Lasso Meets Horseshoe

Anindya Bhadra [*] Purdue University
Jyotishka Datta [†] University of Arkansas
Nicholas G. Polson [‡] and Brandon T. Willard [§]
The University of Chicago Booth School of Business

May 12, 2017

Abstract

The goal of our paper is to survey the major advances for Lasso and Horseshoe sparse signal recovery regularisation methodologies. Lasso and its variants are a gold standard for best subset selection of predictors while Horseshoe is a state-of-the-art Bayesian estimator for sparse signals. Lasso is scalable and fast using convex optimization whilst the Horseshoe penalty is non-convex with theoretical guarantees of minimizing the Bayes risk under quadratic loss. 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 has become routine practice in many Statistics and machine learning applications. This has led to a vast growing literature for both frequentist and Bayesian methodologies for computation of large scale inference estimators. Whilst the general area is too large to cover in a single review article, we revisit two popular approaches to sparse parameter estimation problems, the classical Lasso (Tibshirani, 1996) and the Bayes horseshoe estimator (Carvalho et al., 2010). We compare 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 zero or 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, yet different areas:

1. Estimation of the underlying sparse parameter vector.
2. Multiple testing where the number of tests is much larger than the sample size, and
3. Subset selection in linear model where the number of covariates p is much larger than the sample size n.

Current research provides a rich variety of methodologies for high-dimensional inference based on regularisation which implicitly or explicitly penalizes models based on their dimensionality. The gold standard is Lasso (acronym for 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 its computational efficiency based on the Least Angle Regression method due to Efron et al. (2004) as well as its ability to produce a sparse solution, with optimality properties for both estimation and variable selection. Bühlmann and van de Geer (2011), Hastie et al. (2015), James et al. (2013) provide excellent references for different aspects of Lasso and its various modifications.

Bayesian alternatives are numerous and can be broadly classified into two categories: discrete mixtures or "two-groups" model or "spike-and-slab" priors (vide Bogdan et al. (2011), Efron (2008, 2010), Johnstone and Silverman (2004)) and shrinkage priors Armagan et al. (2011, 2013), Carvalho et al. (2009, 2010), Castillo and van der Vaart (2012), Griffin and Brown (2010), Polson and Scott (2010b)). The first approach places a point mass at zero and an absolutely continuous prior on the non-zero elements of the parameter vector. The second approach 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 we discuss the trade-offs associated with choosing one over the other a little later. It should also be noted that most of the penalization approaches can be interpreted in a Bayesian sense, by considering the mode of the posterior distribution under an appropriate shrinkage prior.

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

The rest of the paper is organized as follows: Section 2 provides some historical background for the normal means or the Gaussian compound decision problem and the normal regression problem. 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

Sparse Normal Means: As a starting point, suppose we observe data from the probability model $(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, 1)$ 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. 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.

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 situation where $p \gg n$ and "most" of θ_i's are zero. Like the sparse normal means problem, our goal is to identify the non-zero entries of $\boldsymbol{\theta}$ as well as estimate it. There is 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 most popular of these methods is the Lasso that uses an ℓ_1 penalty, i.e. $p_\lambda(\theta_j) = -\lambda|\theta_j|$, that 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 also act as a tool for variable selection and estimation. Penalized methods can also be treated as finding the posterior mode under a prior which relates to the penalty in (2.1) via $p_\lambda(\boldsymbol{\theta}) = -\log(\pi(\boldsymbol{\theta}))$. We discuss the penalization methods from a Bayesian viewpoint in the next section.

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 does uniformly well when the true is dense. Lasso's success has inspired many estimation methods that rely on convexity and sparsity that entails.

