Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Ganeshkhind, Pune, MH, during January 02–04, 2020; pp 1–20

Horseshoe and Strawderman-Berger Estimator for Constrained Normal Means

Neha Agarwala [1], Junyong Park [2], Anindya Roy[1]
[1]*Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, USA*
[2]*Department of Mathematics and Statistics, Seoul National University, Seoul, South Korea*

Abstract

Key words: Constrained normal means, Skrinkage estimators, mixture distribution, skew normal distribution, MCMC

1. Introduction

Traditional statistical theory have mostly focused on methods developed for large samples and small number of features. Modern scientific world, however, is moving fast towards the regime of high dimensional data. In high dimensional setting, often one deals with the case when only few variables are relevant. Thus it has become increasingly important to identify true signals as the data tends to be sparse. Probably the most common of such high dimensional sparse estimation problems is estimation of normal mean when sample size is small compared to the dimension. It is the proverbial needle in a haystack problem that has received much attention in the literature. The setting of the problem is simple. Given data $y_1, \ldots, y_n$ arising independently from the model

$$y_i|\mu_i \sim N(\mu_i, \sigma^2),$$

one wishes to estimate the entire vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Of course, given that there are only n independent observations for $(n+1)$ unknown parameters, additional assumptions need to be made for meaningful estimation of the mean vector. Usually some level of sparsity if assumed for the true mean vector. Both Bayesian and frequentist estimators have been developed for this problem, the most well known being the shrinkage estimators starting with James and Stein (1961) [9] thresholding estimators starting with [6] Donoho and Johnston (1994), penalized estimators such as Lasso (Tibshirani, 1996) [17], SCAD (Fan and Li, 2001) [7] and many other variants of them.

In the Bayesian setting, the popular approaches are using the spike and slab priors and continuous shrinkage priors. Formulation of sparse mean vector scenarios as a combination of two regimes where the mean values are zero or arising from a measure which allows for possibly large values naturally leads to a mixture prior of the form

$$p(\mu) = p\delta_0 + (1-p)g(\mu)$$

Corresponding Author: Anindya Roy
Email: anindya@umbc.edu

where the point mass p as $\mu = 0$ is the *spike* and the probability density $g(\cdot)$ allowing μ to take possibly large non-zero values is the *slab*. Mitchell and Beauchamp (1988) [13] considered it in the context of variable selection in Gaussian regression. Since then such priors have gained popularity in many contexts including variable selection, covariance matrix estimation, false discovery rate estimation. Many authors have advocated the use of such point mixture prior for normal mean estimation. Strawderman-Berger (SB) prior (Strawderman and Berger, 1996) [3] explicitly considered in this article is an example of such a spike-and-slab prior in a hierarchical setting where the hyper-parameters governing the slab $g(\cdot)$ is allowed to change according to some prior for each μ_i. Specifically, they propose the following model

$$\begin{aligned} \mu_i|\tau,\lambda_i &\sim N(0,\tau^2\lambda_i^2) \\ p(\lambda_i) &\propto \lambda_i(1+\lambda_i^2)^{1/2} \\ p(\tau) &\sim \sigma C[\sigma,\sigma] I(\tau > \sigma) \end{aligned} \tag{1}$$

where $C[a,b]$ is the Cauchy density.

A version of the spike-slab prior considered recently is the non-local prior recommended by Johnson and Rossell (2010, 2012) [11], [10] where the slab is well separated from the spike at zero.

Another class of priors considered for sparse estimation of mean are the shrinkage priors or the global-local priors. Park and Casella (2008) [14] proposed a scale mixture of Gaussian prior that they called the *Bayesian Lasso*. However, these priors do not have sufficient prior mass near zero to work well in the very sparse regime. Carvalho *et al.* (2010) [5] proposed the *horseshoe* (HS) prior defined as

$$\begin{aligned} \mu_i|\tau,\lambda_i &\sim N(0,\tau^2\lambda_i^2) \\ p(\lambda_i) &\propto C[0,1]_+ \\ p(\tau) &\sim \sigma C[0,1]_+ \end{aligned} \tag{2}$$

where $C[0,1]_+$ is the half-Cauchy density, the standard Cauchy truncated to the positive half. The horseshoe prior has only one component as opposed to the two separate components of the spike-and-slab priors but overcomes the deficiency of the Bayesian Lasso in sparse regime by allowing infinite prior density at zero.

While full Bayesian analysis is possible, empirical Bayes solutions have also been discussed for the two component mixture priors such as SB and the single component shrinkage priors such as HS, empirical Bayes solutions for high-dimensional sparse mean estimation have been also looked at in the literature; see Johnston and Silverman (2004) [12], Brown and Greenshtein (2009) [4]

Often one has prior knowledge on the range of possible values for the mean parameter, such as the parameter is non-negative. One way of estimating such a parameter is to first obtain an unrestricted estimate of the parameter and then truncate it so that the estimate lies in the constrained parameter space. Intuitively, the performance of the estimator is expected to be much better if such constraint conditions are incorporated in the model. Constrained estimation of normal mean restricted to convex cones has been discussed in Sen and Silvapulle (2001) [16]. Danaher *et al.* (2012) provide an example of Bayesian estimation of normal mean when the mean is constrained to a convex polytope.

