Lasso Meets Horseshoe

Anindya Bhadra[*]
Purdue University

Jyotishka Datta[†]
University of Arkansas

Nicholas G. Polson [‡]
University of Chicago
Booth School of Business

Brandon T. Willard[§]
University of Chicago
Booth School of Business

June 30, 2017

Abstract

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

Keywords: Sparsity; regression; Lasso; global-local priors; horseshoe; horseshoe+; regularization; hyper-parameter tuning.

[*]*Address:* 250 N. University St., West Lafayette, IN 47907, email: bhadra@purdue.edu.
[†]*Address:* SCEN 309, 1 University of Arkansas, Fayetteville, AR, 72701, email: jd033@uark.edu.
[‡]*Address:* 5807 S. Woodlawn Ave., Chicago, IL 60637, email: ngp@chicagobooth.edu.
[§]*Address:* 5807 S. Woodlawn Ave., Chicago, IL 60637, email: bwillard@uchicago.edu.

1 Introduction

High-dimensional variable selection and sparse signal recovery have become routine practice in many statistics and machine learning applications. This has led to a vast growing literature for both classical and Bayesian methodologies for computation of large scale inference estimators. Whilst this area is too large to cover in a single review article, we revisit two popular sparse parameter estimation approaches, the classical Lasso (Tibshirani, 1996) and Bayes horseshoe estimation (Carvalho et al., 2010). We focus on and contrast three areas: performance in high-dimensional data, theoretical optimality and computational efficiency.

Formal definitions of sparsity and ultra-sparsity relies on the property of a few large signals among many (nearly) zero noisy observations. A common goal in high dimensional inference problems is to identify the low-dimensional signals observed in white noise. This encompasses three related areas:

(i) Estimation of the underlying sparse parameter vector.

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

(iii) Subset selection in regression where the number of covariates p is far larger than n.

The current literature provides a rich variety of methodologies for high-dimensional inference based on regularization which implicitly or explicitly penalizes models based on their dimensionality. The gold standard is Lasso (Least Absolute Shrinkage and Selection Operator) that produces a sparse point estimate by constraining the ℓ_1 norm of the parameter vector. Lasso's widespread popularity is due to a number of factors, particularly its computational efficiency with Least Angle Regression method due to Efron et al. (2004) and simple coordinate descent approaches of Friedman et al. (2007) as well as its ability to produce a sparse solution, with optimality (oracle) properties for both estimation and variable selection. Bühlmann

and van de Geer (2011), Hastie et al. (2015), James et al. (2013) provide excellent references for different aspects of Lasso and its various modifications.

Broadly speaking, Bayes procedures are classified into two categories: discrete mixtures or "two-groups" model or "spike-and-slab" priors (Bogdan et al., 2011, Efron, 2008, 2010, Johnstone and Silverman, 2004) and shrinkage priors (Armagan et al., 2011, 2013, Carvalho et al., 2009, 2010, Castillo and van der Vaart, 2012, Griffin and Brown, 2010, Polson and Scott, 2010b). The first places a point mass at zero and an absolutely continuous prior on the non-zero elements of the parameter vector. The second entails placing absolutely continuous shrinkage priors on the entire parameter vector, that shrink the entire coefficient towards zero. 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 and Scott, 2016).

Both Lasso and the Bayesian procedures come with strong theoretical gurantees for estimation, prediction and variable selection. Lasso and Horseshoe possess oracle properties - risk attainable only under some knowledge about the underlying model. 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 4). Horseshoe has been shown to achieve oracle properties in variable selection (Datta and Ghosh, 2013) and near-minimaxity in estimation (van der Pas et al., 2017) and improved prediction performance (Bhadra et al., 2016a), although theoretical studies of the horseshoe is still an active area.

The rest of the paper is organized as follows: Section 2 provides some historical background for the normal means (a.k.a the Gaussian compound decision problem) and the Gaussian linear 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 Lasso and the horseshoe method and we provide discussion and directions for future work in Section 5.

2   Sparse Means, Regression and Variable Selection

2.1   Sparse Normal Means:

As a starting point, suppose 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" or "nearly black" vector case occurs when the parameter vector θ lies in the set $l_0[p_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq p_n\}$ with the upper bound on the number of non-zero parameter values $p_n = o(n)$ as $n \to \infty$.

