Horseshoe Regularization for Feature Subset Selection

Anindya Bhadra [1] Jyotishka Datta [2] Nicholas G. Polson [3] and Brandon Willard [3]

Abstract

Feature subset selection arises in many high-dimensional applications of statistics, such as compressed sensing and genomics. The ℓ_0 penalty is ideal for this task, the caveat being it requires the NP-hard combinatorial evaluation of all models. A recent area of considerable interest is to develop efficient algorithms to fit models with a non-convex ℓ_γ penalty for $\gamma \in (0,1)$, which results in sparser models than the convex ℓ_1 or lasso penalty, but is harder to fit. We propose an alternative, termed the horseshoe regularization penalty for feature subset selection, and demonstrate its theoretical and computational advantages. The distinguishing feature from existing non-convex optimization approaches is a full probabilistic representation of the penalty as the negative of the logarithm of a suitable prior, which in turn enables efficient expectation-maximization and local linear approximation algorithms for optimization and MCMC for uncertainty quantification. In synthetic and real data, the resulting algorithms provide better statistical performance, and the computation requires a fraction of time of state-of-the-art non-convex solvers.

Keywords: Bayesian methods; feature selection; horseshoe estimator; non-convex regularization; scale mixtures.

1 Introduction

Feature subset selection is typically performed by convex penalties such as the lasso (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005), or their variants. Convex penalties enjoy a number of advantages, such as uniqueness of solution, efficient computation and relatively straightforward theoretical analysis. Convex penalties, however, suffer from some undesirable features. For example, the lasso, which is based on a soft thresholding operation, leaves a constant bias that does not go to zero for large signals. A consequence is poor mean squared error in estimation. The lasso also suffers from problems in presence of correlated variables. Non-convex penalties, on the other hand, can result in optimal theoretical performances for variable selection (Fan and Li, 2001). However, the computational burden of fitting non-convex penalties is more challenging. In this article, we take a Bayesian view of the optimization problem as finding the posterior mode under a given prior. Our approach is probabilistic, which enables a latent variable representation and results in efficient expectation-maximization (Dempster et al., 1977) and local linear approximation (Zou and Li, 2008) algorithms for optimization, as well as a Markov chain Monte Carlo (MCMC) scheme for posterior simulation. The performance comparison in simulations reveals the proposed regularization provides better statistical performance, while allowing much faster computation compared to state-of-the-art non-convex solvers.

Is this characteristic of local optimization (looking for mode instead of averaging over the whole posterior?)

[1] *Address:* Department of Statistics, Purdue University, 250 N. University St., West Lafayette, IN 47907, USA.
[2] *Address:* Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA.
[3] *Address:* Booth School of Business, The University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637, USA.

1.1 Related Works in Non-Convex Regularization

Consider the sparse normal means model where we observe $(y_i \mid \theta_i) \overset{ind}{\sim} \mathcal{N}(\theta_i, 1)$ for $i = 1, \ldots, n$, where $\#(\theta_i \neq 0) \leq p_n$ and $p_n = o(n)$ as $n \to \infty$. Non-convex regularization problems arise from a need to correctly identify the zero components in $\theta = (\theta_1, \ldots, \theta_n)$, given observations $y = (y_1, \ldots, y_n)$, also known as subset selection. The ℓ_0 penalty, defined as $||\theta||_0 = \sum_{i=1}^n 1(|\theta_i| > 0)$, is ideal for this task, and the more commonly used lasso or convex ℓ_1 penalty, $||\theta||_1 = \sum_{i=1}^n |\theta_i|$, tends to select a denser model (Mazumder et al., 2012). Unfortunately, naïvely using the ℓ_0 penalty requires a combinatorial evaluation of all 2^n models, which is NP-hard (Natarajan, 1995). Penalties of the form ℓ_γ for $\gamma \geq 1$ give rise to convex problems and efficient solvers are available. It remains a challenge to fit models with ℓ_γ penalties for $\gamma \in (0, 1)$. While this does not necessarily a present combinatorial problem, the regularization problem is non-convex. Thus, the general purpose tools for convex optimization do not apply, nor is a unique solution guaranteed (see, e.g., Boyd and Vandenberghe, 2004, Chapter 1). Non-convex penalties include the smoothly clipped absolute deviation or SCAD (Fan and Li, 2001) and the minimax concave penalty or MCP (Zhang, 2010). Recent computational advances in fitting models with non-convex penalties include Breheny and Huang (2011) and Mazumder et al. (2012). Both works use coordinate descent approaches to fit SCAD and MCP and provide conditions for convergence. Alternatively, an overview of proximal algorithms for non-convex optimization is given by Parikh and Boyd (2014) and Polson et al. (2015). Recent works have also demonstrated the equivalence between fitting a model with MCP penalty and evaluating the posterior mode in a Bayesian hierarchical model under a suitable prior (Schifano et al., 2010; Strawderman et al., 2013). Following along these lines, we show that evaluating the posterior mode under a suitable approximation to the horseshoe prior of Carvalho et al. (2009, 2010) solves a non-convex optimization problem with desirable theoretical properties and derive fast computational algorithms.

2 The Horseshoe Prior and Penalty

Many penalized optimization problems in statistics take the form

$$\underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \{l(\theta; y) + \pi(\theta)\}, \tag{1}$$

where $l(\theta; y)$ is a measure of fit of parameter θ to data y (also known as the empirical risk), and $\pi(\theta)$ is a penalty function. Let $p(y \mid \theta) \propto \exp\{-l(\theta; y)\}$ and $p(\theta) \propto \exp\{-\pi(\theta)\}$, where p denotes a generic density. If $l(\theta; y)$ is proportional to the negative of the log likelihood function under a suitable model, one arrives at a Bayesian interpretation to the optimization problem: finding the mode of the posterior density $p(\theta \mid y)$ under prior density $p(\theta)$ (Polson and Scott, 2016). The properties of the penalty are then induced by those of the prior. The horseshoe prior (Carvalho et al., 2010) is defined as global-local Gaussian scale mixture under a half-Cauchy prior, with density

$$p_{HS}(\theta_i \mid \tau) = \int_0^\infty \frac{1}{u_i \tau} \phi\left(\frac{\theta_i}{u_i \tau}\right) \frac{2}{\pi(1 + u_i^2)} du_i, \tag{2}$$

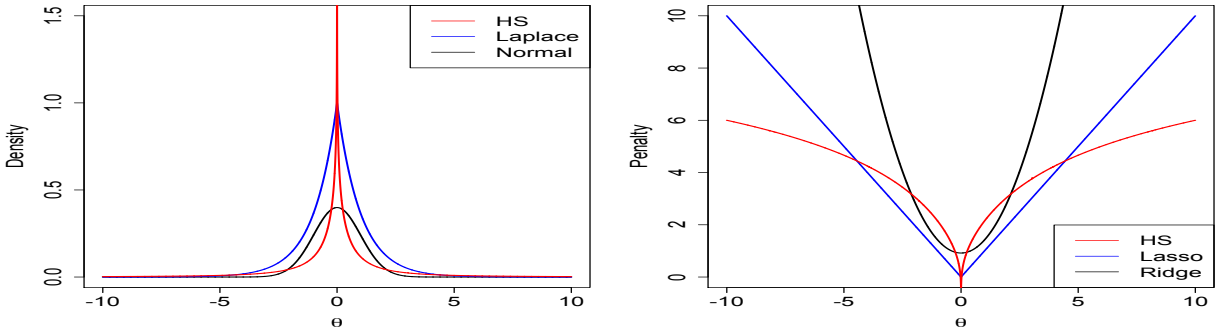

Figure 1: Some densities and penalties given by the negative of their logarithms. Left panel: the horseshoe (HS) with $\tau = 1$, the standard Laplace and standard normal densities. Right panel: the corresponding non-convex horseshoe and the convex lasso and ridge penalties.

where $\tau > 0$ and $\phi(\cdot)$ denotes the standard normal density. Equivalently,

$$\theta_i \mid u_i, \tau \overset{ind}{\sim} \mathcal{N}(0, u_i^2\tau^2), \quad u_i \mid \tau \overset{ind}{\sim} C^+(0,1), \quad \tau > 0,$$

where C^+ denotes a half-Cauchy random variable. The u_i are local shrinkage parameters which shrink irrelevant signals to zero while keeping the magnitude of true signals. The parameter τ is a global shrinkage parameter which gauges the level of sparsity. Several optimality results are available when the posterior mean under the horseshoe prior is used as an estimator, such as minimax optimality in estimation under ℓ_2 loss (van der Pas et al., 2014) and asymptotic optimality in testing under 0–1 loss (Datta and Ghosh, 2013). However, little is known of the properties of the posterior mode under the horseshoe prior, which amounts to a solution of (1) under the horseshoe penalty using a squared error empirical risk, given by $\sum_{i=1}^n (\theta_i - y_i)^2$.