A parallel surge of Bayesian methods has also been undertaken for sparse regression problems with an underlying variable selection procedure. A typical hierarchical Bayesian method 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, 2000, George and Foster, 2000, Ishwaran and Rao, 2005, Mitchell and Beauchamp, 1988) and more recently Rcková 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, **fl** 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 and van der Vaart, 2012, Castillo et al., 2015), exploring the full posterior using Markov chain Monte Carlo (MCMC) is typically more challenging using point mass mixture priors. Rcková and George (2016) commented on the inefficiency of the stochastic search algorithms for exploring the posterior even for moderate dimensions and developed a deterministic alternative to quickly find the maximum a-posteriori model. We note that (i) increasing the efficiency in computation in the spike-and-slab model remains an active area of research (see, e.g., Rcková 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 has a different motto: instead of putting a prior on the model space to yield a sparse estimator, they model the posterior inclusion probabilities $P(\theta_i \neq 0 \mid y_i)$ directly, thus

leading to fast computation. In a series of papers 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 provide a brief survey and comparison of these different priors and introduce a recently proposed horseshoe-like prior in Section 3.3.

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

$$\underset{\theta \in \Re^d}{\text{minimise}} \quad l(\theta) + \text{pen}_\lambda(\theta) \ . \tag{3.1}$$

The probabilistic approach leads to a Bayesian hierarchical model

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

The solution to the minimisation problem estimated by regularisation (3.1)

corresponds to the posterior mode, $\hat{\theta} = \arg\max_\theta\ p(\theta|y)$, where $p(\theta|y)$ denotes the posterior distribution (Polson and Scott, 2016). The properties of the penalty are then induced by those of the prior. For example, regression with a least squares log-likelihood subject to a penalty such as an ℓ_2-norm (ridge) (Hoerl and Kennard, 1970) Gaussian probability model or ℓ_1-norm (lasso) (Tibshirani, 1996) double exponential probability model.

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. Shrinkage often wins in terms of predictive performance. But shrinkage methods do not give exact zeros, which is preferred over dichotomous models by some practitioners (Stephens and Balding, 2009). Thus, 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 componentwise Laplace prior, as given below:

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

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 ex-

ponential 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. In fact, 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, these are deficiencies of the double exponential prior. For example, for shrinking small observations while maintaining robustness to the large observations is noted by various authors, including Datta and Ghosh (2013), Polson and Scott (2010b), the key property of global-local priors. See the shrinkage profile plots in Fig. 2 as well as in the aforementioned papers.

For correlated predictors, Zou and Hastie (2005) proposed a family of convex penalty 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. (2016d). 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

In a series of remarkable papers, Carvalho, Polson, and Scott (2009, 2010), Polson and Scott (2010b, 2012) introduced a continuous "one-group" shrinkage rule based on what they call the horseshoe prior for multiple testing and model selection. The horseshoe prior for θ_i, given a global shrinkage parameter τ is given by:

$$\begin{aligned}(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2),\ ,(\theta_i \mid \lambda_i, \tau) \sim \mathcal{N}(0, \lambda_i^2\tau^2) \\ \lambda_i^2 \sim \mathrm{C}^+(0,1), \quad i = 1, \ldots, n.\end{aligned} \tag{3.5}$$

As discussed before, 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' is attributed to the shape of the $\mathrm{Be}(\frac{1}{2}, \frac{1}{2})$ prior density of the shrinkage weight, κ_i, for each observation. 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 test-

ing rule attains the Bayes oracle up to a multiplicative constant.

We focus now on the behaviour of the posterior distribution under the horseshoe prior. Although the prior density under the horseshoe prior doesn't admit a closed form, we can write the horseshoe posterior mean by using the Tweedies' formula $\mathbb{E}(\theta \mid y) = y + \frac{\partial \ln m(y)}{\partial y}$ as:

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

where Φ_1 is the degenerate hypergeometric function of two variables. 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) 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 and Scott, 2010b):

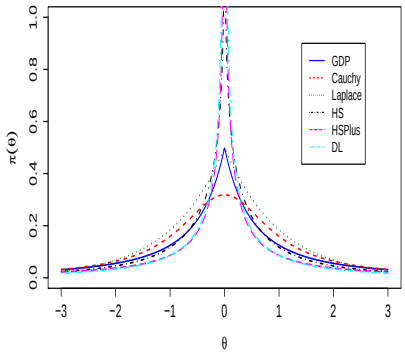
$$(\mathbf{y} \mid \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I});\ \theta_i \sim \mathcal{N}(0, \lambda_i^2\tau^2)$$
$$\lambda_i^2 \sim \pi(\lambda_i^2);\ (\tau, \sigma^2) \sim \pi(\tau^2, \sigma^2), i = 1, \ldots, n.$$