A natural Bayesian solution for inference under sparsity is the two-groups model that puts a non-zero probability spike at zero and a suitable prior on the non-zero 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 and Silverman (2004) showed that a thresholding-based estimator for θ under the two-groups model with an empirical Bayes estimate for μ attains the minimax rate in ℓ_q norm for $q \in (0, 2]$ for θ that are either nearly black or belong to an ℓ_p ball of 'small' radius. Castillo and van der Vaart (2012) treated a full Bayes version of the problem and again found an estimate that is minimax in ℓ_q norm for mean vectors that are either nearly black or have bounded weak ℓ_p norm for $p \in (0, 2]$.

We now turn to the problem of variable selection in sparse regression.

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 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Our focus is on the sparse solution

4

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:

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

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

The popular 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 and 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 θ. We discuss the penalization methods from a Bayesian viewpoint in the next section.

2.3 Variable Selection:

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

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 Castillo et al. (2015):

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

Bayesian approaches for sparse linear regression include (George and Foster, 2000, George, 2000, Ishwaran and Rao, 2005, Mitchell and Beauchamp, 1988) and more recently Ročková and George (2016), who introduced 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, γ 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.

Despite the attractive theoretical properties outlined above, the discrete indicators in spike-and-slab models give rise to a combinatorial problem. While some posterior point estimates such as the posterior mean or quantiles might be easily computable for spike-and-slab (Castillo et al., 2015, Castillo and van der Vaart, 2012), exploring the full posterior using Markov chain Monte Carlo (MCMC) is typically more challenging using point mass mixture priors. Ročková and 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 and 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.

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, thus leading to fast computation. Carvalho et al. (2009, 2010), Polson and Scott (2010b, 2012) 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 Section 3.3.

The estimators resulting from the one-group shrinkage priors are very different from the shrinkage estimator due to James and Stein (1961), who showed that maximum likelihood estimators for multivariate normal means are inadmissible beyond $\mathbb{R}^2$. 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

3.1 Bayesian Regularization : A Useful Duality

Regularization requires the researcher to specify a measure of fit, denoted by $l(\theta)$ and a penalty function, denoted by $\phi(\theta)$. Probabilistically, $l(\theta)$ and $\text{pen}_\lambda(\theta)$ correspond to the negative logarithms of the likelihood and prior distribution, respectively.

Regularization leads to an optimization problem of the form

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

and the probabilistic approach leads to a Bayesian hierarchical model

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

For appropriate $l(y \mid \theta)$ and $\text{pen}_\lambda(\theta)$, the solution to (3.1) corresponds to

the posterior mode of (3.2), $\hat{\theta} = \text{argmax}_\theta \, p(\theta \mid y)$, where $p(\theta \mid y)$ 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 (Ridge) corresponds to a (Hoerl and Kennard, 1970) Gaussian prior–under the same observation distribution, and an ℓ_1 penalty (Lasso) (Tibshirani, 1996) corresponds to a double-exponential prior (Park and 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 multicollinearity in the design (Hastie et al., 2009, ch.3). Shrinkage often wins in terms of predictive performance. However, shrinkage methods do not give exact zeros, which is preferred over dichotomous models by some practitioners (Stephens and Balding, 2009). Ultimately, both selection and shrinkage have their advantages and disadvantages.

3.2 Lasso Penalty and Prior

The Lasso based estimate of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that maximizes the ℓ_1 penalized log-likelihood, or equivalently, the posterior mode under a component-wise Laplace prior, as given below:

$$\text{(Penalty)}: \ \text{pen}_\lambda(\boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} |\theta_j| \ \Leftrightarrow \ \pi_\lambda(\boldsymbol{\theta}) = \exp(-\lambda \sum_{j=1}^{p} |\theta_j|) \ \text{(Prior)} \ (3.3)$$

The posterior mode is therefore the same as the classical Lasso-based estimate, and the mode inherits the optimal properties of Lasso proved by Bühlmann and van de Geer (2011). For example, the Oracle inequality in Bühlmann and van de Geer (2011, Eq. (2.8), Th. (6.1)) states that up to $O(\log(p))$ and a compatibility constant ϕ_0^2, the mean squared prediction error is of the same order as if one knew active set $S_0 = \{j : \theta_j^0 \neq 0\}$.) Lasso also exhibits other desirable properties such as computational tractability, consistency of point estimates of $\boldsymbol{\theta}$ for suitably tuned λ, and optimality

results on variable selection.