Carvalho et al. (2010) show that the horseshoe prior density admits tight upper and lower bounds

$$\frac{\log\left(1+\frac{4\tau^2}{\theta_i^2}\right)}{\tau(2\pi)^{3/2}} < p_{HS}(\theta_i \mid \tau) < \frac{2\log\left(1+\frac{2\tau^2}{\theta_i^2}\right)}{\tau(2\pi)^{3/2}}, \tag{3}$$

for $\theta_i \in \mathbb{R}$, $\tau > 0$ and prove $\lim_{|\theta_i|\to 0} p(\theta_i) = \infty$ and $\lim_{|\theta_i|\to\infty} p(\theta_i) \sim (\theta_i)^{-2}$ for any fixed τ. The corresponding penalty is $\pi_{HS}(\theta \mid \tau) = \sum_{i=1}^n \pi_{HS}(\theta_i \mid \tau)$, where,

$$\pi_{HS}(\theta_i \mid \tau) = -\log p_{HS}(\theta_i \mid \tau) = -\log\log\left(1+\frac{2\tau^2}{\theta_i^2}\right), \tag{4}$$

up to terms involving θ_i. Since the horseshoe prior density has tails decaying as θ_i^{-2}, the corresponding penalty behaves as the logarithmic penalty for large values of $|\theta_i|$ and can be seen to be non-convex. Figure 1 right panel shows the non-convex horseshoe penalty, in combination with convex lasso and ridge penalties. They are respectively obtained by taking the negative of the logarithm of the horseshoe, the Laplace and normal densities, shown on the left panel. It can be seen that the horseshoe penalty is more aggressive near zero compared to the convex penalties, en-

couraging sparsity. In fact, both the density and penalty are unbounded for the horseshoe at zero, suggesting a global solution to the optimization problem in (1) that is identically equal to zero. However, this does not preclude the possibility of other local solutions, and in fact encourages one to look for local optimization algorithms that might lead to solutions that are more interesting than all zeros. For values far away from zero, the horseshoe penalizes lightly, a fact that can also be attributed to the heavy tails of the local u_i terms. This suggests the horseshoe penalty bridges the gap between ℓ_0 and ℓ_1 penalties, a property also shared by other non-convex penalties such as SCAD or MCP.

2.1 Properties of the Horseshoe Penalty

Fan and Li (2001) describe the desirable properties for a penalty function and list conditions for checking whether those properties hold. These are as follows:

1. *(Near) Unbiasedness.* The resultant estimator is (nearly) unbiased when the true parameter is large. A sufficient condition is that the penalty satisfies $\pi'(|\theta|) = 0$ when $|\theta| \to \infty$ where π' is the first derivative of the penalty π.

2. *Sparsity.* The resultant estimator is sparse. A sufficient condition is that $\inf_\theta\{|\theta| + \pi'(|\theta|)\} > 0$.

3. *Continuity.* The resultant estimator is continuous in the data y to encourage stability in prediction. A necessary and sufficient condition is that $\text{argmin}_\theta\{|\theta| + \pi'(|\theta|)\} = 0$.

Property (3) is violated by hard thresholding rules, whereas Property (1) is violated by the lasso and associated soft thresholding rules and also by all penalties of the form ℓ_γ for $\gamma > 1$. Penalties that satisfy Properties (1)–(3) include MCP and SCAD, however the computational algorithms used to fit models employing these penalties are quite challenging and can suffer from numerical issues. We first show that the horseshoe penalty enjoys Properties (1)–(2) above, arguing for its theoretical advantage; before preceding to develop efficient computational algorithms.

PROPOSITION 2.1. *The horseshoe posterior mode, defined as* $\text{argmin}_\theta\{(\theta - y)^2/2 + \pi_{HS}(\theta)\}$, *where* $\pi_{HS}(\theta)$ *denotes the horseshoe penalty, satisfies Properties (1)–(2) above, but not Property (3).*

Proof. It is enough to check the properties for a single coordinate θ_i. First, note from (4) that

$$\pi'_{HS}(|\theta_i|) = \frac{4\tau^2/|\theta_i|^3}{(1 + 2\tau^2/\theta_i^2)\log(1 + 2\tau^2/\theta_i^2)},$$

and Property (1) follows. Next, for Property (2),

$$|\theta_i| + \pi'_{HS}(|\theta_i|) \quad = \quad |\theta_i| + \frac{4\tau^2/|\theta_i|^3}{(1 + 2\tau^2/\theta_i^2)\log(1 + 2\tau^2/\theta_i^2)}.$$

Thus, $|\theta_i| + \pi'_{HS}(|\theta_i|) \to \infty$ as $|\theta_i| \to 0, \infty$ for any given $\tau > 0$. For $|\theta_i| \neq 0, \infty$, the denominator of the second term is strictly positive. Thus, we need to check

$$\theta_i^4(1 + 2\tau^2/\theta_i^2)\log(1 + 2\tau^2/\theta_i^2) + 4\tau^2 > 0.$$

Result 4.1.33 of Abramowitz and Stegun (1965) gives

$$x < (1+x)\log(1+x), \quad x > -1, x \neq 0.$$

Using $x = 2\tau^2/\theta_i^2$ yields

$$\theta_i^4(1+2\tau^2/\theta_i^2)\log\left(1+2\tau^2/\theta_i^2\right) + 4\tau^2 > 2\tau^2\theta_i^2 + 4\tau^2,$$

which is strictly positive for any $\tau > 0$, proving Property (2) holds. Since $\lim_{|\theta_i|\to 0}\{|\theta_i| + \pi'_{HS}(|\theta_i|)\} = \infty$, Property (3) fails to hold. □

An implication of this result is that the resultant estimator is sparse and is nearly unbiased in estimating large signals. However, the lack of continuity means the estimator suffers from some of the same issues as hard thresholding. We verify the hard thresholding-like behavior of the estimator via simulations and show that if the posterior mean is used as an estimator rather than the posterior mode, then it solves the continuity problem and usually results in an estimator with better squared error loss. However, the posterior mean does not result in a sparse solution and hence, is not suitable for subset selection.

3 The Horseshoe-Like Prior and Its Scale Mixture Representation

There is no closed form for the horseshoe density and numerical integration over u_i in (2) is required to evaluate the density at any given θ_i. The tight upper and lower bounds in (3) are also not densities. However, a proper prior density that mimics the behavior of the horseshoe density with a pole at the origin and polynomial tails is given by

$$p_{\widetilde{HS}}(\theta_i \mid a) = \frac{1}{2\pi a^{1/2}}\log\left(1+\frac{a}{\theta_i^2}\right), \tag{5}$$

for $\theta_i \in \mathbb{R}$, $a > 0$. We call this the horseshoe-like prior. Setting $a = 2\tau^2$ and $a = 4\tau^2$ in (5) one recovers the bounds in (3) that differ only by a constant factor. Since the bounds in (3) are tight in θ_i, and a constant multiplicative factor of the density (or, equivalently, a constant additive term to the penalty) has no bearings on the solutions to the optimization problem, one can use $\pi_{\widetilde{HS}}(\theta_i) = -\log(p_{\widetilde{HS}}(\theta_i))$ as a useful surrogate of the horseshoe penalty. The chief advantage of using a proper density is that it enables one to use the technique of latent variables to solve the optimization problem, such as the EM algorithm or the techniques based on data augmentation (Tanner and Wong, 1987), provided one can find a suitable probabilistic representation.

The methodology developed in the remainder of this article relies on the following key result. For a real valued function $f(\cdot)$, the Frullani integral identity (Jeffreys and Swirles, 1972, pp. 406–407) gives

$$\int_0^\infty \frac{f(cx) - f(dx)}{x}dx = \{f(0) - f(\infty)\}\log(d/c),$$

for $c > 0, d > 0$. Using $f(x) = \exp(-x)$ yields a latent variable representation for the global-local

scale mixture for $p_{\widetilde{HS}}(\theta_i \mid a)$ in (5) as:

$$\begin{aligned}\frac{1}{2\pi a^{1/2}}\log\left(1+\frac{a}{\theta_i^2}\right) &= \int_0^\infty \exp\left(-\frac{u_i\theta_i^2}{a}\right)\frac{(1-e^{-u_i})}{2\pi a^{1/2}u_i}du_i \\ &= \int_0^\infty \left(\frac{u_i}{a\pi}\right)^{1/2}\exp\left(-\frac{u_i\theta_i^2}{a}\right)\frac{(1-e^{-u_i})}{2\pi^{1/2}u_i^{3/2}}du_i, \quad a>0.\end{aligned}$$

or equivalently,

$$(\theta_i \mid u_i, a) \overset{ind}{\sim} \mathcal{N}\left(0, \frac{a}{2u_i}\right),\ p(u_i) = \frac{1-e^{-u_i}}{2\pi^{1/2}u_i^{3/2}}, \tag{6}$$

for $0 < u_i < \infty,\ a > 0$. Once again, the u_i terms act as local scale parameters and the global term a controls the overall level of sparsity. A useful outcome of this probabilistic representation is that the u_i terms can be viewed as latent variables and thus suggests the possibility of EM and MCMC schemes, provided the appropriate quantities in the posterior can be easily computed.

