical Planning and Inference*, **138**(12), 3758–3773.

Datta, Jyotishka, & Dunson, David B. 2016. Bayesian Inference on Quasi-Sparse Count Data. *Biometrika*, **103**(4), 971–983.

Datta, Jyotishka, & Ghosh, Jayanta K. 2013. Asymptotic Properties of Bayes Risk for the Horseshoe Prior. *Bayesian Analysis*, **8**(1), 111–132.

Datta, Jyotishka, & Ghosh, Jayanta K. 2015. In Search of Optimal Objective Priors for Model Selection and Estimation. *Current Trends in Bayesian Methodology with Applications*, 225.

Donoho, David L. 2006. Compressed Sensing. *IEEE Transactions on information theory*, **52**(4), 1289–1306.

Donoho, David L., & Johnstone, Iain M. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

Donoho, David L, & Johnstone, Iain M. 1995. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, **90**(432), 1200–1224.

Donoho, David L, Johnstone, Iain M, Hoch, Jeffrey C, & Stern, Alan S. 1992. Maximum Entropy and the Nearly Black Object. *Journal of the Royal Statistical Society. Series B*, **54**, 41–81.

Efron, Bradley. 2008. Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science*, **23**(1), 1–22.

Efron, Bradley. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Vol. 1. Cambridge University Press.

Efron, Bradley, Hastie, Trevor, Johnstone, Iain, & Tibshirani, Robert. 2004. Least Angle Regression. *The Annals of statistics*, **32**(2), 407–499.

Fan, Jianqing, & Li, Runze. 2001. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.

Faulkner, James R., & Minin, Vladimir N. 2015. Bayesian trend filtering: adaptive temporal smoothing with shrinkage priors. Dec.

Friedman, Jerome, Hastie, Trevor, Höfling, Holger, Tibshirani, Robert, & others. 2007. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, **1**(2), 302–332.

Friedman, Jerome, Hastie, Trevor, & Tibshirani, Robert. 2008. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**(3), 432–441.

Friedman, Jerome, Hastie, Trevor, & Tibshirani, Rob. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), 1–22.

Gelman, Andrew, Hwang, Jessica, & Vehtari, Aki. 2014. Understanding predictive information criteria for Bayesian models. *Statistics and computing*, **24**(6), 997–1016.

George, Edward I. 2000. The Variable Selection Problem. *Journal of the American Statistical Association*, **95**(452), 1304–1308.

George, EdwardI., & Foster, Dean P. 2000. Calibration and Empirical Bayes Variable Selection. *Biometrika*, **87**(4), 731–747.

Ghosal, Subhashis, Ghosh, Jayanta K., & van der Vaart, Aad W. 2000. Convergence Rates of Posterior Distributions. *The Annals of Statistics*, **28**(2), 500–531.

Ghosh, Prasenjit, & Chakrabarti, Arijit. 2017. Asymptotic Optimality of One-Group Shrinkage Priors in Sparse High-dimensional Problems. *Bayesian Anal.* Advance publication.

Ghosh, Prasenjit, Tang, Xueying, Ghosh, Malay, & Chakrabarti, Arijit. 2016. Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Anal.*, **11**(3), 753–796.

Gordy, Michael B. 1998. Computationally convenient distributional assumptions for common-value auctions. *Computational Economics*, **12**(1), 61–78.

Gramacy, Robert B, Pantaleo, Ester, & others. 2010. Shrinkage Regression for Multivariate Inference with Missing Data, and an Application to Portfolio Balancing. *Bayesian Analysis*, **5**(2), 237–262.

Griffin, Jim E, & Brown, Philip J. 2010. Inference with Normal-Gamma Prior Distributions in Regression Problems. *Bayesian Analysis*, **5**(1), 171–188.

Hahn, P. Richard, & Lopes, Hedibert. 2014. Shrinkage priors for linear instrumental variable models with many instruments. *arXiv preprint arXiv:1408.0462*, Aug.

Hahn, P. Richard, He, Jingyu, & Lopes, Hedibert. 2016. *Elliptical Slice Sampling for Bayesian Shrinkage Regression with Applications to Causal Inference*. Tech. rept.

Hahn, P Richard, He, Jingyu, & Lopes, Hedibert F. 2017. Efficient sampling for Gaussian linear regression with arbitrary priors.

Hans, Chris. 2011. Elastic Net Regression Modeling with the Orthant Normal Prior. *Journal of*

the American Statistical Association, **106**(496), 1383–1393.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, Hastie, T, Friedman, J, & Tibshirani, R. 2009. *The Elements of Statistical Learning*. Vol. 2. Springer.

Hastie, Trevor, Tibshirani, Robert, & Wainwright, Martin. 2015. *Statistical Learning with Sparsity*. CRC press.

Hoerl, Arthur E., & Kennard, Robert W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55–67.