As previously discussed, the posterior mode of $\boldsymbol{\theta}$ under the double exponential prior will have all the above optimal properties of the Lasso estimate of $\boldsymbol{\theta}$. Unfortunately, the same is not expected to hold for the posterior mean, which is the Bayes estimate under squared error loss. 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 Datta and Ghosh (2013), Polson and Scott (2010b), and motivates the key properties of global-local priors. Figure 1 provides profile plots as a diagnostic of shrinkage behaviour for different priors.

Consider the normal means model: $y_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1), \theta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \lambda_i^2 \tau^2)$, 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 identifying 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 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 recovering large signals.

For correlated predictors, Zou and Hastie (2005) proposed a family of convex penalty called 'elastic net', which is a hybrid between Lasso and
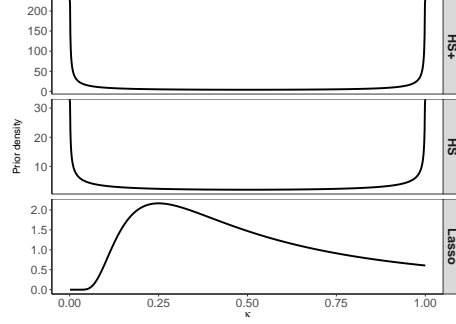
Figure 1: Shrinkage profile for Horseshoe, Horseshoe+, and Laplace prior.

Ridge. The penalty term is $\sum_{j=1}^{p} \lambda p_\alpha(\theta_j)$, where

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

Both Lasso and Elastic net facilitate efficient Bayesian computation via a global-local scale mixture representation (Bhadra et al., 2016c). The Lasso penalty arises as a Laplace global-local mixture (Andrews and Mallows, 1974), while the elastic-net regression can be recast as a global-local mixture with a mixing density belonging to the orthant-normal family of distributions (Hans, 2011). The orthant-normal prior on θ_i, given hyper-parameters λ_1 and λ_2, has a density function with the following form:

$$p(\theta_i \mid \lambda_1, \lambda_2) = \begin{cases} \phi(\theta_i \mid \frac{\lambda_1}{2\lambda_2}, \frac{\sigma^2}{\lambda_2})/2\Phi\left(-\frac{\lambda_1}{2\sigma\lambda_2^{1/2}}\right), & \theta_i < 0, \\ \phi(\theta_i \mid \frac{-\lambda_1}{2\lambda_2}, \frac{\sigma^2}{\lambda_2})/2\Phi\left(-\frac{\lambda_1}{2\sigma\lambda_2^{1/2}}\right), & \theta_i \geq 0. \end{cases} \tag{3.4}$$

3.3 Horseshoe Penalty and Prior

The horseshoe prior is a continuous "one-group" shrinkage rule based on what they call the horseshoe prior for multiple testing and model selection. Specifically, the horseshoe prior for θ_i, given a global shrinkage parameter τ, is given by the hierarchical model

$$(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2), \, , (\theta_i \mid \lambda_i, \tau) \sim \mathcal{N}(0, \lambda_i^2 \tau^2)$$

10

$$\lambda_i^2 \sim C^+(0,1), \quad i = 1,\ldots,n. \tag{3.5}$$

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:

$$\mathbb{E}(\theta_i \mid y_i) = (1 - \mathbb{E}(\kappa_i \mid y_i))y_i \text{ where } \kappa_i = 1/(1 + \lambda_i^2\tau^2) \tag{3.6}$$

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'. 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 and 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

$$\mathbb{E}(\theta_i \mid y_i, \tau) = y_i \left(1 - \frac{2\Phi_1(\frac{1}{2}, 1, \frac{5}{2}, \frac{y_i^2}{2\sigma^2}, 1 - \frac{1}{\tau^2})}{3\Phi_1(\frac{1}{2}, 1, \frac{3}{2}, \frac{y_i^2}{2\sigma^2}, 1 - \frac{1}{\tau^2})}\right) , \qquad (3.7)$$

where Φ_1 is the bivariate confluent hypergeometric function. This enables one to rapidly calculate the posterior mean estimator under the horseshoe prior via a "plug-in" approach with estimated values of the hyper-parameter τ. In a series of fundamental papers, van der Pas et al. (2014, 2016a, 2017, 2016c) 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 Section 4.