3.1 Alternative Scale Mixtures and the Marginal Density Under the Horseshoe-Like Prior

The horseshoe-like prior density on θ_i, given by (5), can be represented as a scale mixture of both Cauchy and Laplace densities on θ_i, as the following two lemmas show.

LEMMA 3.1. *For a fixed τ, the horseshoe-like prior can be written as a uniform scale mixture of a Cauchy prior on θ_i, i.e. $\theta_i \mid \lambda_i, \tau \sim \mathcal{C}(0, \lambda_i\tau)$ and $\lambda_i \sim \mathcal{U}(0,1)$.*

$$p_{\widetilde{HS}}(\theta_i \mid \tau^2) = \frac{1}{2\pi\tau}\log\left(1+\frac{\tau^2}{\theta_i^2}\right) = \frac{1}{\pi}\int_0^1 \frac{\lambda_i\tau}{\lambda_i^2\tau^2+\theta_i^2}d\lambda_i. \tag{7}$$

The proof is elementary and therefore omitted. The Cauchy scale mixture representation provides a natural adaptive sparsity model for the horseshoe-like prior. The horseshoe-like prior can be also expressed as a mixture of Laplace densities on θ_i due to a result by Steutel and Van Harn (2003).

LEMMA 3.2. *The horseshoe-like prior density in (5) can be written as a scale mixture of a double exponential or Laplace prior on θ_i, as given below:*

$$\begin{aligned}p_{\widetilde{HS}}(\theta_i \mid \tau^2) = \frac{1}{2\pi\tau}\log\left(1+\frac{\tau^2}{\theta_i^2}\right) &= \frac{1}{2\tau}\int_0^\infty \lambda_i \exp\{-\lambda_i|\theta_i|/\tau\}h(\lambda_i)d\lambda_i \\ \text{where } h(\lambda) = \frac{2}{\pi}\left(\frac{1-\cos(\lambda)}{\lambda^2}\right) &= \frac{1}{2\pi}\left(\frac{\sin(\lambda/2)}{\lambda/2}\right)^2,\ 0 \le \lambda < \infty.\end{aligned} \tag{8}$$

Here the mixing density on λ_i is a special type of density arising from Polya characteristic functions, called the Fejer-de la Vallee Poussin (or FVP) density (Devroye, 1986, Theorem 6.9).

A useful outcome of Lemma 3.1 is the following result for the marginal density on y_i under an Inverse-Gamma$(1/2, 1/2)$ prior on σ^2. We can use a Cauchy convolution result (Bhadra et al., 2016b) to prove the following:

PROPOSITION 3.1. *Let the observations* $(y_i \mid \theta_i, \sigma^2) \sim \mathcal{N}(\theta_i, \sigma^2)$ *and* $\sigma^2 \sim$ *Inverse-Gamma*$(1/2, 1/2)$, *where the* θ_i*'s are given the horseshoe-like prior in* (5), *i.e.* $p(\theta_i \mid \tau) = \frac{1}{2\pi\tau} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right)$. *Then the marginal density of* y_i *is given by:*

$$m(y_i \mid \tau) = \frac{1}{2\pi\tau} \log\left(1 + \frac{\tau^2}{1 + y_i^2}\right). \tag{9}$$

A proof is given in Appendix A. A consequence is that the marginal density $m(y_i \mid \tau)$ behaves similar to the prior density $p(\theta_i \mid \tau)$ for large values of $|y_i|$ and thus also displays heavy tails.

Implications of the Laplace scale mixture in Lemma 3.2 are discussed in Section 4.3, where it is used to derive a local linear approximation (LLA) algorithm.

4 Computational Algorithm I: Algorithms for Feature Selection

We derive fast computational algorithms for evaluating the maximum a-posteriori (MAP) estimate under the horseshoe-like prior. According to Sections 2 and 3, the solution to the this problem is identical to that of the optimization problem in (1) under the horseshoe penalty $\pi_{HS}(\theta)$, where the empirical risk measure is the squared error loss. The proposed technique uses the latent variable representation in (6) to derive an EM algorithm for MAP estimation.

4.1 EM for Subset Selection in Normal Means Model

First, consider the model: $(y_i \mid \theta_i) \overset{ind}{\sim} \mathcal{N}(0, 1)$. From (6), the hierarchical model for $i = 1, \ldots, n$ is

$$(y_i \mid \theta_i) \overset{ind}{\sim} \mathcal{N}(\theta_i, 1),\ (\theta_i \mid u_i, a) \overset{ind}{\sim} \mathcal{N}\left(0, \frac{a}{2u_i}\right),\ p(u_i) = \frac{1 - e^{-u_i}}{2\pi^{1/2} u_i^{3/2}},\ 0 < u_i < \infty, \quad a > 0.$$

The complete data posterior is

$$p(\theta_i, u_i \mid y_i, a) \quad \propto \quad \exp\left\{-\frac{(y_i - \theta_i)^2}{2}\right\} \exp\left(-\frac{u_i \theta_i^2}{a}\right) \frac{(1 - e^{-u_i})}{u_i}.$$

If one views the u_i terms as latent variables, the E-step consists of computing their posterior expectations. It is given by $\tilde{u}_i = E(u_i \mid \theta_i, y_i, a)$, where,

$$\tilde{u}_i = \frac{1}{2\pi a^{1/2}} \int_0^\infty u_i \exp\left(-\frac{u_i \theta_i^2}{a}\right) \frac{(1 - e^{-u_i})}{u_i} du_i = \frac{1}{2\pi a^{1/2}} \left(\frac{a}{\theta_i^2} - \frac{a}{\theta_i^2 + a}\right).$$

The M-step maximizes the complete data posterior jointly in (θ, a) with the u_i terms replaced by $\tilde{u}_i$. While the joint maximization does not have a closed-form solution, the conditional maximizations $(\theta \mid a)$ and $(a \mid \theta)$ are simple. The optimal θ_i for a given a is simply the Gaussian posterior mode,

$$\hat{\theta}_i \mid a = \left(1 + \frac{2\tilde{u}_i}{a}\right)^{-1} y_i.$$

Maximization of a with given θ is easy due to the fact that

$$\theta_i \sqrt{2u_i} \mid a \sim \mathcal{N}(0, a).$$

Thus,

$$\hat{a} \mid \theta = \frac{1}{n} \sum_{i=1}^{n} 2\tilde{u}_i \theta_i^2 = \frac{a^{3/2}}{n\pi} \sum_{i=1}^{n} \frac{1}{\theta_i^2 + a}.$$

Thus, the $(t+1)^{\text{th}}$ expectation-maximization recursion for $t \geq 0$ is given by as a coordinate descent, or as expectation-conditional maximization (Meng and Rubin, 1993), as

$$\hat{a}^{(t+1)} \mid \hat{\theta}_1^{(t)}, \ldots, \hat{\theta}_n^{(t)} = \frac{\{\hat{a}^{(t)}\}^{3/2}}{n\pi} \sum_{i=1}^{n} \left(\frac{1}{\{\hat{\theta}_i^{(t)}\}^2 + \hat{a}^{(t)}} \right),$$

$$\hat{\theta}_i^{(t+1)} \mid \hat{a}^{(t+1)} = y_i \left(1 + \frac{\{\hat{a}^{(t+1)}\}^{1/2}}{\pi \{\hat{\theta}_i^{(t)}\}^2 \left[\{\hat{\theta}_i^{(t)}\}^2 + \hat{a}^{(t+1)} \right]} \right)^{-1},$$

for $i = 1, \ldots, n$, which is repeated until convergence and $\hat{\theta}^{(0)}$ and $\hat{a}^{(0)}$ are suitable initial values. Since the penalty is unbounded at zero, the global solution to the optimization problem is given by $\hat{\theta}_i = 0$ for all i. However, since the EM is a local, deterministic algorithm, it converges once a local mode is identified. In fact, the existence of a global mode identically equal to zero provides arguments against using a global optimization algorithm, such as simulated annealing (Kirkpatrick et al., 1983). The convergence of the EM algorithm of course depends on the choice of starting values. However, the fact that there is no unique solution is a result of the non-convex penalty itself, rather than an artifact caused by a failure of the optimization algorithm. Local solutions can be compared by evaluating the likelihoods at the solutions, or by their squared error estimates. If the algorithm converges to the uninteresting all zero solution, it can be restarted with a different choice of starting values.