These priors are collectively called the "global-local" shrinkage priors after Polson and Scott (2010b) as 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., 2016c) and the Dirichlet-Laplace (Bhattacharya et al., 2015b) priors. A natural question is *how do we compare between these priors?* It is known due to several authors (Bhadra et al., 2016b, Polson and Scott, 2010b, van der Pas et al., 2016, e.g.) that the key features of

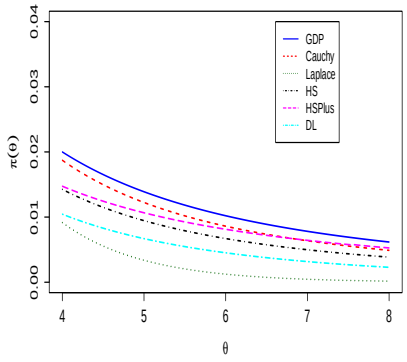
Prior	Origin Behavior	Tails
Horseshoe	$-\log(\lvert\theta\rvert)$	$\lvert\theta\rvert^{-2}$
Horseshoe+	$-\log(\lvert\theta\rvert)$	$\lvert\theta\rvert^{-1}$
Horseshoe-like	$-\lvert\theta\rvert^{1-\epsilon}\log(\lvert\theta\rvert)$	$\lvert\theta\rvert^{1-\epsilon}\ \epsilon \geq 0$
GDP	Bounded at origin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

Table 1: Different Priors: Behaviour near origin and tails

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 behaviour near origin and the tails. A detailed list of shrinkage priors proposed in the recent past is deferred to §7.



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



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

One way to judge a prior is by the penalty it induces in a regularisation 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, 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 form 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 form of the horseshoe-like prior and the general family and a key representation theorem, and direct the interested readers to Bhadra et al. (2017) for more details:

Horseshoe-like prior The horseshoe-like prior Bhadra et al. (2017) has the following marginal prior density for θ_i:

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

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

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

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 Pareto$(1/2)$ density, yielding the following representation theorem:

THEOREM 3.1. *The horseshoe-like prior in* (3.8) *has the following global-local scale mixture representation:*

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

3.4 Shrinkage Profiles and Sparsity

The better performance of horseshoe and horseshoe+ priors compared to double exponential or normal can also be explained by their shrinkage profile. Consider the marginal likelihood for the normal means model: $p(y_i \mid \kappa_i, \tau) = \kappa_i^{1/2} \exp\left(-\kappa_i y_i^2/2\right)$. 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 to κ-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.

4 Statistical Risk Properties

History of Shrinkage Estimation: Inadmissibility of MLE The story of shrinkage estimation goes back to the proof in Stein (1956) that the max-

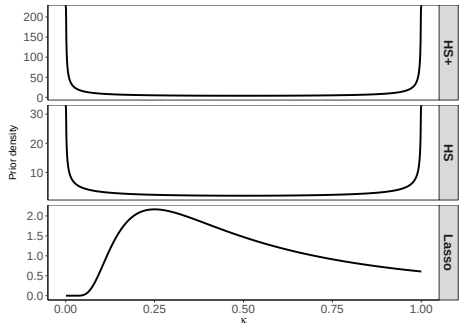


Figure 2: Shrinkage profile for Horseshoe, Horseshoe+, and Laplace prior.

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

imum likelihood estimators for normal data are inadmissible beyond $\mathbb{R}^2$. The James-Stein 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 $m({}^{\tau^2}/_{\tau^2+1})$. James and Stein 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, $\hat{\theta}_{JS}(\mathbf{y})$. This estimator uniformly dominates the traditional sample mean estimator, $\hat{\theta}$. For all values of the true parameter θ and for $n > 2$, we have the classical mean squared error (MSE) risk bound:

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

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 param-

eter 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

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

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 nowehere 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.