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

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

These priors are collectively called "global-local" shrinkage priors in Polson and 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., 2013), the three-parameter beta (Armagan et al., 2011), and the more recent horseshoe+ (Bhadra et al., 2016d) and the Dirichlet-Laplace (Bhattacharya et al., 2015b) priors. A natural question is *how do we compare these priors?* It is known due to several authors (Bhadra et al., 2016b, Polson and Scott, 2010b, 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. Below we list a few popular global-local shrinkage priors along with their behavior near origin and the tails. A detailed list of shrinkage priors proposed in the recent past is deferred to Section 7.

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

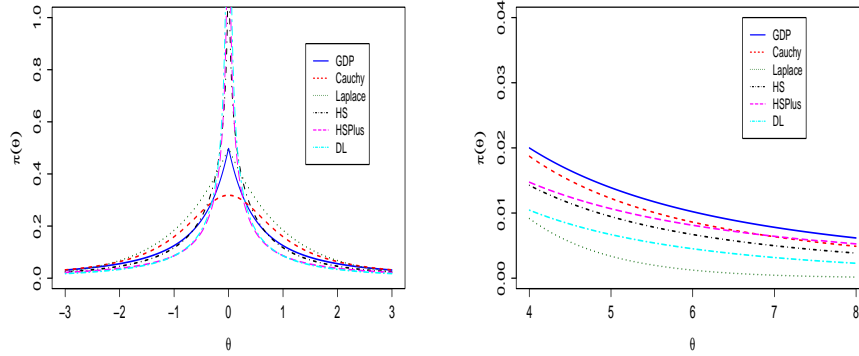Table 1: Different Priors: Behavior near origin and tails

Figure 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.

One way to judge a prior is by the penalty it induces in a regularization framework (3.1). For a prior $p(\theta)$, the induced penalty is given by $-\log p(\theta)$ as described in (3.2). Although the horseshoe prior leads to optimal performance as a shrinkage prior, the induced penalty does not admit a closed form as the marginal prior 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. (2017) 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_i, yet with a spike at origin and heavy tails and more importantly, admits a global-local scale mixture representation. The scale mixture representation supports both a traditional MCMC sampling for uncertainty quantification in full Bayes inference and EM/MM or proximal learning when computational efficiency is the primary concern. Since the aim of designing a sparsity prior is achieving higher spike near zero while maintaining regularly varying tails, a useful strategy is to split the range of the prior into disjoint intervals: $[0, 1)$ and $[1, \infty)$, and aim for higher spike in one and heavier tail in the other. This leads to a class of 'horseshoe-like' priors with more flexibility in shape than any single shrinkage prior. We provide the general form of horseshoe-like priors and a key representation theorem. See Bhadra et al. (2017) for the proofs and more details.

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

$$\tilde{p}_{\tilde{H}S}(\theta_i \mid \tau^2) = \frac{1}{2\pi\tau} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right), \quad \theta_i \in \mathbb{R},\ \tau > 0. \quad (3.8)$$

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

$$p_{hs}(\theta_i \mid \tau^2) \propto \begin{cases} \frac{1}{\theta_i^{1-\epsilon}} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right) & \text{if } |\theta_i| < 1 \\ \theta_i^{1-\epsilon} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right) & \text{if } |\theta_i| \geq 1, \end{cases} \quad \epsilon \geq 0, \tau > 0. \quad (3.9)$$

Normal scale mixture The horseshoe-like prior (3.8) is a Gaussian scale mixture with a Slash normal density, which is in turn a Normal scale mixture of $\text{Pareto}(1/2)$ density, yielding the following representation theorem:

THEOREM 3.1 (Bhadra et al. (2017))**.** *The horseshoe-like prior in* (3.8)

has the following global-local scale mixture representation:

$$(\theta_i \mid t_i, \tau) \sim \mathcal{N}\left(0, \frac{\tau^2}{t_i^2}\right), \quad (t_i \mid s_i) \sim \mathcal{N}(0, s_i),$$

$$s_i \sim \text{Pareto}\left(\frac{1}{2}\right), \quad t_i \in \mathbb{R}, \ \tau \geq 0. \tag{3.10}$$