In this paper we particularly look at the case when the dimension is large and the mean vector is assumed to be sparse. We focus on the high dimensional normal means estimation problem where the mean vector is constrained to be in a closed convex polyhedral cone. Let $\boldsymbol{y} = (y_1, \ldots, y_n)' \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where the parameter of interest $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ is assumed to belong to the convex cone

$$\mathcal{K} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{\mu} \geq 0\} \tag{3}$$

where $\boldsymbol{A}$ is some fixed $r \times n$ matrix. We assume that $\mathcal{K}$ has non-zero interior volume with respect to the n dimensional Lebesgue measure. Of course, one of the most interesting question is how to specify sparsity in constrained spaces such as $\mathcal{K}$. However, the scope of this paper is very limited. Without getting into a discourse about sparsity in constrained sets such as $\mathcal{K}$, we simply compare the performance of sparsity generating spike-and-slab priors such as Strawderman-Berger and shrinkage priors such as Horseshoe, when the priors are defined in terms of scale mixtures of truncated normal instead of normal. This straightforward generalization is probably not optimal, particularly if the conic geometry is very different from that of the entire space. However, for most part we will consider $\mathcal{K}$ to be the positive orthant, a case which bears similarity with the unconstrained case to a large extent.

In Section 2 we discuss the Bayes estimators for the Strawderman-Berger and the Horseshoe prior extended to the convex cone case. In Section 3 we present results of a limited numerical experiment comparing the performance of posterior quantities obtained using different priors along with that of the maximum likelihood estimator (MLE) projected to the convex cone. We end with some discussions in Section 4.

2. Methods

We consider the model $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{\mu} \in \mathcal{K} = \{\boldsymbol{\mu} : \boldsymbol{A}\boldsymbol{\mu} \geq 0\}$. First we consider a straightforward generalization of the global-local prior to the constrained case when the prior on $\boldsymbol{\mu}$ is supported on $\mathcal{K}$. Extension of the two-component mixture prior to the convex cone restriction is more nuanced and is discussed second.

2.1 Horseshoe Extension

Specifically, let

$$\begin{aligned} \boldsymbol{\mu}|\boldsymbol{\lambda}, \tau &\sim TN(0, \tau^2 \boldsymbol{\Lambda}, \mathcal{K}) \\ (\lambda_1, \ldots, \lambda_n) &\sim \prod_{i=1}^{n} p_\lambda(\lambda_i) \\ \tau &\sim p_\tau(\tau) \end{aligned} \tag{4}$$

where $\boldsymbol{\Lambda} = diag\{\lambda_1^2, \ldots, \lambda_n^2\}$ and $TN(\psi, \boldsymbol{\Sigma}, \mathcal{K})$ denotes normal with the distribution of a multivariate normal with mean ψ, variance matrix $\boldsymbol{\Sigma}$ and truncated to the cone $\mathcal{K}$.

It is interesting is to investigate the effect of the conic geometry on the Bayes estimates. Of course, the truncated normal prior will be conjugate, yielding a truncated normal posterior. If the model is $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$, and the prior density for $\boldsymbol{\mu}$ is $TN(\boldsymbol{0}, \boldsymbol{V}, \mathcal{K})$ and $\boldsymbol{Q} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{V}^{-1})^{-1}$, then the posterior is $\boldsymbol{\mu}|\boldsymbol{y} \sim TN(\boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y}, \boldsymbol{Q}, \mathcal{K})$. Thus, the posterior mean could be directly computed for that of a truncated normal, albeit truncated to a general convex polyhedral cone. We derive the expression for the

posterior mean using a slightly different argument which is instructive in the sense it provides explicit expressions for the marginal of $\boldsymbol{y}$ using hidden truncation argument.

Before we give our main result, we define some useful notation. Let $\Phi^{(r)}(\boldsymbol{z};\boldsymbol{\xi},\boldsymbol{W}) = P(\boldsymbol{Z} \leq \boldsymbol{z})$ for $\boldsymbol{Z} \sim N(\boldsymbol{\xi},\boldsymbol{W})$ where Φ is the standard normal cdf. Also, for $\boldsymbol{x} = (x_1,\ldots,x_n)'$, let $\phi^{(n)}(\boldsymbol{x}) = \prod_{i=1}^{n}\phi(x_i)$ where ϕ is the standard normal pdf.

The following result provides some insight to how the half-plane restrictions, $\boldsymbol{A}$ appears in the expression for the posterior mean.

Theorem 1. *Let $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu},\boldsymbol{\Omega})$ where it is known a priori that $\boldsymbol{\mu} \in \mathcal{K}$, a polyhedral convex cone defined by $\mathcal{K} = \{\boldsymbol{\mu} : \boldsymbol{A\mu} \geq \boldsymbol{0}\}$ for some matrix $\boldsymbol{A}$. Let the prior on $\boldsymbol{\mu}$ be $\boldsymbol{\mu} \sim N(0,\boldsymbol{V})_{\mathcal{K}}$. Let $\boldsymbol{F} = \boldsymbol{AQA'} = \boldsymbol{DRD'}$, say, where $\boldsymbol{D}$ be a diagonal matrix with entries equal to the square root of the diagonal entries of $\boldsymbol{F}$ and $\boldsymbol{Q} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{V}^{-1})^{-1}$. Also for $i = 1,\ldots,r$, let $\boldsymbol{R}_{-i}$ be $\boldsymbol{R}$ without the ith column and the ith row and let $\boldsymbol{r}_{-i}$ denote the ith column of $\boldsymbol{R}$ without the ith diagonal element. Let $\boldsymbol{B}_i = [\boldsymbol{I} : -\boldsymbol{r}_{-i}]$ where $\boldsymbol{I}$ is the identity matrix of dimension $(r-1)$ and let $\boldsymbol{u} = (u_1,\ldots,u_r)' = \boldsymbol{D}^{-1/2}\boldsymbol{AQ\Omega}^{-1}\boldsymbol{y}$. Assuming $\boldsymbol{\Omega}$ and $\boldsymbol{V}$ are fixed and given, we have*