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:

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

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 is due to Ghosal et al. (2000) who 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, 2016,?, 2017). Specifically, the horseshoe estimator achieves

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

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. (2016) and Ghosh et al. (2016a) 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 van der Pas et al. (2016)'s results apply for a wider class of priors, including the horseshoe+ prior (Bhadra et al., 2016c) and spike-and-slab Lasso (Rcková and George, 2016), Ghosh et al. (2016a) attain sharper mean-squared error bounds.

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. iid 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 $\boldsymbol{\theta}_0$ (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*'.

Asymptotic Bayes Optimality under Sparsity One of main reasons behind the widespread popularity of Lasso is its in-built mechanism for per-

forming simultaneous shrinkage and selection. The frequentist Lasso or the equivalent MAP estimator under i.i.d. Laplace priors 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 the Laplace 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 interpreted as a pseudo-inclusion probability $P(\theta_i \neq 0 \mid y_i)$, and it induces a multiple testing rule.

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

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

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

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 called ABOS (asymptotic Bayes optimal under sparsity) 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, i.e. $\tau = O(\pi)$. Datta and Ghosh (2013)'s proof hinges on the concentration of posterior distribution near 0 or 1, depending on the trade-off between signal strength and sparsity. In their numerical experiments, Datta and Ghosh (2013) also confirmed the idea that the horseshoe prior induced rule outperforms the shrinkage rule induced by the Laplace prior under different 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 et al. (2016a,b), van der Pas et al. (2016) have shown that usual estimates of τ adapts to sparsity, a condition that also guarantees near-minimaxity in estimation. Ghosh et al. (2016b) 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

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., 2016,?). 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 ($p_n \log(n)$) 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:

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

More importantly, 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. Of course, one needs care to prevent the estimator to be 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. Like 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.

Marginal Likelihood: We 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:

$$m(y \mid \tau) = \prod_{i=1}^{n}(1+\lambda_i^2\tau^2)^{-\frac{1}{2}} \exp\left\{-\frac{y_i^2}{2(1+\lambda_i^2\tau^2)}\right\}(1+\lambda_i^2)^{-1}d\lambda_i \tag{5.1}$$

As Tiao and Tan (1965) point out, 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)'s approach, 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 in van der Pas et al. (2017)'s approach, we first calculate the marginal prior of θ_i after integrating out λ_i^2 in (3.5):

$$p_\tau(\theta_i) = \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\theta_i^2}{2\lambda^2\tau^2}\right\} \frac{1}{\lambda\tau} \frac{2}{\pi(1+\lambda^2)} d\lambda \tag{5.2}$$

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

$$\hat{\tau}_M = \underset{\tau\in[1/n,1]}{\operatorname{argmax}} \prod_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_i-\theta_i)^2}{2}\right\} p_\tau(\theta_i) d\theta_i \tag{5.3}$$

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 $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ 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 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 na$\acute{\text{i}}$ve 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.

7 Applications and Extensions

Applications and Extensions of Lasso: Since the inception of Lasso as a regularisation method for linear regression in 1996, a great deal of exten-

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. (2016)
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)

sions 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):

Applications and Extensions of Horseshoe: As discussed in §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 et al.

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 et al. (2011b)
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. (2011a)
Square Root Lasso	Belloni et al. (2011)
Scaled Lasso	Sun and Zhang (2012)
Minimum concave penalty	Zhang (2010)
SparseNet	Mazumder et al. (2012)

(2016a,b), van der Pas et al. (2016) 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 problem , 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. 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 few 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. (2016c)
Horseshoe-like	Bhadra et al. (2017)
Spike-and-Slab Lasso	Rcková and George (2016)
R2-D2	Zhang et al. (2016)

8 Discussion

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

1. So many global-local priors; what is the unifying theme? Spike and heavy tails Polson and Scott (2010b), van der Pas et al. (2016).

2. How close to minimax constant can we get in estimation? Only near-minimaxity and the best constant is $4\sigma^2$ (van der Pas et al., 2017).

3. How close to oracle risk can we get in testing? We can get exact rate (Ghosh et al., 2016b).

4. Is prediction performance optimal?

5. Yet to rigorously show global-local priors have any information theoretic properties in default Bayes problems that reference priors (Berger and Bernardo, 1992) enjoy.