Table 2: 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	$2/\left\{\pi\tau(1+(\lambda_i/\tau)^2)\right\}$	$\frac{\tau}{\sqrt{\kappa_i(1-\kappa_i)}}\frac{1}{(1+\kappa_i(\tau^2-1))}$
Horseshoe+	$\frac{4\log\lambda_i/\tau}{\left\{\pi^2\tau(\lambda_i/\tau)^2-1)\right\}}$	$\frac{\tau}{\sqrt{\kappa_i(1-\kappa_i)}}\frac{\log\left\{(1-\kappa_i)/\kappa_i\tau^2\right\}}{(1-\kappa_i(\tau^2+1))}$
Double Exponential	$\lambda_i\exp(-\lambda_i^2/2)$	$\kappa_i^{-2}\exp-\frac{1}{2\kappa_i}$

4 Statistical Risk Properties

4.1 History of Shrinkage Estimation: 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 $\hat{\theta}^{JS} = (1 - (m-2)/\|\mathbf{y}\|^2)\mathbf{y}$ with posterior mean $\hat{\theta}_{\text{Bayes}} = (\tau^2/\tau^2+1)\mathbf{y}$, which corresponds to the Bayes risk of $m(\tau^2/\tau^2+1)$. estimate. James and Stein (1961) proved that this estimator dominates the MLE in terms of the expected total squared error for every choice of $\boldsymbol{\theta}$, i.e. it outperforms the MLE no matter what the true $\boldsymbol{\theta}$ is. To motivate the need for developing new prior distributions, consider the classic James–Stein "global" shrinkage rule, $\hat{\theta}_{JS}(\mathbf{y})$. The JS estimator uniformly dominates the traditional sample mean estimator, $\hat{\theta}$. For all values of the true parameter θ and for $n > 2$, we have the classical mean squared error (MSE) risk bound:

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

For sparse signal problem the standard James–Stein shrinkage rule, $\hat{\theta}_{JS}$, 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 and Silverman (2004) show that $E\|\hat{\theta}^{JS} - \theta\| \leq n$ with risk 2 at the origin. Moreover, we can bound (for $\sigma^2 = 1$)

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

and so $\hat{\theta}_{JS}(\mathbf{y})$ 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 and Johnstone (1994) showed, simpler rules such as the hard-thresholding and soft-thresholding estimates given by $\hat{\theta}^H(y, \lambda) = yI\{|y| \geq \lambda\}$ and $\hat{\theta}^S(y, \lambda) = sgn(y)(|y| - \lambda)_+$ satisfy an Oracle inequality. In particular, when the thresholding sequence is close to $\sqrt{2 \log n}$ ('universal threshold'), these estimators attain the "oracular 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 ℓ_2 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 $\delta(Y)$, we have a lower bound:

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0}\|\delta(Y) - \theta_0\|^2 \geq 2p_n \log(n/p_n)(1 + o(1)) \ (\sigma^2 = 1) \qquad (4.1)$$

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.

Horseshoe estimators 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, 2017, 2016c). Specifically, the horseshoe estimator achieves

$$\sup_{\theta \in l_0[p_n]} \mathbb{E}_{y|\theta} \|\hat{\theta}_{HS}(y) - \theta\|^2 \asymp p_n \log\left(n/p_n\right), \; (\sigma^2 = 1) \tag{4.2}$$

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. 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. Independently, van der Pas et al. (2016a) and Ghosh and Chakrabarti (2017) showed that these optimality properties are not unique features of the horseshoe prior and they hold for a general class of global-local shrinkage priors. While the results of van der Pas et al. (2016a) apply to a wider class of priors, including the horseshoe+ prior (Bhadra et al., 2016d) and spike-and-slab Lasso (Ročková and George, 2016). We note that Ghosh and Chakrabarti (2017) attain sharper mean-squared error bounds.

4.3 Lasso:

A natural question is how does the Lasso fare in these aspects? While the MAP estimator under the Bayesian formulation of Lasso (i.e. i.i.d. Laplace prior on θ_i's) enjoys all the desirable properties of the frequentist Lasso, it is known to be sub-optimal for recovery of the underlying θ_0 (Castillo and van der Vaart, 2012). In fact, Castillo and van der Vaart (2012) show that unlike the mode, the full posterior distribution under the Laplace prior does not contract at the optimal rate, making it 'useless for *uncertainty quantification*'.