4.2 EM for Subset Selection in High-Dimensional Regression

A similar computational algorithm is also applicable to feature selection in high-dimensional regression. Consider the following regression model for $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \theta \in \mathbb{R}^p$ where $p > n$:

$$(y \mid X, \theta) \overset{ind}{\sim} \mathcal{N}(X\theta, 1), \ (\theta_i \mid u_i, a) \overset{ind}{\sim} \mathcal{N}\left(0, \frac{a}{2u_i}\right), \ p(u_i) = \frac{1 - e^{-u_i}}{2\pi^{1/2} u_i^{3/2}}, \ 0 < u_i < \infty, \quad a > 0.$$

The normal means model $(y_i \mid \theta_i) \overset{ind}{\sim} \mathcal{N}(\theta_i, 1)$ of Section 4.1 can be seen to be a special case of the regression model with $p = n$ and $X = I_n$, where I_n is the identity matrix of size n. Since there is no change in the hierarchy compared to the normal means model for the latent u_i terms, their conditional expectations remain unchanged. Similarly, the posterior mode of a has the same form as Section 4.1. The only change is that the posterior mode for θ is now given by

$$\hat{\theta} \mid a = \left\{ X^T X + \text{diag}\left(\frac{2\tilde{u}_i}{a} \right) \right\}^{-1} X^T y.$$

Consequently, the $(t+1)^{\text{th}}$ EM iteration for $t \geq 0$ is:

$$\hat{a}^{(t+1)} \mid \hat{\theta}_1^{(t)}, \ldots, \hat{\theta}_p^{(t)} = \frac{\{\hat{a}^{(t)}\}^{3/2}}{p\pi} \sum_{i=1}^{p} \left(\frac{1}{\{\hat{\theta}_i^{(t)}\}^2 + \hat{a}^{(t)}} \right),$$

$$\hat{\theta}^{(t+1)} \mid \hat{a}^{(t+1)} = \left\{ X^T X + \text{diag} \left(\frac{2\tilde{u}_i^{(t)}}{\hat{a}^{(t+1)}} \right) \right\}^{-1} X^T y,$$

where, as in Section 4.1,

$$\tilde{u}_i^{(t)} = \frac{1}{2\pi a^{1/2}} \left(\frac{a}{\theta_i^2} - \frac{a}{\theta_i^2 + a} \right),$$

computed at $a = \hat{a}^{(t)}, \theta_i = \hat{\theta}_i^{(t)}$. The computationally limiting step is the calculation of the inverse of the $p \times p$ matrix of the form $(X^T X + D^{-1})^{-1}$ where D^{-1} is a $p \times p$ positive definite diagonal matrix, which in our case is diag $(2\tilde{u}_i/a)$. The naive computational complexity is $O(p^3)$. However, an application of the Woodbury matrix identity gives

$$(X^T X + D^{-1})^{-1} = D - DX^T (XDX^T + I_n)^{-1} XD.$$

This involves the computing the inverse of an $n \times n$ matrix, which is $O(n^3)$, and the computation of matrix products DX^T and $X^T D$, which are $O(np^2)$. Thus, the resultant computational complexity is $O(np^2)$ when $p > n$, which is an improvement over $O(p^3)$.

4.3 One-step Estimator Using the LLA Algorithm

We now discuss the implications of the horseshoe-like prior as a Laplace scale mixture (see Lemma 3.2), and show that it is useful for sparse parameter learning via the local linear approximation (LLA) algorithm of Zou and Li (2008) that improves upon the local quadratic approximation (LQA) of Fan and Li (2001). In particular, Zou and Li (2008) provided an EM algorithm and an optimal one-step estimator by using an inverse Laplace transform on the bridge penalty, which is equivalent to a Laplace mixture of a stable law. In general, any sparsity-inducing prior that admits a Laplace mixture representation falls into the LLA–LQA framework, a notable example being the generalized double Pareto prior (Armagan et al., 2011).

Fan and Li (2001); Hunter and Li (2005) and Zou and Li (2008) discuss LQA and LLA algorithms. Hunter and Li (2005) discuss the relationship of the LQA and minorize-majorize (MM) algorithms which are extensions of the EM algorithm. When penalties can be written as a cumulant transformation, equivalently a scale mixture of normals, these algorithms are exact. Polson and Scott (2016) discuss the duality between mixture and envelope representation from a Bayesian perspective of hierarchical modeling and present several useful conditions for such duality to hold.

We discuss these strategies for the horseshoe-like prior after a brief description of the framework in the context of a penalized likelihood model. Specifically, consider the regularization problem

$$Q(\theta) = \underset{\theta \in \mathbb{R}^p}{\text{argmax}} \left\{ \sum_{i=1}^{n} l_i(\theta) - n \sum_{j=1}^{p} \pi_\tau \left(|\theta_j| \right) \right\},$$

where $l_i(\theta)$ is the log likelihood of the ith observation, n is the number of observations, p is the model space dimension and π_τ is the penalty applied to each coefficient, although in principle they could be component-specific. The LQA algorithm uses a quadratic Taylor approximation for $\pi_\tau\left(|\theta_j|\right)$ whereas the LLA algorithm (Zou and Li, 2008, Equation (2.6)) uses

$$\pi_\tau\left(|\theta_j|\right) \approx \pi_\tau\left(|\theta_j^{(0)}|\right) + \pi'_\tau\left(|\theta_j^{(0)}|\right)\left(|\theta_j| - |\theta_j^{(0)}|\right), \text{ for } \theta_j \approx \theta_j^{(0)}.$$

Hence, LLA leads to the following iterative algorithm that can be solved with the LARS algorithm (Efron et al., 2004) for LASSO. Set the initial value $\theta_j^{(0)}$ to be the un-penalized maximum likelihood estimate. For each $k = 1, 2, \ldots$, solve the iterative system of equations until convergence of the $\{\theta^{(k)}\}$ sequence.

$$\theta^{(k+1)} = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} l_i(\theta) - n\sum_{j=1}^{p} \pi'_\tau\left(|\theta_j^{(k)}|\right)|\theta_j| \right\}. \tag{10}$$

This scheme is called the LLA algorithm of Zou and Li (2008), which has a unque advantage of producing sparse intermediate and final estimates $\theta^{(k)}$ unlike the LQA algorithm. Zou and Li (2008) also showed that the LLA algorithm can be recast as an EM algorithm under certain conditions. Suppose the exponentiated (negative) penalty function $\exp(-n\pi_\tau(\cdot))$ admits the following Laplace mixture representation:

$$\exp(-n\pi_\tau(|\theta_j|)) = \int_0^\infty \frac{1}{2\omega_j} \mathrm{e}^{-|\theta_j|/\omega_j} p(\omega_j) d\omega_j. \tag{11}$$

Then, maximizing $Q(\boldsymbol{\theta})$ becomes equivalent to calculating the posterior mode of $p(\boldsymbol{\theta} \mid y)$ by treating $\exp(-n\pi_\tau(|\theta_j|))$ as the prior on $\boldsymbol{\theta}$ after marginalizing the hyperparameters. This property holds true for the penalty induced by the horseshoe-like prior as it satisfies the Laplace mixture representation (*vide* Lemma 3.2). For the general prior-penalty in (11), the exact EM step for LLA algorithm is given by (Zou and Li, 2008, Equation (2.13)):

$$\theta^{(k+1)} = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmax}} \left[\sum_{i=1}^{n} l_i(\theta) + \sum_{j=1}^{p} \left\{ -|\theta_j| \, \mathbb{E}\left(\omega_j^{-1} \mid \theta_j^{(k)}, y\right) \right\} \right], \; k = 1, 2, \ldots.$$

As the posterior moment comes from a scale mixture, the expectation can be derived without an explicit knowledge of the mixing measure. A computationally efficient alternative to the aforesaid EM procedure is the one-step estimator $\hat{\theta}_{ose}$ proposed by Zou and Li (2008), that automatically incorporates sparsity. For the linear regression model, taking $\theta^{(0)}$ to be the ordinary least squares estimator, we get:

$$\theta^{(1)}_{\text{lin-reg}} = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}||y - X\theta||^2 + n\sum_{j=1}^{p} \pi'_\tau\left(|\theta_j^{(0)}|\right)|\theta_j| \right\},$$

and for a general likelihood model, assuming $\theta^{(0)} = \hat{\theta}(\text{mle})$, the corresponding one-step estimators

are given as:

$$\theta^{(1)}_{\text{log-lik}} = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2}(\theta - \theta^{(0)})'[-\nabla^2 \ell(\theta^{(0)})](\theta - \theta^{(0)}) + n \sum_{j=1}^{p} \pi'_\tau \left(|\theta_j^{(0)}| \right) |\theta_j| \right\},$$

For the horseshoe-like prior, $\pi'_\tau \left(|\theta_j^{(0)}| \right)$ is given as:

$$\pi'_\tau \left(|\theta_j| \right) = \frac{4\tau^2/|\theta_j|^3}{\left(1 + 2\tau^2/\theta_j^2\right) \log\left(1 + 2\tau^2/\theta_j^2\right)},$$

Hence, the one-step estimator for the horseshoe-like prior for the normal means problem can be written as:

$$\theta^{(1)} = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} ||y - \theta||^2 + 4\tau^2 n \sum_{j=1}^{p} \frac{|\theta_j|}{|\theta^{(0)}|^3 (1 + 2\tau^2/(\theta_j^{(0)})^2) \log(1 + 2\tau^2/(\theta_j^{(0)})^2)} \right\}. \tag{12}$$

The one-step estimator in (12) has a superficial similarity with the adaptive LASSO (Zou, 2006) in that the weights of $|\theta_j|$ are decreasing function of $\hat{\theta}(ols)$. The one-step estimator can be rapidly computed by exploiting the LARS algorithm (Efron et al., 2004).