$$E(\boldsymbol{\mu}|\boldsymbol{y}) = \boldsymbol{Q}[\boldsymbol{\Omega}^{-1}\boldsymbol{y} + \boldsymbol{A'D}^{-1/2}\boldsymbol{v}]$$

where $\boldsymbol{v} = (v_1,\ldots,v_r)'$ and $v_i = \phi(-u_i)\Phi^{(r-1)}(\boldsymbol{B}_i\boldsymbol{u};\boldsymbol{0},\boldsymbol{R}_{-i})/\Phi^{(r)}(\boldsymbol{u};\boldsymbol{0},\boldsymbol{R})$.

Proof. The joint model for $\boldsymbol{y}$ and $\boldsymbol{\mu}$ is

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{\mu}\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix} \quad \begin{pmatrix}\boldsymbol{\Omega}+\boldsymbol{V} & \boldsymbol{V}\\ \boldsymbol{V} & \boldsymbol{V}\end{pmatrix}\right)$$

Hence that of $\boldsymbol{y}$ and $\boldsymbol{A\mu}$ is

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{\mu}\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix} \quad \begin{pmatrix}\boldsymbol{\Omega}+\boldsymbol{V} & \boldsymbol{VA'}\\ \boldsymbol{AV} & \boldsymbol{AVA'}\end{pmatrix}\right)$$

Then following Arnold (2009) [2], the marginal density formula for $\boldsymbol{y}$ under the hidden truncation $\boldsymbol{A\mu} \geq \boldsymbol{0}$, is

$$p_y(\boldsymbol{y}) = |\boldsymbol{\Sigma}_{11}|^{-1/2}\phi^{(n)}(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y})\frac{\Phi^{(r)}(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y};\boldsymbol{0},\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})}{\Phi^{(r)}(\boldsymbol{0};\boldsymbol{0},\boldsymbol{\Sigma}_{22})}$$

where

$$\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Omega}+\boldsymbol{V} & \boldsymbol{VA'}\\ \boldsymbol{AV} & \boldsymbol{AVA'}\end{pmatrix}.$$

Then by the multiparameter version of Tweedie's formula (Robbins, 1956) [15], we have

$$\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}) = \boldsymbol{y} + \boldsymbol{\Omega}\nabla_{\boldsymbol{y}}\log p_y(\boldsymbol{y})$$

The gradient of $\log p_y(\boldsymbol{y})$ has two parts, The first part is $\nabla_{\boldsymbol{y}}\phi^{(n)}(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y})$. By the chain rule of vector differentiation, we get

$$\nabla_{\boldsymbol{y}}\phi^{(n)}(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y}) = -\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y} = -(\boldsymbol{\Omega}+\boldsymbol{V})^{-1}\boldsymbol{y}.$$

Therefore,

$$\begin{aligned} \mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}) &= \boldsymbol{y} - \boldsymbol{\Omega}(\boldsymbol{\Omega}+\boldsymbol{V})^{-1}\boldsymbol{y} + \boldsymbol{\Omega}\frac{\nabla_{\boldsymbol{y}}\Phi^{(r)}(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y};\boldsymbol{0},\boldsymbol{F})}{\Phi^{(r)}(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y};\boldsymbol{0},\boldsymbol{F})}, \\ &= \boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y} + \boldsymbol{\Omega}\frac{\nabla_{\boldsymbol{y}}\Phi^{(r)}(-\boldsymbol{A}\boldsymbol{z};\boldsymbol{0},\boldsymbol{F})}{\Phi^{(r)}(-\boldsymbol{A}\boldsymbol{z};\boldsymbol{0},\boldsymbol{F})} \end{aligned}$$

where $\boldsymbol{F} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \boldsymbol{A}\boldsymbol{Q}\boldsymbol{A}'$. To compute the gradient we use the standard formula for partial derivatives of the multivariate cdf of a random vector $\boldsymbol{X} = (X_i, \ldots, X_k)$ given by $\frac{\partial}{\partial x_i}F(x_1, \ldots, x_k) = f_i(x_i)F_{-i|i}(x_{-i})$ where f_i is the marginal density of X_i, $F_{-i|i}$ is the conditional cdf of the rest of the components of $\boldsymbol{X}$ given X_i and $\boldsymbol{x}_{-i}$ is the vector $\boldsymbol{x} = (x_1, \ldots, x_k)$ without the ith component. Using the result for multivariate normal cdf we have $\frac{\partial}{\partial u_i}\Phi^{(r)}(\boldsymbol{u};\boldsymbol{0},\boldsymbol{R}) = \phi(u_i)\Phi^{(r-1)}(\boldsymbol{B}_i\boldsymbol{u};\boldsymbol{0},\boldsymbol{R}_{-i})$. The result follows using the chain rule of vector differentiation. □

The expression for the posterior mean has two parts. The first part $\boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y}$ is the usual Bayes estimator normal-normal conjugacy which is the unbiased estimator $\boldsymbol{y}$ plus a Bayes correction. However, under the conic constraint the second term acts as a correction for the restriction to the convex cone.