4.4 Asymptotic Bayes Optimality under Sparsity:

One of main reasons behind the widespread popularity of Lasso is the in-built mechanism for performing simultaneous shrinkage and selection. The frequentist Lasso or the equivalent MAP estimator under i.i.d. The double-exponential prior induces automatic sparsity and can be easily adjusted to achieve model selection consistency. The horseshoe estimator, on the other hand, is a shrinkage rule that induces a selection rule through thresholding the pseudo posterior inclusion probabilities. Thus, we can compare their relative performance for multiple testing under the two-groups model and a 0-1 additive loss framework. It turns out 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 can pushes most of the density to 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.

$$\text{Reject the } i^{th} \text{ null hypothesis } H_{0i} : \theta_i = 0 \text{ if } 1 - \hat{\kappa}_i > \frac{1}{2} . \qquad (4.3)$$

Under the two-groups model (A.2), and a 0-1 loss, the Bayes risk is

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

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 "oracular 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 asymptotic framework of Bogdan et al. (2011) is

$$p_n \to 0, \ u_n = \psi_n^2 \to \infty, \ \text{and} \ \log(v_n)/u_n \to C \in (0, \infty) \qquad (4.4)$$

where $v_n = \psi_n^2 (\frac{1-p_n}{p_n})^2$. The Bayes risk for the Bayes oracle under the above framework (4.4) 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 oracular 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 and Ghosh (2013) first showed that the decision rule (4.3) 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 and Ghosh (2013) hinges on the concentration of the posterior distribution near 0 and 1, depending on the trade-off between signal strength and sparsity. In numerical experiments, Datta and Ghosh (2013) also confirmed that the horseshoe prior induced 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 and Ghosh (2013), Ghosh and 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 $\lim_{n\to\infty} \tau/p \in (0, \infty)$.

5 Hyper-parameter Tuning

5.1 Optimization and Cross-validation

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 and 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, a fully Bayesian approach that specifies a hyper-prior on τ and second, an empirical Bayesian approach that estimates τ from the data using a simple thresholding or maximum marginal likelihood approach (MMLE). 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 $n = 100$. The success of the MMLE estimator, both theoretically and numerically, challenges the notion that for the horseshoe prior an empirical Bayes parameter estimate of τ cannot replace a full Bayes estimate of τ. In reality, one must take care to prevent the estimator from getting too close to zero.

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 zero in such situations. Yet another approach for handling τ was proposed by Piironen and 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.

5.2 Marginal Likelihood:

We now 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 in (3.5) after marginalising out θ_i from the model as:

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

(5.1)

Tiao and 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 (Datta and Ghosh, 2013, Polson and Scott, 2010b). 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 pre-empt this behaviour. 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 (3.5):

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

(5.2)

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

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

(5.3)

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

6   Computation and Simulation

Over the last few years, several different implementation of the horseshoe prior for normal means and regression model has been proposed. The MCMC based implementations usually proceed via block-updating θ, λ 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 and 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. (2015a) who used a Gaussian sampling alternative to the naïve Cholesky decomposition to reduce the computational burden from $O(p^3)$ to $O(n^p)$. A very recent paper by Johndrow and Orenstein (2017) claims to improve this even further by implementing a block update strategy but using a random 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 3.

Table 3: 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. (2015a)
GPU accelerated Gibbs sampling	Terenin et al. (2016)
bayesreg + MATLAB code in paper	Makalic and Schmidt (2016)
MATLAB code	Johndrow and Orenstein (2017)

7 Applications and Extensions

7.1 Applications and Extensions of Lasso:

Since the inception of Lasso as a regularisation 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. Although a comprehensive list of regularization methods that extend the idea of Lasso and even move beyond the convex penalty is beyond the scope of this article, we give a list of popular regularization methods in Table 4, which is adapted from Tibshirani (2014).