5 Computational Algorithm II: MCMC for Posterior Exploration

In addition to fast EM and LLA algorithms for MAP estimates, one may wish to explore the entire posterior for a full Bayes solution and uncertainty quantification. The hierarchy for the horseshoe-like prior in (6) can be reparameterized by taking $t_i^2 = 2u_i$ and $\tau^2 = a$ to yield the following:

$$(\theta_i \mid t_i, \tau) \sim \mathcal{N}\left(0, \frac{\tau^2}{t_i^2}\right), \; p(t_i) = \frac{(1 - e^{-\frac{1}{2}t_i^2})}{\sqrt{2\pi} t_i^2}, \tag{13}$$

where $t_i \in \mathbb{R}, \tau^2 > 0$ and the prior density $p(t_i)$ in (13) is known as the standard slash-normal ($SN(0,1)$) density, given by:

$$p_{SN}(x) = \frac{\phi(0) - \phi(x)}{x^2} = \frac{1 - e^{-\frac{1}{2}x^2}}{\sqrt{2\pi} x^2}, \quad x \in \mathbb{R},$$

where $\phi(\cdot)$ is the density of a standard normal. The $SN(0,1)$ density can be also written as a normal variance mixture with a Pareto$(1/2)$ mixing density (Barndorff-Nielsen et al., 1982; Gneiting, 1997). The following result provides a scale mixture representation for the type II modulated normal density, which reduces to the slash-normal density for $b = 1/2$.

PROPOSITION 5.1. *(Gneiting, 1997). Suppose $p(x)$ is a scale mixture of normal with density*

$$p(x) = \int_0^\infty \frac{1}{(2\pi v)^{\frac{1}{2}}} \exp\left(-\frac{x^2}{2v}\right) dF(v),$$

where F is a distribution function on $[0,\infty]$. The modulated normal distributions of type II arise when $F(\cdot)$ is a Pareto distribution on $[1,\infty)$ with parameter $b>0$. The Pareto distribution has density b/ν^{b+1} for $\nu>1$, and the resulting normal scale mixture has density:

$$p(x)=\frac{b}{(2\pi)^{1/2}}\left(\frac{x^2}{2}\right)^{-(b+\frac{1}{2})}\gamma\left(b+\frac{1}{2},\frac{x^2}{2}\right).$$

Here $\gamma(\alpha,x)=\int_0^x t^{\alpha-1}\mathrm{e}^t dt$ denotes the lower incomplete gamma function.

Hence, the following lemma is immediate.

LEMMA 5.1 (Hierarchy for slash-normal)**.** *Slash-normal random variables can be generated as the $X=ZV^{\frac{1}{2}}$, where Z is a standard normal and V follows a Pareto distribution on $[1,\infty)$ with parameter $1/2$.*

Thus, the final scale mixture representation for the horseshoe-like prior is:

$$(\theta_i\mid t_i,\tau)\sim\mathcal{N}\left(0,\frac{\tau^2}{t_i^2}\right),(t_i)\sim SN(0,1),\ t_i\in\mathbb{R},\ \tau^2>0, \tag{14}$$

or,

$$(\theta_i\mid t_i,\tau)\sim\mathcal{N}\left(0,\frac{\tau^2}{t_i^2}\right),(t_i\mid s_i)\sim\mathcal{N}(0,s_i),\ s_i\sim\text{Pareto}(1/2),\ t_i\in\mathbb{R},\ \tau^2>0. \tag{15}$$

5.1 Complete Conditionals and an MCMC Sampler

We use the scale-mixture representation of $SN(0,1)$ mixing density from Result 5.1:

$$\frac{(1-e^{-\frac{1}{2}t_i^2})}{\sqrt{2\pi}t_i^2}=\int_1^\infty\frac{1}{\sqrt{2\pi s_i}}\exp\left(-\frac{t_i^2}{2s_i}\right)\frac{1}{2s_i^{3/2}}ds_i=\int_0^1\frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{\nu_i t_i^2}{2}\right)d\nu_i,\ \text{where } \nu_i=s_i^{-1}.$$

We need to either specify a prior on the hyper-parameter τ (full Bayes) or treat it as a tuning parameter (empirical Bayes). Since τ is a scale parameter for $p(\theta_i)$, one option is a $C^+(0,1)$ prior on τ. We first present the steps in the MCMC scheme conditional on τ, where full conditionals of the other parameters are in closed form and then discuss simulation of τ, which requires a slice sampling step. Together, these steps constitute a Metropolis within Gibbs approach. Conditional on τ, the joint density is:

$$p(y,\theta,t,\nu\mid\tau)\ \propto\ \prod_{i=1}^n\exp\left\{-\frac{(y_i-\theta_i)^2}{2}\right\}\frac{|t_i|}{|\tau|}\exp\left(-\frac{t_i^2}{2\tau^2}\theta_i^2\right)\exp\left(-\frac{\nu_i t_i^2}{2}\right)\mathbf{1}\{0<\nu_i<1\}.$$

The complete conditionals given τ for $i=1,\ldots,n$ are:

$$(\theta_i\mid y_i,t_i,\nu_i,\tau)\sim\mathcal{N}\left(\left(1+\frac{t_i^2}{\tau^2}\right)^{-1}y_i,\left(1+\frac{t_i^2}{\tau^2}\right)^{-1}\right),$$

$$(t_i^2\mid y_i,\theta_i,\nu_i,\tau)\sim\text{Gamma}\left(\text{shape}=\frac{3}{2},\text{rate}=\frac{\theta_i^2}{2\tau^2}+\frac{\nu_i}{2}\right),$$

$$(\nu_i\mid y_i,t_i,\theta_i,\tau)\sim\text{Exponential}\left(\text{rate}=\frac{t_i^2}{2}\right)\mathbf{1}\{0<\nu_i<1\}.$$

Under the half Cauchy prior for τ, $p(\tau) \propto (1+\tau^2)^{-1}$, the conditional distribution of $\eta = 1/\tau^2$ is given by:

$$p(\eta \mid y, \theta, t, \nu) \propto \frac{1}{1+\eta} \eta^{\frac{n-1}{2}} \exp\left(-\frac{\eta}{2} \sum_{i=1}^{n} t_i^2 \theta_i^2\right).$$

Thus, the slice sampling steps for sampling η are:

1. Sample $(u \mid \eta)$ uniformly on $[0, (1+\eta)^{-1}]$.
2. Sample $(\eta \mid u) \sim \text{Gamma}((n+1)/2, \sum_{i=1}^{n} t_i^2 \theta_i^2/2)$, a Gamma density, truncated to have zero probability outside the interval $[0, (1-u)u^{-1}]$.

6 Simulation Study

We performed simulation studies to compare feature selection performances with the normal means model of Section 4.1 and the linear regression model of Section 4.2.

6.1 Normal Means Model

We take $n = 1000$. In true θ, components 1–10 are of magnitude 3, components 11–20 are of magnitude -3, followed by 980 zeros. Then we generate data as $(y_i \mid \theta_i) \overset{ind}{\sim} \mathcal{N}(\theta_i, 1)$ for $i = 1, \ldots, n$. We compare the horseshoe posterior mode obtained by the proposed EM algorithm of Section 4.1, posterior mean obtained from the MCMC algorithm of Section 5.1, SCAD, MCP and lasso. The results are summarized in Figure 2 and Table 1. The posterior mode correctly identifies 600 out of 880 zero components, which is the highest among all methods. It also identifies 19 of the 20 true non-zero features, indicating a good performance in subset selection. Since a method that performs well in subset selection can have poor ℓ_2 estimation properties (e.g., hard-thresholding), we also compare the methods for the sum of squared errors (SSE), defined as $\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2$, where $\hat{\theta}_i$ is the estimate of θ_i. The mode performs better than the two other non-convex penalties (SCAD and MCP). Lasso performs reasonably well in terms of SSE, but poorly in terms of detection of zeros and non-zeros, resulting in a denser solution. This behavior is well documented for convex penalties. The horseshoe posterior mean does not result in exact zero solutions. However, in terms of the SSE, it has the best performance among all methods. The reason for this can be seen from Figure 2, second panel from left. The bias of the horseshoe posterior mean goes to zero for large signals, whereas for smaller signals, there is stronger shrinkage compared to the lasso, but a smooth shrinkage profile (unlike the mode, SCAD or MCP). Lasso leaves a small but constant bias in the estimates, due its soft thresholding behavior. Finally, in terms of computational time, the proposed EM algorithm is orders of magnitude faster than state of the art non-convex solvers such as the R package `ncvreg`, which implements both SCAD and MCP or `sparsenet`, which implements coordinate descent algorithm to fit MCP.