Let the entries of the covariance matrices be functions of some lower dimensional parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\}$. For example, for the usual Horseshoe prior formulation, $\boldsymbol{\theta} = \{\sigma^2, \lambda_1^2, \ldots, \lambda_n^2, \tau^2)$. Even though the expression for the posterior mean in Theorem 1 is derived with fixed $\boldsymbol{\Omega}$ and $\boldsymbol{V}$, it is instructive to write the posterior mean as $\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y},\boldsymbol{\theta})$. If priors are specified on $\boldsymbol{\theta}$, then the posterior mean for $\boldsymbol{\mu}$ an be obtained as

$$\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}) = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y},\boldsymbol{\theta})),$$

where the first expectation on the right hand side is taken over the marginal posterior of $\boldsymbol{\theta}$.

The marginal distribution of $\boldsymbol{y}$ given the truncated normal prior is $p_y(\boldsymbol{y})$ and it belonds to the *closed (fundamental) skew normal* family; see Gonzalez-Farias *et al.* (2004) [8], Arellano-Vallea and Genton, (2005) [1]. The marginal distribution can be used for estimation of hyper-parameter to obtain the marginal posterior of $\boldsymbol{\theta}$. For example, one could use the fundamental skew normal likelihood directly.

2.2 Strawderman-Berger Extension

Extending the two-component prior to the constrained requires a more careful approach. In the unrestricted case, the sparse regime is generated using a binomial experiment of binary strings $(Z_1, \ldots, Z_n)$ on the n dimensional Hamming space, where the mean value is chosen to be zero for the cases where the binary string has one. The probability of one is constant and equal to p, e.g. $Z_i \overset{iid}{\sim} Bernoulli(p)$.

For the case when the mean parameter $\boldsymbol{\mu}$ is restricted to $\mathcal{K}$ and its components are chosen to be sparse, the choice of zero values could mean choosing the vector to lie on a lower dimensional face of the cone. By Minkowski-Weyl thoerem, the polyhedral cone is finitely generated by set of extreme rays $\boldsymbol{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s\}$, say. The prior on $\boldsymbol{\mu}$ could be invoked through a reparamterization in terms of $\boldsymbol{\mu} \leftrightarrow \sum \gamma_i \boldsymbol{u}_i$.

The faces of $\mathcal{K}$ are, up to symmetries, formed by conic combinations of all possible subsets of $\boldsymbol{U}$. The Hamming space $(Z_1, \ldots, Zs)$ could be identified with all possible subsets of $\{1, 2, \ldots, s\}$ and for a given realization of the binary indicators, the mean vector would lie on a face

$$F = \{ \sum_{i:Z_i=0} \gamma_i \boldsymbol{u}_i : \gamma_i \geq 0\}$$

The nonzero part would then be sampled from the $TN(\boldsymbol{0}, \boldsymbol{V}, \mathcal{K})$ restricted to the face F, ie. $TN(\boldsymbol{0}, \boldsymbol{V}_F, F)$, where $\boldsymbol{V}_F$ is the sub-matrix of $\boldsymbol{V}$. Thus, essentially a two-component prior is a mixture of a measure choosing the lower dimensional face F and then a second measure on $\mathcal{K}$ whose restriction to F provides the prior distribution of $\boldsymbol{\mu}$.

2.3 Non-negative Orthant

The form of the marginal is a product of skew-normal densities when F and Σ_{11} become diagonal. For example, when $\mathcal{K} = \mathbb{R}^n_+$, we have $\boldsymbol{A} = \boldsymbol{I}$. in this case the marginal factorizes as product of skew-normal densities.

Horseshoe prior:
The extension of the horseshoe prior to the convex cone restriction would be to assume half Cauchy prior for the λ_i and τ (check this):

$$\begin{aligned} \mu_i|\tau, \lambda_i &\sim N(0, \tau^2\lambda_i^2)_+, \\ \lambda_i &\sim C(0, 1)_+ \end{aligned} \tag{5}$$

where $N(0, 1)_+$ represent a standard normal distribution truncated from below at 0 and $C(0, 1)_+$ represent a standard half-Cauchy distribution on the positive reals. One can estimate σ and τ using an Empirical Bayes approach. We use a Jeffrey's prior on σ and standard half-Cauchy prior with scale equal to σ on τ.

$$\begin{aligned} \pi(\sigma) &\propto \frac{1}{\sigma}, \\ \tau|\sigma &\sim C(0, \sigma)_+ \end{aligned} \tag{6}$$

Carvalho *et al* (2010) [5] described λ_i as the local shrinkage parameter and τ the global shrinkage parameter. Horseshoe prior essentially is a scale mixture of truncated normals, scale being a function of a common variance component, τ and an individual variance component, λ_i for each μ_i.