Table 4: A few regularization methods

Method	Authors
Adaptive Lasso	Zou (2006)
Compressive sensing	Candes (2008), Donoho (2006)
Dantzig selector	Candes and Tao (2007)
Elastic net	Zou and Hastie (2005)
Fused Lasso	Tibshirani et al. (2005)
Generalized Lasso	Tibshirani and Taylor (2011)
Graphical Lasso	Friedman et al. (2008)
Grouped Lasso	Yuan and Lin (2006)
Hierarchical interaction models	Bien et al. (2013)
Matrix completion	Candès and 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 and Zhang (2012)
Minimum concave penalty	Zhang (2010)
SparseNet	Mazumder et al. (2012)

7.2 Applications and Extensions of Horseshoe:

As discussed in Section 3.3, the Horseshoe prior belongs to a wider class of global-local shrinkage priors (Polson and 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 Ghosh and Chakrabarti (2017), Ghosh et al. (2016), van der Pas et al. (2016a) have provided conditions for optimality of one-group continuous priors for estimation of sparse normal means and multiple testing. Table 5 provides a sampling of a few continuous shrinkage priors popular in the literature.

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. (2016b) 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. (2016a) show how to use global-local priors for prediction and provide theoretical and numerical evidence that it performs better than a variety of competitors including Lasso, Ridge, PCR and sparse PLS.

Moving beyond Gaussianity, Datta and Dunson (2016) re-discovered the Gauss-hypergeometric prior for flexible shrinkage needed for quasi-sparse count data, with a tighter control on false discoveries. Piironen and Vehtari (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 and 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.

Table 5: A catalog of global-local shrinkage priors

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

Table 6: Applications of the horseshoe prior

Application	Authors
Fadeout method for mean-field variational inference under non-centered parameterizations and stochastic variational inference for undirected graphical model.	Ingraham and Marks (2016)
Linear regression for Causal inference and Instrumental variable models	Hahn et al. (2016), Hahn and Lopes (2014)
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 and Minin (2015)
Quasi-Sparse Count Data	Datta and Dunson (2016)
Variable Selection under the projection predictive framework	Piironen and Vehtari (2015)

8 Discussion

Lasso and Horseshoe prior regularization–and the general class of global-local shrinkage priors–are widely used for sparse signal recovery in high dimensional data. 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 is its heavy tails and probability spike at zero, which makes it adaptive to sparsity and robust to large signals. A number of MCMC sampling algorithms 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 3. 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 (Bhadra et al., 2017).

We have discussed the theoretical optimality properties for both Lasso and horseshoe estimator. The optimality properties of Lasso are well-known and they depend on "neighbourhood stability" or "irrepresentability" condition and "beta-min" condition. Informally, these conditions guarantee against ill-posed design matrix and separability of signal and noise parameters. The horseshoe estimator enjoys Kullback–Leibler super-efficiency in true density recovery, near-minimaxity in estimation, and asymptotic Bayes optimality for multiple testing using the pseudo-inclusion probabilities as discussed in Section 4. The horseshoe priors are also good default priors for many-to-one functionals as shown in Bhadra et al. (2016b), but a thorough study of horseshoe prior for default Bayes problems is still an unexplored area.

This is still a fruitful area for future research. For example, the horseshoe prior and related classes can be approached with new computational strategies, such as proximal algorithms. One promising area involves observation distributions from the exponential family, and whether or not the optimality properties carry over to the non-Gaussian cases. Another interesting direction could include structured sparsity under the horseshoe prior, such as grouped variable selection and Gaussian graphical models.

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 $\omega_i = P(\theta_i \neq 0|y_i)$. Formally, in a two-groups model θ_i's are modeled

as

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

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.

References

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

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

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Berzuini, C., Guo, H., Burgess, S., and Bernardinelli, L. (2016). Mendelian Randomization with Poor Instruments: a Bayesian Approach. *arXiv:1608.02990 [math, stat]*. arXiv: 1608.02990.

Bhadra, A., Datta, J., Li, Y., Polson, N. G., and Willard, B. (2016a). Prediction risk for the horseshoe regression. *arXiv preprint arXiv:1605.04796*.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016b). Default Bayesian Analysis with Global-Local Shrinkage Priors. *Biometrika*, to appear.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016c). Global-Local Mixtures. *arXiv preprint arXiv:1604.07487*.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016d). The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Analysis*, to appear.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). Horseshoe Regularization for Feature Subset Selection. *arXiv preprint arXiv:1702.07400*.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2015a). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *arXiv:1506.04778 [stat]*.