6.2 Linear Regression Model

We take $n = 70, p = 350$. The true $\theta \in \mathbb{R}^p$ has components 1–10 are of magnitude 3, components 11–20 are of magnitude -3, followed by 330 zeros. The matrix of predictors $X \in \mathbb{R}^{n \times p}$ is generated

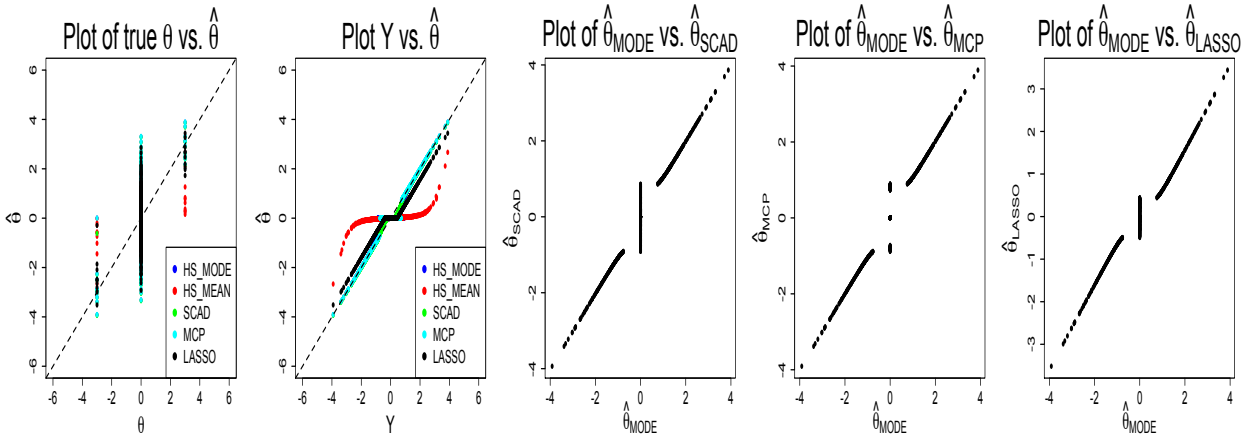

Figure 2: Simulation results for competing methods on sparse normal means model.

	MODE	MEAN	SCAD	MCP	LASSO
SSE	890.67	111.8	1006.45	977.98	540.5
COR_Z	600	NA	163	515	310
COR_NZ	19	NA	20	19	20
TIME	0.82	10.75	13.21	11.637	3.16

Table 1: Performance comparisons in normal means model for HS posterior mode, HS posterior mean, SCAD, MCP and LASSO. The rows are sum of squared error (SSE), zeros and non-zeros correctly detected (COR_Z & COR_NZ) and time in s. (TIME).

from i.i.d. standard normals. Finally, the observations are generated as $Y_i \overset{ind}{\sim} \mathcal{N}(X\theta, 1)$ for $i = 1, \ldots, n$. Figure 3 and Table 2 document the results. Here the posterior mode has the second best performance in detection of zeros. SCAD detects the highest number of true zeros correctly, but this comes at the expense of a poor performance in detection of non-zeros (7 out of 20) and a poor SSE. The horseshoe posterior mode results in sparser solution compared to MCP and lasso. The mode, MCP and lasso all perform well in the detection of non-zeros. Computation times for all methods (except MCMC) are comparable. As before, the mean has the lowest SSE, but does not give a sparse solution. The poor fit of SCAD in this case can be verified from the second panel from left of Figure 3, where the fitted $\hat{Y} = X\hat{\theta}$ values can be seen to be far away from the actual Y values for SCAD.

	MODE	MEAN	SCAD	MCP	LASSO
SSE	91.03	44.35	143.2	42.55	66.93
COR_Z	302	NA	323	276	292
COR_NZ	18	NA	7	20	20
TIME	0.248	14.978	0.226	0.561	0.177

Table 2: Performance comparisons in the regression model for HS posterior mode, HS posterior mean, SCAD, MCP and LASSO. The rows are sum of squared error (SSE), zeros and non-zeros correctly detected (COR_Z & COR_NZ) and time in s. (TIME).

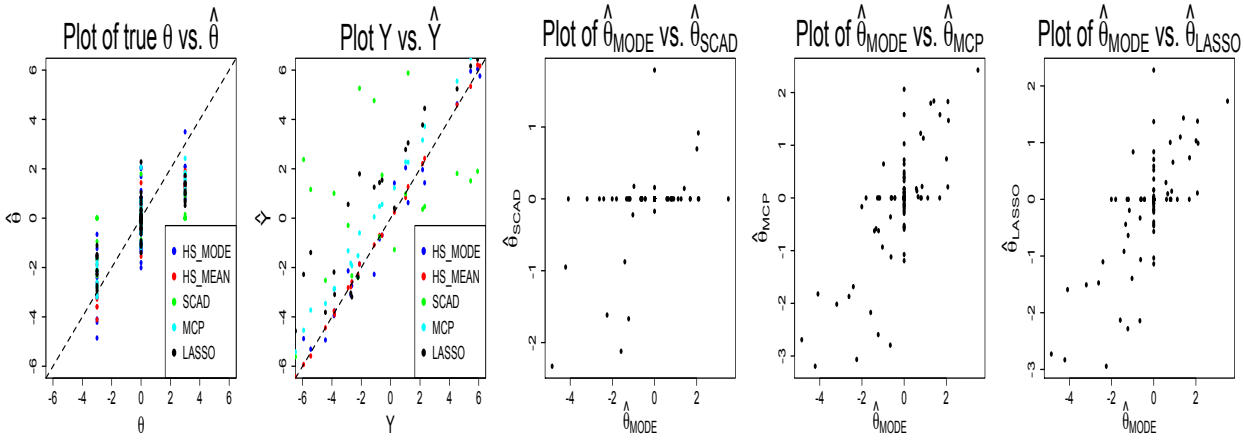

Figure 3: Simulation results for competing methods on sparse linear regression model.

6.3 Comparisons With The Horseshoe Prior

Since the horseshoe-like prior is a close approximation to the horseshoe prior, it is perhaps instructive to take a closer look at a comparison between the two. We first demonstrate the performance of the horseshoe-like prior in a simulation study for estimating a sparse normal mean vector with $(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2)$ and two different choices of $\boldsymbol{\theta}$: (1) $\boldsymbol{\theta}_1 = (\underbrace{7, \ldots, 7}_{q_n=10}, \overbrace{0, \ldots, 0}^{n-q_n=90})$ and (2) $\boldsymbol{\theta}_2 = (\underbrace{7, \ldots, 7}_{q_n=10}, \underbrace{3, \ldots, 3}_{r_n=10}, \overbrace{0, \ldots, 0}^{n-q_n-r_n=80})$. The choices are made to test the performance of horseshoe-like prior with sparse signals near the 'verge of detectability' $\sqrt{2 \log n}$ (Bogdan et al., 2011) as well as for signals with a relatively large magnitude, e.g. $2\sqrt{2 \log n}$ away from origin. Similar to the horseshoe prior, the horseshoe-like prior should be able to identify the signals in both cases. Figure 4 shows the estimated $\hat{\theta}$ under the horseshoe-like prior, along with the observations y_is and the 95% credible intervals. It is evident that the true signals are recovered in both the cases.

It is also instructive to compare the shrinkage profile of the horseshoe-like prior with that of the horseshoe prior for the second example. Figure 5 shows that although the shrinkage profile for the two priors are very similar, the horseshoe-like prior exerts a slightly stronger shrinkage on the noise observations near zero, but does not shrink the signals both near and far from the $\sqrt{2 \log n}$ boundary.