Conditional on σ, τ and λ_i's, $\mu_i|\boldsymbol{y}$ is independently distributed with

$$\mu_i|\lambda_i, \tau, \sigma, \boldsymbol{y} \sim N\left(\frac{\frac{y_i}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2\lambda_i^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2\lambda_i^2}} \right)_+ \tag{7}$$

Let $m_i = \left(1 - \frac{\frac{1}{\lambda_i^2\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\lambda_i^2\tau^2}}\right) y_i$ and $s_i^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2\lambda_i^2}}$.

$$E(\mu_i|\lambda_i, \tau, \sigma, \boldsymbol{y}) = m_i + \frac{\phi(\frac{-m_i}{s_i})}{1 - \Phi(\frac{-m_i}{s_i})} s_i \tag{8}$$

The Bayes estimator of μ_i is given by

$$\hat{\mu}_i = E(\mu_i|\boldsymbol{y}) \quad = E_{\lambda_i,\tau,\sigma|\boldsymbol{y}} E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}) \tag{9}$$

For $t > 0$,

$$t \;<\; \frac{\phi(t)}{1-\Phi(t)} \;<\; \frac{1+t^2}{t} \tag{10}$$

- $\hat{\mu}_i > 0$ since $E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}) > 0$ using the inequality in (10).
- $\hat{\mu}$ is non-decreasing in y. (see Appendix)
- For $y > 0$ and large, the non-linear part in equation (8) is close to 0.

Strawderman-Berger Prior:

The extension of Strawderman-Berger prior for non-negative orthant puts a truncated normal distribution in place of the usual normal distribution.

$$\begin{aligned} \pi(\mu_i) &= p\delta_o + (1-p)\; N(0,\tau^2\lambda_i^2)_+ \\ \pi(\lambda_i) &\propto \lambda_i(1+\lambda_i^2)^{\frac{3}{2}} \\ p &\sim \text{Unif}(0,1) \end{aligned} \tag{11}$$

Similar to Horseshoe, we use a Jeffrey's prior on σ and for τ, we use a truncated Cauchy prior with location and scale both equal to σ bounded below at σ.

$$\begin{aligned} \tau|\sigma &\sim C(\sigma,\sigma)\; 1(\tau \geq \sigma) \\ \pi(\sigma) &\propto \frac{1}{\sigma} \end{aligned} \tag{12}$$

Conditional on λ_i,τ,p,σ, the posterior distribution of μ_i is a mixture distribution where the posterior probability of $\mu_i = 0$, c_i acts as local shrinkage.

$$\pi(\mu_i|\lambda_i,\tau,p,\sigma,\boldsymbol{y}) = c_i\delta_o + (1-c_i)\; N(m_i,s_i^2)_+ \tag{13}$$

where

$$c_i = \frac{\frac{p}{\sigma}\phi(\frac{y_i}{\sigma})}{\frac{p}{\sigma}\phi(\frac{y_i}{\sigma}) + \frac{2(1-p)}{l_i}\phi(\frac{y_i}{l_i})\Phi(\frac{m_i}{s_i})}$$

and $l_i^2 = \sigma^2 + \lambda_i^2\tau^2$.

$$E(\mu_i|\boldsymbol{\lambda},\tau,\sigma,p,\boldsymbol{y}) = (1-c_i)\left(m_i + \frac{\phi\left(\frac{-m_i}{s_i}\right)}{\Phi\left(\frac{m_i}{s_i}\right)}s_i\right)$$

The Bayes estimator for μ_i is the posterior mean, $E(\mu_i|\boldsymbol{y}) = E_{\boldsymbol{\lambda},\tau,p|\boldsymbol{y}} E(\mu_i|\boldsymbol{\lambda},\tau,\sigma,p,\boldsymbol{y})$ which is obtained using Metropolis-Hastings.

The following results hold for the posterior mean based on Strawderman-Berger prior.

- $\hat{\mu}_i > 0$ since $E(\mu_i|\lambda_i, \tau, p, \sigma, \boldsymbol{y}) > 0$ using the inequality in (10).
- $\hat{\mu}$ is non-decreasing in y. (see Appendix)
- For $y > 0$ and large, the non-linear part in equation (8) is close to 0.

In the two component model, the posterior mean could be computed in a manner similar to that computed for the Horseshoe type prior. However, for the spike-and-slab type prior, it is more interesting to look at the componentwise posterior median. Thus, for an additive loss

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum |\mu_i - \hat{\mu}_i|$$

it makes sense to look at the componentwise posterior median, $\hat{\boldsymbol{\mu}}_M$. One could show that the posterior median $\hat{\mu}_i(y_i)$ is a continuous shrinkage soft thresholding rule, in the sense that

1. $\hat{\mu}_i(y_i) \leq y_i$ for all $y_i > 0$.
2. For each y, these exists $T(y)$ such that $\hat{\mu}_M(y) = 0$ if $y < T(y)$.

$$P(\mu_i|\boldsymbol{y}) = E_{\boldsymbol{\theta}|\boldsymbol{y}}\Big(c_i 1(\mu_i = 0) + (1 - c_i)\frac{1}{s_i}\,\phi\big(\frac{\mu_i - m_i}{s_i}\big)\Big) \tag{14}$$