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

A Two-groups Model

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

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

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

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

As can be seen from (A.2), the two-groups model leads to a sparse estimate, i.e., it puts exact zeros in the model.

References

David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102, 1974.

Artin Armagan, Merlise Clyde, and David B Dunson. Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems*, pages 523–531, 2011.

Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143, 2013.

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

James O Berger and José M Bernardo. On the development of reference priors. *Bayesian statistics*, 4(4):35–60, 1992.

Anindya Bhadra, Jyotishka Datta, Yunfan Li, Nicholas G Polson, and Brandon Willard. Prediction risk for global-local shrinkage regression. *arXiv preprint arXiv:1605.04796*, 2016a.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. Default Bayesian Analysis with Global-Local Shrinkage Priors. *Biometrika*, to appear, 2016b.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Analysis*, to appear, 2016c.

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

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

Anirban Bhattacharya, Antik Chakraborty, and Bani K. Mallick. Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *arXiv:1506.04778 [stat]*, June 2015a. URL http://arxiv.org/abs/1506.04778.

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015b.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

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

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

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

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

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

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

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480, 2010.

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

Ismail Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, October 2015. URL http://dx.doi.org/10.1214/15-AOS1334.

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

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

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

David L Donoho and Iain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, pages 425–455, 1994.

David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B*, 54:41–81, 1992.

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

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

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. URL http://projecteuclid.org/euclid.aos/1083178935.

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

Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, and others. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

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

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

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

EdwardI. George and Dean P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000. URL http://biomet.oxfordjournals.org/content/87/4/731.abstract.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, April 2000. URL http://dx.doi.org/10.1214/aos/1016218228.

Prasenjit Ghosh, Arijit Chakrabarti, and others. Asymptotic Optimality of One-Group Shrinkage Priors in Sparse High-dimensional Problems. *Bayesian Analysis*, 2016a.

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

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

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

P. Richard Hahn, Jingyu He, and Hedibert Lopes. Elliptical slice sampling for Bayesian shrinkage regression with applications to causal inference. 2016. URL http://faculty.chicagobooth.edu/richard.hahn/JCGS_submit.pdf.

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2nd edition, 2009. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

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

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1): 55–67, 1970. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

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

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

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

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

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

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

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106:1125–1138, 2012.

T. J. Mitchell and J. J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, December 1988. ISSN 01621459. URL http://dx.doi.org/10.2307/2290129.

Tomi Peltola, Aki S Havulinna, Veikko Salomaa, and Aki Vehtari. 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, 2014.

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

Nicholas G Polson and James G Scott. Large-scale simultaneous testing with hypergeometric inverted-beta priors. *arXiv preprint arXiv:1010.5223*, 2010a.

Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010b.

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

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

Veronika Rcková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, (just-accepted), 2016.

James G. Scott. Parameter expansion in local-shrinkage models. *arXiv preprint arXiv:1010.5265*, 2010. URL http://arxiv.org/abs/1010.5265. bibtex: scott_parameter_2010.

Charles Stein. 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, 1956. URL http://www.stat.yale.edu/~hz68/619/Stein-1956.pdf.

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

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

Alexander Terenin, Shawfeng Dong, and David Draper. GPU-accelerated Gibbs Sampling. *arXiv:1608.04329 [cs, stat]*, August 2016. URL http://arxiv.org/abs/1608.04329.

George C Tiao and WY Tan. Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance-components. *Biometrika*, pages 37–53, 1965.

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

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

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

Ryan J Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011a.

Ryan J Tibshirani, Jonathan Taylor, and others. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011b.

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

SL van der Pas, BJK Kleijn, and AW van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618, 2014.

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

Stephanie van der Pas, James Scott, Antik Chakraborty, and Anirban Bhattacharya. Horseshoe: Implementation of the Horseshoe Prior. November 2016. URL https://cran.r-project.org/web/packages/horseshoe/index.html.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. How many needles in the haystack? Adaptive inference and uncertainty quantification for the horseshoe. *arXiv:1607.01892*, 2016.

Stéphanie van der Pas, Botond Szabó, van der Vaart, and Aad. Adaptive posterior contraction rates for the horseshoe. *arXiv:1702.03698*, 2017.

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

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

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

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

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