7 Leukemia Data Example

We consider a popular microarray gene expression data set with 3051 genes and 38 leukemia samples (Dudoit et al., 2002; Golub et al., 1999). This is a two class study where the goal is to identify genes that significantly differ between the 27 acute lymphoblastic leukemia (ALL) cases and 11 acute myeloid leukemia (AML) cases. The multiple testing for this data is carried out as follows: first a two-sample t-test with 36 degrees of freedom was performed for each 3,051

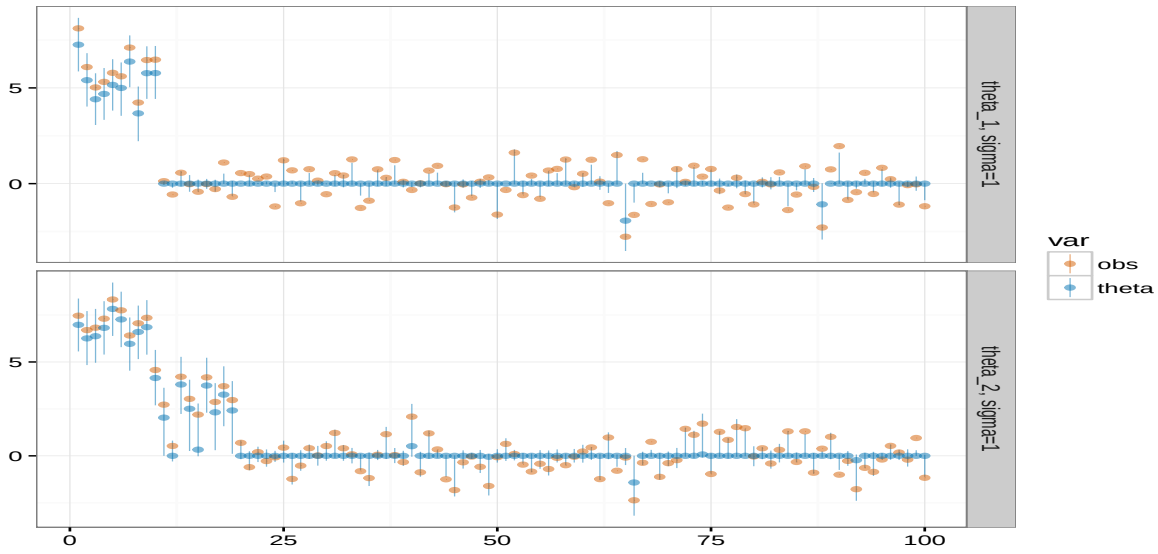

Figure 4: Comparison of posterior mean estimates for two different sparse normal means, $\boldsymbol{\theta}_1 \sim 0.1\delta_{\{7\}} + 0.9\delta_{\{0\}}$ (top) and $\boldsymbol{\theta}_2 \sim 0.1\delta_{\{7\}} + 0.1\delta_{\{3\}} + 0.8\delta_{\{0\}}$ (bottom) under the horseshoe-like prior.

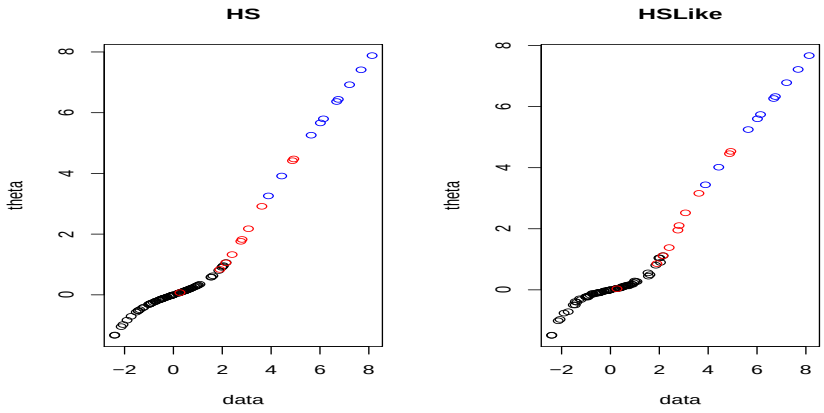

Figure 5: Comparison of posterior mean estimates under the horseshoe (HS) and horseshoe-like (HSLike) priors for $\boldsymbol{\theta} \sim 0.1\delta_{\{7\}} + 0.1\delta_{\{3\}} + 0.8\delta_{\{0\}}$. The black, red and blue circles represent true $\theta_i = 0, 3$ and 7 respectively.

genes and the t-test statistics are converted to z-test statistics using the quantile transformation $z_i = \Phi^{-1}(T_{36\text{ d.f.}}(t_i))$ for $i = 1, \ldots, 3051$. The i^{th} null hypothesis H_{0i} posits no difference in the gene expression levels for the i^{th} gene between the ALL and AML cases, and under the global null hypothesis $\cap H_{0i}$ the histogram of the z-values should mimic a $\mathcal{N}(0,1)$ curve closely. The histogram of the z-values along with the standard normal curve and a fitted normal density are shown in Figure 6. The departure of the histogram from the normal density curve suggests presence of many genes *differing* between the two classes.

The three classical multiple testing procedures, *viz.* Bonferroni, Benjamini–Hochberg and Bejamini–Yekutieli, identify 98, 681 and 269 genes as significant, by adjusting p-values obtained from the test statistics. Given the size of the data, the Bonferroni procedure is overly conservative

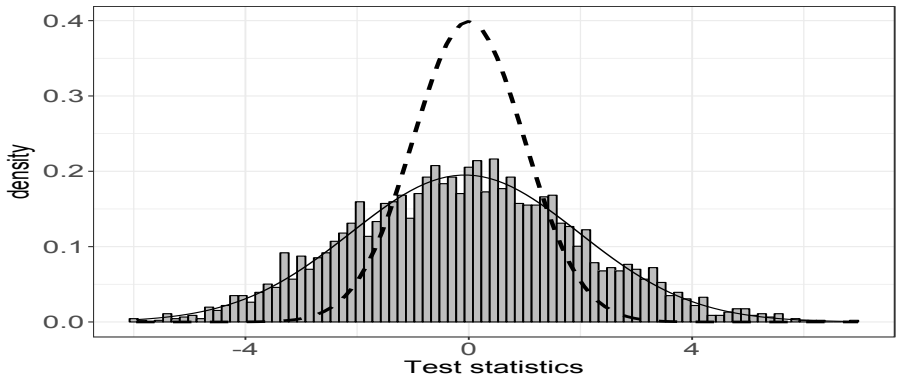

Figure 6: Histogram of z-values along with a dashed $\mathcal{N}(0,1)$ and a solid $\mathcal{N}(\bar{z},s)$ density curve, where $\bar{z}$ and s are the sample mean and standard deviation of the z-values.

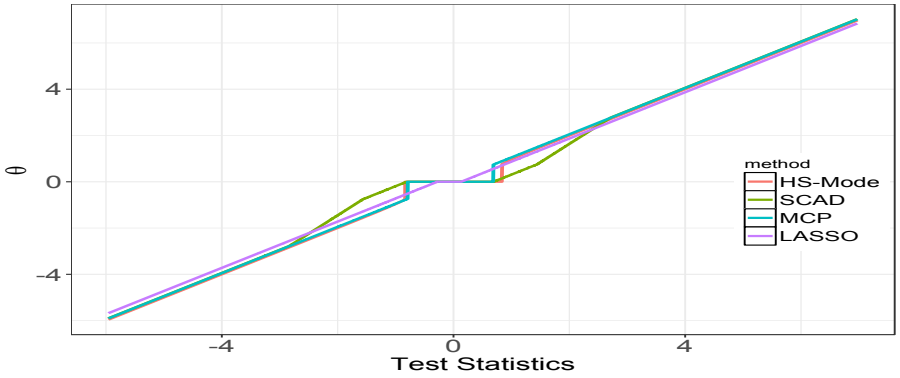

Figure 7: The posterior estimates for the competing methods versus the observed test statistics.

for large scale testing and Benjamini–Hochberg can be thought of as a recognized gold standard (see, e.g., Efron, 2010). In order to perform subset selection for this data, we compare the horseshoe posterior mode obtained by the proposed EM algorithm along with SCAD, MCP and lasso. It is also possible to use the posterior mean of the horseshoe-like prior with a thresholding rule as in Datta and Ghosh (2013) for performing multiple testing, but we do not consider it here since it is not a formal subset selection algorithm. Figure 7 compares the thresholding nature for the candidate methods and shows that the lasso is least conservative (declares 1,395 genes significant) and the horseshoe posterior mode is the most conservative (declares 987 genes significant) among them. Also, it appears that the three methods except the lasso induce somewhat similar thresholding rules. Figure 8 plots the estimated mean parameter $\hat{\theta}_i$s underlying the normal observations z_is with the points color-coded according to the Benjamini–Hochberg multiple testing rule. Once again, it seems that the horseshoe posterior mode performs similarly to SCAD and MCP and the lasso acts in an anti-conservative way, potentially leading to many false discoveries.