For each y_i, let us define $T(y_i) := \frac{s_i}{1-k_i}\Phi^{-1}\Big\{\frac{p}{2(1-p)}\Big\}$, where $k_i = \frac{\frac{1}{\lambda_i^2\tau^2}}{\frac{1}{\sigma^2}+\frac{1}{\lambda_i^2\tau^2}}$. Then,

$$\begin{aligned}
y_i &\leq T(y_i) \\
\implies p - 2(1-p)\Phi(\frac{m_i}{s_i}) &\geq 0 \\
\implies E_{\boldsymbol{\theta}|\boldsymbol{y}}\left\{\frac{p - 2(1-p)\Phi(\frac{m_i}{s_i})}{p + 2(1-p)\Phi(\frac{m_i}{s_i})}\right\} &\geq 0 \\
\implies E_{\boldsymbol{\theta}|\boldsymbol{y}}\left\{\frac{\frac{p}{\sigma}\phi(\frac{y_i}{\sigma})}{\frac{p}{\sigma}\phi(\frac{y_i}{\sigma}) + \frac{2(1-p)}{\sigma}\phi(\frac{y_i}{\sigma})\Phi(\frac{m_i}{s_i})} - \frac{1}{2}\right\} &\geq 0 \\
\implies E_{\boldsymbol{\theta}|\boldsymbol{y}}(c_i) &\geq 0.5 \\
\implies \hat{\mu}_M(y_i) &= 0
\end{aligned}$$

3. Posterior Computation and Numerical Results

We use Metropolis-Hastings algorithm to generate random samples from the posterior distribution, $\mu_i|\boldsymbol{\theta}, \boldsymbol{y}$ and thus obtain the Bayes estimator $E(\mu_i|\boldsymbol{y})$ by taking average over the randomly generated samples of $\boldsymbol{\theta}$. For the horseshoe prior, $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \tau, \sigma\}$ and for the Strawderman-Berger prior, $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \tau, p, \sigma\}$.

3.1 Simulation for HS Posterior

The Bayes estimator of μ_i is given by

$$\hat{\mu}_i = E(\mu_i|\boldsymbol{y}) \quad = E_{\boldsymbol{\lambda},\tau,\sigma|\boldsymbol{y}} E(\mu_i|\boldsymbol{\lambda},\tau,\sigma,\boldsymbol{y}) \tag{15}$$

We want to simulate from the posterior distribution of $\boldsymbol{\lambda},\tau,\sigma|\boldsymbol{y}$ which is given below:

$$\pi(\boldsymbol{\lambda},\tau,\sigma|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda},\tau,\sigma)\ \pi(\boldsymbol{\lambda})\ \pi(\tau|\sigma)\ \pi(\sigma)$$

where

$$\begin{aligned}
\pi(\boldsymbol{y}|\boldsymbol{\lambda},\tau,\sigma) &= \int \pi(\boldsymbol{y},\boldsymbol{\mu}|\boldsymbol{\lambda},\tau,\sigma)d\boldsymbol{\mu} \\
&= \int \pi(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\lambda},\tau,\sigma)\pi(\boldsymbol{\mu}|\boldsymbol{\lambda},\tau,\sigma)d\boldsymbol{\mu} \\
&= \prod_{i=1}^{n} \int \pi(y_i|\mu_i,\lambda_i,\tau,\sigma)\pi(\mu_i|\lambda_i,\tau,\sigma)d\mu_i \\
&= \prod_{i=1}^{n} \left[\frac{1}{l_i}\phi(y_i/l_i)\Phi(m_i/s_i)\right] \\
&= \prod_{i=1}^{n} \pi(y_i|\lambda_i,\tau,\sigma,p)
\end{aligned}$$

Simulating $\boldsymbol{\lambda_i}$: The conditional distribution of $\lambda_i|\tau,\sigma,\boldsymbol{y}$ for $i = 1,\ldots,n$, is given by

$$\pi(\lambda_i|\tau,\sigma,\boldsymbol{y}) \propto \pi(y_i|\lambda_i,\tau,\sigma)\ \pi(\lambda_i)$$

For each iteration j, we generate

$$\lambda_i^* \sim N(\lambda_i^{(j-1)},1)1(\lambda_i^* > 0) \equiv q_{\lambda_i}(\lambda_i^*|\lambda_i^{(j-1)})$$

and $u \sim U(0,1)$ where $q_{\lambda_i}(x^*|x)$ is the proposal density. We accept λ_i^*, if $R < u$, where R is the acceptance ratio given by

$$R = \frac{\pi(\lambda_i^*|\tau,\sigma,\boldsymbol{y})\ q_{\lambda_i}(\lambda_i^{(j-1)}|\lambda_i^*)}{\pi(\lambda_i^{(j-1)}|\tau,\sigma,\boldsymbol{y})\ q_{\lambda_i}(\lambda_i^*|\lambda_i^{(j-1)})}$$

Otherwise, we accept the $\lambda_i^{(j-1)}$.

Simulating $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}$: The conditional distribution of $\tau,\sigma|\boldsymbol{\lambda},\boldsymbol{y}$ is given by

$$\pi(\tau,\sigma|\boldsymbol{\lambda},\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda},\tau,\sigma)\ \pi(\tau|\sigma)\ \pi(\sigma)$$

For each iteration j, we generate

$$\tau^* \sim N(\tau^{(j-1)},1)1(\tau^* > 0) \equiv q_\tau(\tau^*|\tau^{(j-1)})$$

$$\sigma^* \sim N(\sigma^{(j-1)},1)1(\sigma^* > 0) \equiv q_\sigma(\sigma^*|\sigma^{(j-1)})$$

and $u \sim U(0,1)$ where $q_\tau(x^*|x)$ is the proposal density for τ and $q_\sigma(x^*|x)$ is the proposal density for σ. We accept τ^* and σ^*, if $R < u$, where R is the acceptance ratio given by

$$R = \frac{\pi(\tau^*,\sigma^*|\boldsymbol{\lambda},\boldsymbol{y})\ q_\tau(\tau^{(j-1)}|\tau^*)\ q_\sigma(\sigma^{(j-1)}|\sigma^*)}{\pi(\tau^{(j-1)},\sigma^{(j-1)}|\boldsymbol{\lambda},\boldsymbol{y})\ q_\tau(\tau^*|\tau^{(j-1)})\ q_\sigma(\sigma^*|\sigma^{(j-1)})}$$

Otherwise, we accept the $\tau^{(j-1)}$ and $\sigma^{(j-1)}$.