Ingraham, John B., & Marks, Debora S. 2016. Bayesian Sparsity for Intractable Distributions. *arXiv preprint arXiv:1602.03807*.

Ishwaran, Hemant, & Rao, J. Sunil. 2005. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.*, **33**(2), 730–773.

James, Gareth, Witten, Daniela, Hastie, Trevor, & Tibshirani, Robert. 2013. *An Introduction to Statistical Learning*. Vol. 6. Springer.

James, William, & Stein, Charles. 1961. Estimation with Quadratic Loss. *Pages 361–379 of: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1.

Javanmard, Adel, & Montanari, Andrea. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, **15**(1), 2869–2909.

Jeffreys, Harold, & Swirles, Bertha. 1972. *Methods of Mathematical Physics*. 3 edn. Cambridge: Cambridge university press.

Johndrow, James E., & Orenstein, Paulo. 2017. Scalable MCMC for Bayes Shrinkage Priors. *arXiv preprint arXiv:1705.00841*.

Johnstone, Iain M, & Silverman, Bernard W. 2004. Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences. *Annals of Statistics*, **32**, 1594–1649.

Jolliffe, Ian T, Trendafilov, Nickolay T, & Uddin, Mudassir. 2003. A Modified Principal Component Technique Based on the LASSO. *Journal of computational and Graphical Statistics*, **12**(3), 531–547.

Kowal, Daniel R, Matteson, David S, & Ruppert, David. 2017. Dynamic Shrinkage Processes. *arXiv preprint arXiv:1707.00763*.

Li, Ker-Chau. 1989. Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 1001–1008.

Li, Yunfan, Craig, Bruce A, & Bhadra, Anindya. 2017. The Graphical Horseshoe Estimator for Inverse Covariance Matrices. *arXiv preprint arXiv:1707.06661*.

Liu, Hanzhong, & Yu, Bin. 2013. Asymptotic properties of Lasso+ mLS and Lasso+ Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, **7**, 3124–3169.

Louizos, Christos, Ullrich, Karen, & Welling, Max. 2017. Bayesian compression for deep learning. *Pages 3290–3300 of: Advances in Neural Information Processing Systems*.

Magnusson, Mns, Jonsson, Leif, & Villani, Mattias. 2016. DOLDA - a regularized supervised topic model for high-dimensional multi-class regression. *arXiv preprint arXiv:1602.00260*, Jan.

Makalic, Enes, & Schmidt, Daniel F. 2016. High-Dimensional Bayesian Regularised Regression with the BayesReg Package. *arXiv preprint arXiv:1611.06649*.

Mazumder, Rahul, Hastie, Trevor, & Tibshirani, Robert. 2010. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of machine learning research*, **11**(Aug), 2287–2322.

Mazumder, Rahul, Friedman, Jerome H, & Hastie, Trevor. 2012. SparseNet: Coordinate Descent with Nonconvex Penalties. *Journal of the American Statistical Association*, **106**, 1125–1138.

Mitchell, T. J., & Beauchamp, J. J. 1988. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, **83**(404), 1023–1032.

Nalenz, Malte, & Villani, Mattias. 2017. Tree Ensembles with Rule Structured Horseshoe Regularization. *arXiv preprint arXiv:1702.05008*.

Nickl, Richard, van de Geer, Sara, *et al.* 2013. Confidence sets in sparse regression. *The Annals of Statistics*, **41**(6), 2852–2876.

Park, Trevor, & Casella, George. 2008. The Bayesian Lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

Peltola, Tomi, Havulinna, Aki S, Salomaa, Veikko, & Vehtari, Aki. 2014. Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction. *Pages 79–88 of: Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop-Volume 1218*. CEUR-WS. org.

Piironen, Juho, & Vehtari, Aki. 2015. Projection predictive variable selection using Stan+R. *arXiv preprint arXiv:1508.02502*, Aug.

Piironen, Juho, & Vehtari, Aki. 2016. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. *arXiv preprint arXiv:1610.05559*.

Piironen, Juho, & Vehtari, Aki. 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, **11**(2), 5018–5051.

Polson, Nicholas G, & Scott, James G. 2010a. Large-Scale Simultaneous Testing with Hypergeometric Inverted-Beta Priors. *arXiv preprint arXiv:1010.5223*.

Polson, Nicholas G, & Scott, James G. 2010b. Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. *Bayesian Statistics*, **9**, 501–538.

Polson, Nicholas G, & Scott, James G. 2012a. Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(2), 287–311.

Polson, Nicholas G, & Scott, James G. 2012b. On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis*, **7**(4), 887–902.

Polson, Nicholas G, & Scott, James G. 2016. Mixtures, Envelopes and Hierarchical Duality. *Journal of the Royal Statistical Society. Series B*, **78**, 701–727.

Polson, Nicholas G, Scott, James G, & Willard, Brandon T. 2015. Proximal Algorithms in Statistics and Machine Learning. *Statistical Science*, **30**(4), 559–581.