8 Conclusions and Future Work

We developed novel theoretical insights and fast computational algorithms for subset selection using the horseshoe regularization penalty. Our approach has a probabilistic representation, which

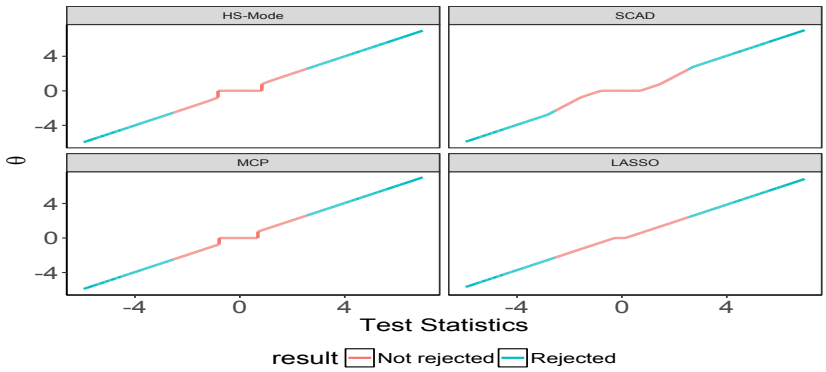

Figure 8: The posterior estimates versus the observed test statistics for different methods, with the points color-coded according to the Benjamini-Hochberg decision rule.

allows for simulating the entire posterior via MCMC, (Section 5), in addition to developing EM and LLA algorithms for identifying MAP point estimates (Section 4). In turn, this allows us to contrast the respective strengths and weaknesses of posterior mean and posterior mode. The former typically performs best in estimation under squared error loss, but is not sparse. These attributes are exactly reversed for the latter. In terms of both computational speed and statistical performance, horseshoe regularization outperforms state of the art non-convex solvers such as MCP or SCAD.

There are a number of directions for future work. For example, some other global-local priors that have shown promise in sparse Bayesian inference include the generalized beta (Armagan et al., 2011), the horseshoe+ (Bhadra et al., 2016a,c) and the Dirichlet–Laplace (Bhattacharya et al., 2015), to name a few. An open question is how these priors perform in terms of subset selection and whether fast computational algorithms are available. Following the recommendation of Gelman (2006), we used a standard half-Cauchy ($C^+(0,1)$) prior in (2), similar to the original horseshoe formulation (Carvalho et al., 2009; Polson and Scott, 2012). However, results in Piironen and Vehtari (2017) indicate it will be interesting to investigate the effect of the hyper-parameter η in a $C^+(0,\eta)$ prior in subset selection.

A more general family of proper prior densities can be constructed as follows:

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

for $\epsilon \geq 0, \tau > 0$, which reduces to the horseshoe-like prior of Equation (5) for $\epsilon = 1$. Furthermore, the density is approximately equal to $\theta_i^{1-\epsilon} \log(\theta_i^{-1})$ near the origin and the tails decay as $\theta_i^{-(1+\epsilon)}$. Thus, the main features of the horseshoe prior, that is, unboundedness at the origin and polynomially decaying tails, are preserved. The parameter ϵ represents a tradeoff between tail-heaviness and peakedness at the origin. For $\epsilon \in (0,1)$, the tails are heavier compared to the horseshoe, but at the cost of a smaller peak at the origin. The opposite is true for $\epsilon > 1$. Detailed investigation of this broader class of priors should be considered future work.

A Proof of Proposition 3.1

The hierarchy for the horseshoe-like prior can be written as:

$$y_i \mid \theta_i, \sigma^2 \sim \mathcal{N}(\theta_i, \sigma^2), \text{ where } \sigma^2 \sim \text{Inverse-Gamma}(\alpha, \beta),$$
$$p(\theta_i \mid \tau) = \frac{1}{2\pi\tau} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right).$$

Here we treat τ^2 as a tuning parameter. The marginal density of y_i is:

$$m(y_i \mid \tau) = \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_i-\theta_i)^2}{2\sigma^2}} \frac{1}{2\pi\tau} \log\left(1 + \frac{\tau^2}{\theta_i^2}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{-\frac{\beta}{\sigma^2}} d\theta_i d\sigma^2. \tag{A.1}$$

First, integrating out σ^2 under the Inverse-Gamma(α, β) hyper-prior gives the marginal likelihood:

$$p(y_i \mid \theta_i, \alpha, \beta) = \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)\Gamma(\frac{1}{2})} \beta^{-1/2} \frac{1}{\left(1 + \frac{1}{2\beta}(y_i - \theta_i)^2\right)^{\alpha+\frac{1}{2}}}, \ \alpha, \beta > 0. \tag{A.2}$$

For the special case $\alpha = \beta = \frac{1}{2}$, the marginal t distribution of y_i in (A.2) is into a Cauchy distribution with location θ_i, i.e.

$$p(y_i \mid \theta_i) = \frac{1}{\pi\{1 + (y_i - \theta_i)^2\}}.$$

Now, using Lemma 3.1, we can write the horseshoe-like prior as a $\mathcal{U}(0,1)$ scale mixture of Cauchy, or, in other words, $(\theta_i \mid \lambda_i, \tau) \sim \mathcal{C}(0, \lambda_i\tau)$ and $\lambda_i \sim \mathcal{U}(0,1)$. This hierarchy implies that the horseshoe-like prior is a member of the global-local mixtures described in Bhadra et al. (2016b), where the local shrinkage parameter has a $\mathcal{U}(0,1)$ prior, commonly used for the global shrinkage parameter τ. In a recent article, van der Pas et al. (2016) argue that restriction of the prior mass of τ to the interval $[1/n, 1]$ helps in achieving near-minimax rates as well as preventing degeneracy of the estimates of τ. Using this scale mixture representation in the hierarchy we can write:

$$m(y_i \mid \tau) = \int_{-\infty}^{\infty}\int_{0}^{1} \frac{1}{\pi\{1 + (y_i - \theta_i)^2\}} \frac{1}{\pi} \frac{\lambda_i\tau}{\lambda_i^2\tau^2 + \theta_i^2} d\lambda_i.$$

Equivalently,

$$Y_i - \theta_i \sim \mathcal{C}(0,1), \ \theta_i \sim \mathcal{C}(0, \lambda_i\tau), \text{ and } \lambda_i \sim \mathcal{U}(0,1),$$
$$\Rightarrow Y_i = (Y_i - \theta_i) + \theta_i \stackrel{\text{D}}{=} \mathcal{C}(0,1) + \lambda_i\tau\, \mathcal{C}(0,1) \stackrel{\text{D}}{=} \mathcal{C}(0, 1 + \lambda_i\tau), \text{ and } \lambda_i \sim \mathcal{U}(0,1).$$

The last equation follows from the following lemma (*vide* Bhadra et al. (2016b) for a proof using the Cauchy-Schlömilch integral identity).

LEMMA A.1. *Let $X_i \sim \mathcal{C}(0,1)$ $(i = 1, 2)$ be Cauchy distributed random variates, then $Z = w_1X_1 + w_2X_2 \sim \mathcal{C}(0, w_1 + w_2)$. where $w_1, w_2 > 0$.*

Hence the marginal of y_i is:

$$\begin{aligned} m(y_i \mid \tau) &= \int_0^1 \frac{1}{\pi(1+\lambda_i\tau)\left[1+\left\{\frac{y_i}{(1+\lambda_i\tau)}\right\}^2\right]} d\lambda_i \\ &= \frac{1}{\pi}\int_0^1 \frac{(1+\lambda_i\tau)}{\left\{(1+\lambda_i\tau)^2+y_i^2\right\}} d\lambda_i \\ &= \frac{1}{2\pi\tau}\int_1^{(1+\tau)^2} \frac{dt}{t+y_i^2} = \frac{1}{2\pi\tau}\log\left(1+\frac{\tau^2}{1+y_i^2}\right). \end{aligned}$$

References

Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions*. Dover Publications, New York.

Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P., Pereira, F. C. N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 523–531.

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review* **50,** 145–159.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016a). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* **103,** 955–969.

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

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

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

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

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5,** 232–253.

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.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39,** 1–38.

Devroye, L. (1986). *Nonuniform random variate generation*. Springer-Verlag, New York.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97,** 77–87.

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

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics* **32,** 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,** 1348–1360.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1,** 515–534.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation* **59,** 375–384.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33,** 1617–1642.

Jeffreys, H. and Swirles, B. (1972). *Methods of Mathematical Physics*. Cambridge University Press, Cambridge, 3rd edition.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220,** 671–680.

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

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing* **24,** 227–234.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization* **1,** 127–239.

Piironen, J. and Vehtari, A. (2017). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *The 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, page to appear.

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

Polson, N. G., Scott, J. G., and Willard, B. T. (2015). Proximal algorithms in statistics and machine learning. *Statistical Science* **30,** 559–581.

Schifano, E. D., Strawderman, R. L., and Wells, M. T. (2010). Majorization-minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics* **4,** 1258–1299.

Steutel, F. W. and Van Harn, K. (2003). *Infinite divisibility of probability distributions on the real line*. CRC Press.

Strawderman, R. L., Wells, M. T., and Schifano, E. D. (2013). Hierarchical bayes, maximum a posteriori estimators, and minimax concave penalized likelihood estimation. *Electronic Journal of Statistics* **7,** 973–990.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association* **82,** 528–540.

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

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

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

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36,** 1509–1533.