3.2 Simulation for SB Posterior

The Bayes estimator of μ_i is given by

$$\hat{\mu}_i = E(\mu_i|\boldsymbol{y}) \quad = E_{\boldsymbol{\lambda},\tau,\sigma,p|\boldsymbol{y}} E(\mu_i|\boldsymbol{\lambda},\tau,\sigma,p,\boldsymbol{y}) \tag{16}$$

We want to simulate from the posterior distribution of $\boldsymbol{\lambda},\tau,\sigma,p|\boldsymbol{y}$ which is given below:

$$\pi(\boldsymbol{\lambda},\tau,\sigma,p|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda},\tau,\sigma,p)\ \pi(\boldsymbol{\lambda})\ \pi(p)\ \pi(\tau|\sigma)\ \pi(\sigma)$$

where

$$\begin{aligned}
\pi(\boldsymbol{y}|\boldsymbol{\lambda},\tau,\sigma,p) &= \int \pi(\boldsymbol{y},\boldsymbol{\mu}|\boldsymbol{\lambda},\tau,\sigma,p)d\boldsymbol{\mu} \\
&= \int \pi(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\lambda},\tau,\sigma,p)\pi(\boldsymbol{\mu}|\boldsymbol{\lambda},\tau,p)d\boldsymbol{\mu} \\
&= \prod_{i=1}^{n} \int \pi(y_i|\mu_i,\lambda_i,\tau,\sigma,p)\pi(\mu_i|\lambda_i,\tau,\sigma,p)d\mu_i \\
&= \prod_{i=1}^{n} [(1-p)\frac{1}{\sigma}\phi(y_i/\sigma) + 2\ p\ \frac{1}{l_i}\phi(y_i/l_i)\ \Phi\Big(\frac{m_i}{s_i}\Big)] \\
&= \prod_{i=1}^{n} \pi(y_i|\lambda_i,\tau,\sigma,p)
\end{aligned}$$

Simulating $\boldsymbol{\lambda_i}$: The conditional distribution of $\lambda_i|\tau,p,\sigma,\boldsymbol{y}$ for $i = 1,\ldots,n$, is given by

$$\pi(\lambda_i|\tau,p,\sigma,\boldsymbol{y}) \propto \pi(y_i|\lambda_i,\tau,p,\sigma)\ \pi(\lambda_i)$$

For each iteration j, we generate

$$\lambda_i^* \sim N(\lambda_i^{(j-1)},1)1(\lambda_i^* > 0) \equiv q_{\lambda_i}(\lambda_i^*|\lambda_i^{(j-1)})$$

and $u \sim U(0,1)$ where $q_{\lambda_i}(x^*|x)$ is the proposal density. We accept λ_i^*, if $R < u$, where R is the acceptance ratio given by

$$R = \frac{\pi(\lambda_i^*|\tau,p,\sigma,\boldsymbol{y})\ q_{\lambda_i}(\lambda_i^{(j-1)}|\lambda_i^*)}{\pi(\lambda_i^{(j-1)}|\tau,p,\sigma,\boldsymbol{y})\ q_{\lambda_i}(\lambda_i^*|\lambda_i^{(j-1)})}$$

Otherwise, we accept the $\lambda_i^{(j-1)}$.

Simulating $\boldsymbol{p}$: The conditional distribution of $p|\boldsymbol{\lambda},\tau,\sigma,\boldsymbol{y}$ is given by

$$\pi(p|\boldsymbol{\lambda},\tau,\sigma,\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda},\tau,\sigma,p)\ \pi(p)$$

For each iteration j, we generate

$$p^* \sim N(p^{(j-1)},1)1(0 < p^* < 1) \equiv q_p(p^*|p^{(j-1)})$$

and $u \sim U(0,1)$ where $q_p(x^*|x)$ is the proposal density for p. We accept p^*, if $R < u$, where R is the acceptance ratio given by

$$R = \frac{\pi(p^*|\boldsymbol{\lambda}, \tau, \sigma, \boldsymbol{y})\ q_p(p^{(j-1)}|p^*)}{\pi(p|\boldsymbol{\lambda}, \tau, \sigma, \boldsymbol{y})\ q_p(p^*|p^{(j-1)})}$$

Otherwise, we accept the $p^{(j-1)}$.