Ročková, Veronika, & George, Edward I. 2016. The Spike-and-Slab Lasso. *Journal of the American Statistical Association*.

Scott, James G. 2010. Parameter Expansion in Local-Shrinkage Models. *arXiv preprint arXiv:1010.5265*.

Stein, Charles. 1956. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Pages 197–206 of: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1.

Stephens, Matthew, & Balding, David J. 2009. Bayesian Statistical Methods for Genetic Association Studies. *Nature Reviews Genetics*, **10**(10), 681–690.

Sun, Tingni, & Zhang, Cun-Hui. 2012. Scaled Sparse Linear Regression. *Biometrika*, **99**(4), 879–898.

Tang, Xueying, Xu, Xiaofan, Ghosh, Malay, & Ghosh, Prasenjit. 2016. Bayesian Variable Selection and Estimation Based on Global-Local Shrinkage Priors. *Sankhya A*, 1–32.

Terenin, Alexander, Dong, Shawfeng, & Draper, David. 2016. GPU-Accelerated Gibbs Sampling. *arXiv:1608.04329 [cs, stat]*, Aug.

Tiao, George C., & Tan, W. Y. 1966. Bayesian analysis of random-effect models in the analysis of variance. II. Effect of autocorrelated errors. *Biometrika*, **53**, 477–495.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, **58**, 267–288.

Tibshirani, Robert, Saunders, Michael A., Rosset, Saharon, Zhu, Ji, & Knight, Keith. 2005. Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **67**(1), 91–108.

Tibshirani, Robert J. 2014. In Praise of Sparsity and Convexity. *Past, Present, and Future of Statistical Science*, 497–505.

Tibshirani, Ryan J., & Taylor, Jonathan. 2011. The solution path of the generalized lasso. *Ann. Statist.*, **39**(3), 1335–1371.

Tibshirani, Ryan J, Hoefling, Holger, & Tibshirani, Robert. 2011. Nearly-Isotonic Regression. *Technometrics*, **53**(1), 54–61.

Tikhonov, Andrey. 1963. Solution of incorrectly formulated problems and the regularization method. *Soviet Meth. Dokl.*, **4**, 1035–1038.

Tseng, Paul. 2001. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of optimization theory and applications*, **109**(3), 475–494.

Van de Geer, Sara, Bühlmann, Peter, Ritov, Yaacov, Dezeure, Ruben, *et al.* 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**(3), 1166–1202.

van der Pas, SL, Kleijn, BJK, & van der Vaart, AW. 2014. The Horseshoe Estimator: Posterior Concentration around Nearly Black Vectors. *Electronic Journal of Statistics*, **8**, 2585–2618.

van der Pas, Stéphanie, Salomond, Jean-Bernard, & Schmidt-Hieber, Johannes. 2016a. Conditions for Posterior Contraction in the Sparse Normal Means Problem. *Electronic Journal of Statistics*, **10**, 976–1000.

van der Pas, Stephanie, Scott, James, Chakraborty, Antik, & Bhattacharya, Anirban. 2016b. *horseshoe: Implementation of the Horseshoe Prior*. R package version 0.1.0.

van der Pas, Stéphanie, Szabó, Botond, & van der Vaart, Aad. 2016c. How Many Needles in the Haystack? Adaptive Inference and Uncertainty Quantification for the Horseshoe. *arXiv:1607.01892*.

van der Pas, Stéphanie, Szabó, Botond, der Vaart, van, & Aad. 2017. Adaptive Posterior Contraction Rates for the Horseshoe. *arXiv:1702.03698*.

Wang, Hao, & Pillai, Natesh S. 2013. On a Class of Shrinkage Priors for Covariance Matrix Estimation. *Journal of Computational and Graphical Statistics*, **22**(3), 689–707.

Wei, Ran. 2017. *Bayesian Variable Selection Using Continuous Shrinkage Priors for Nonparametric Models and Non-Gaussian Data.* Ph.D. thesis, North Carolina State University.

Witten, Daniela M., Tibshirani, Robert, & Hastie, Trevor. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**(3), 515–534.

Yuan, Ming, & Lin, Yi. 2006. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.

Zhang, Cun-Hui. 2010. Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**(2), 894–942.

Zhang, Cun-Hui, & Zhang, Stephanie S. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1), 217–242.

Zhang, Yan, Reich, Brian J, & Bondell, Howard D. 2016. High Dimensional Linear Regression via the R2-D2 Shrinkage Prior. *arXiv preprint arXiv:1609.00046*.

Zhao, Peng, & Yu, Bin. 2006. On model selection consistency of Lasso. *Journal of Machine learning research*, **7**(Nov), 2541–2563.

Zou, Hui. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American statistical association*, **101**(476), 1418–1429.

Zou, Hui, & Hastie, Trevor. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.