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

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111–1141.

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 39(3):1551–1579.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

Candes, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592.

Candès, E. J. and Tao, T. (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., and Scott, J. G. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research W&CP*, 5:73–80.

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

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

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

Datta, J. and Dunson, D. B. (2016). Bayesian inference on quasi-sparse count data. *Biometrika*, 103(4):971–983.

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

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

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

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1. Cambridge University Press.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Faulkner, J. R. and Minin, V. N. (2015). Bayesian trend filtering: adaptive temporal smoothing with shrinkage priors.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., and others (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

George, E. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Ghosh, P. and Chakrabarti, A. (2017). Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Anal.* Advance publication.

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

Gramacy, R. B., Pantaleo, E., and others (2010). Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. *Bayesian Analysis*, 5(2):237–262.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Hahn, P. R., He, J., and Lopes, H. (2016). Elliptical slice sampling for Bayesian shrinkage regression with applications to causal inference. Technical report.

Hahn, P. R. and Lopes, H. (2014). Shrinkage priors for linear instrumental variable models with many instruments. *arXiv preprint arXiv:1408.0462*.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association*, 106(496):1383–1393.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The Elements of Statistical Learning*, volume 2. Springer.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*. CRC press.

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

Ingraham, J. B. and Marks, D. S. (2016). Bayesian Sparsity for Intractable Distributions. *arXiv preprint arXiv:1602.03807*.

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

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 6. Springer.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379.

Johndrow, J. E. and Orenstein, P. (2017). Scalable MCMC for Bayes Shrinkage Priors. *arXiv preprint arXiv:1705.00841*.

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

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of computational and Graphical Statistics*, 12(3):531–547.

Magnusson, M., Jonsson, L., and Villani, M. (2016). DOLDA - a regularized supervised topic model for high-dimensional multi-class regression. *arXiv preprint arXiv:1602.00260*.

Makalic, E. and Schmidt, D. F. (2016). High-Dimensional Bayesian Regularised Regression with the BayesReg Package. *arXiv preprint arXiv:1611.06649*.

Mazumder, R., Friedman, J. H., and Hastie, T. (2012). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106:1125–1138.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322.

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

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

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

Conference on Bayesian Modeling Applications Workshop-Volume 1218, pages 79–88. CEUR-WS. org.

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

Piironen, J. and Vehtari, A. (2016). On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. *arXiv preprint arXiv:1610.05559*.

Polson, N. G. and Scott, J. G. (2010a). Large-scale simultaneous testing with hypergeometric inverted-beta priors. *arXiv preprint arXiv:1010.5223*.

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

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Polson, N. G. and Scott, J. G. (2016). Mixtures, envelopes and hierarchical duality. *Journal of the Royal Statistical Society. Series B*, 78:701–727.

Ročková, V. and George, E. I. (2016). The spike-and-slab lasso. *Journal of the American Statistical Association*, (just-accepted).

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

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Terenin, A., Dong, S., and Draper, D. (2016). GPU-accelerated Gibbs Sampling. *arXiv:1608.04329 [cs, stat]*.

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

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

Tibshirani, R., Saunders, M. A., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(1):91–108.

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

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics*, 53(1):54–61.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.*, 39(3):1335–1371.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

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

van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016a). Conditions for Posterior Contraction in the Sparse Normal Means Problem. *Electronic Journal of Statistics*, 10:976–1000.

van der Pas, S., Scott, J., Chakraborty, A., and Bhattacharya, A. (2016b). *horseshoe: Implementation of the Horseshoe Prior*. R package version 0.1.0.

van der Pas, S., Szabó, B., der Vaart, v., and Aad (2017). Adaptive posterior contraction rates for the horseshoe. *arXiv:1702.03698*.

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

Wang, H. and Pillai, N. S. (2013). On a class of shrinkage priors for covariance matrix estimation. *Journal of Computational and Graphical Statistics*, 22(3):689–707.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Yuan, M. and Lin, Y. (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, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, Y., Reich, B. J., and Bondell, H. D. (2016). High Dimensional Linear Regression via the R2-D2 Shrinkage Prior. *arXiv preprint arXiv:1609.00046*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.