Simulating $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}$: The conditional distribution of $\tau, \sigma|\boldsymbol{\lambda}, p, \boldsymbol{y}$ is given by

$$\pi(\tau, \sigma|\boldsymbol{\lambda}, p, \boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda}, p, \tau, \sigma)\ \pi(\tau|\sigma)\ \pi(\sigma)$$

For each iteration j, we generate

$$\sigma^* \sim N(\sigma^{(j-1)}, 1)1(\sigma^* > 0) \equiv q_\sigma(\sigma^*|\sigma^{(j-1)})$$

$$\tau^*|\sigma^* \sim N(\tau^{(j-1)}, 1)1(\tau^* > \sigma^*) \equiv q_\tau(\tau^*|\tau^{(j-1)})$$

and $u \sim U(0,1)$ where $q_\tau(x^*|x)$ is the proposal density for τ and $q_\sigma(x^*|x)$ is the proposal density for σ. We accept τ^* and σ^*, if $R < u$, where R is the acceptance ratio given by

$$R = \frac{\pi(\tau^*, \sigma^*|\boldsymbol{\lambda}, p, \boldsymbol{y})\ q_\tau(\tau^{(j-1)}|\tau^*)\ q_\sigma(\sigma^{(j-1)}|\sigma^*)}{\pi(\tau^{(j-1)}, \sigma^{(j-1)}|\boldsymbol{\lambda}, p, \boldsymbol{y})\ q_\tau(\tau^*|\tau^{(j-1)})\ q_\sigma(\sigma^*|\sigma^{(j-1)})}$$

Otherwise, we accept the $\tau^{(j-1)}$ and $\sigma^{(j-1)}$.

3.3 Simulation Design

We compare the performances of Strawderman-Berger estimators, Horseshoe estimator and Maximum Likelihood Estimator (MLE) under different degree of sparsity. The MLE when $\boldsymbol{\mu} \in \mathcal{K} = \mathbb{R}^n_+$ for $\Sigma = \sigma^2 I$ is simply the projection of $\boldsymbol{y}$ onto the non-negative orthant .ie., $\hat{\mu}_i = \max(y_i, 0)$. For a general polyhedral cones with Σ other than $\sigma^2 I$, the MLE is difficult to compute.

We analyze the risk properties of the estimators when the mean vector is simulated under strongly sparse signals and weakly sparse signals. For each of the sparsity level, we further consider two scenarios described below.

Strong sparsity: We use a discrete mixture model to generate exact zero entries for the mean vector using the model below:

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \pi(\mu_i) &= p\delta_o + (1-p)\ G(\alpha, \beta) \end{aligned} \tag{17}$$

where α is taken to be 5, β is 0.5 and 80% of the mean vector has exact zero entries. The major concentration of μ_i's is at 0 with an average concentration of $\mu_i > 0$ at mean 10 with variance 20. Two possible values of σ are considered: $\sigma = 1$ and $\sigma = 3$. The y_i's at $\mu_i = 0$ and $\mu_i > 0$ have a much clearer separation for $\sigma = 1$ than $\sigma = 3$.

Weak sparsity: For weakly sparse signals, we generate $\boldsymbol{\mu}$ which decays according to the power law but none of its components are exactly zero.

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i|\eta, \alpha &\sim \text{Unif}(0, \eta c_i) \\ \eta &\sim \text{Ex}(2) \\ \alpha &\sim \text{Unif}(a, b) \end{aligned} \tag{18}$$

where $c_i = (n/i)^{1/\alpha}$ for $i = 1, \ldots, n$. For simulation purposes, $\sigma = 1$ is chosen and two possible scenarios of $\alpha \sim \text{Unif}(a, b)$ are considered: $a = 0.5, b = 1$ and $a = 1, b = 2$. The first scenario yields relatively large mean entries than the second scenario depending on the randomly generated values of η and α. When $\alpha \sim \text{Unif}(1, 2)$, one can expect the concentration around 0 to be more dense than when $\alpha \sim \text{Unif}(0.5, 1)$ determined by the speed of decay, α.

For each of the scenarios, we simulate 1000 data sets from the corresponding model of dimension $n = 300$ using MCMC with 50000 runs and a burn-in period of 10000. We report the median risk under squared error loss and absolute error loss along with the average risk ratios between the estimators in Table 1 and Table 2.

Figure 1 shows the plots for MLE estimates, posterior mean under horseshoe prior and posterior mean and posterior median under Strawderman-Berger prior for strongly sparse signals for $\sigma = 1$ and $\sigma = 3$. The dimension of the mean vector is 300. Figure 2 presents the same under weakly sparse signals for the two scenarios when $\alpha \sim \text{Unif}(0.5, 1)$ and $\alpha \sim \text{Unif}(1, 2)$.

3.4 Simulation Results

From figure 1 and figure 2, we see that the posterior mean based on Strawderman-Berger prior shrinks more than the horseshoe posterior mean estimator. Also, posterior median for Strawderman-Berger prior is exactly zero for a positive y whereas MLE is equal to zero only for negative y's. The posterior mean for Strawderman-Berger prior does a better shrinkage than that for horseshoe prior, though both of them is always positive and never attains exact zero.

Table 1 shows that the risk performance of SB posterior median and posterior mean is better than the MLE and horseshoe posterior mean both in terms of squared error loss and absolute loss for the strong sparsity case. In particular, the horseshoe posterior mean has at least 50% more risk than both the Strawderman-Berger posterior mean and posterior median. However, the risk for horseshoe posterior mean under squared error loss is $20\% - 35\%$ less than the Strawderman-Berger estimators when $\sigma = 3$.

From table 2, we see that the risk of horseshoe posterior mean is consistently less than that of MLE and SB posterior mean and posterior median. Specifically, horseshoe posterior mean has of $6\% - 40\%$ more risk than the Strawderman-Berger estimators. However, when $\alpha \sim U(1, 2)$, horseshoe estimator has 63% more risk than the SB posterior mean and approximately 411% more risk than SB posterior median, although the median squared error risk is less for horseshoe than the other estimators.

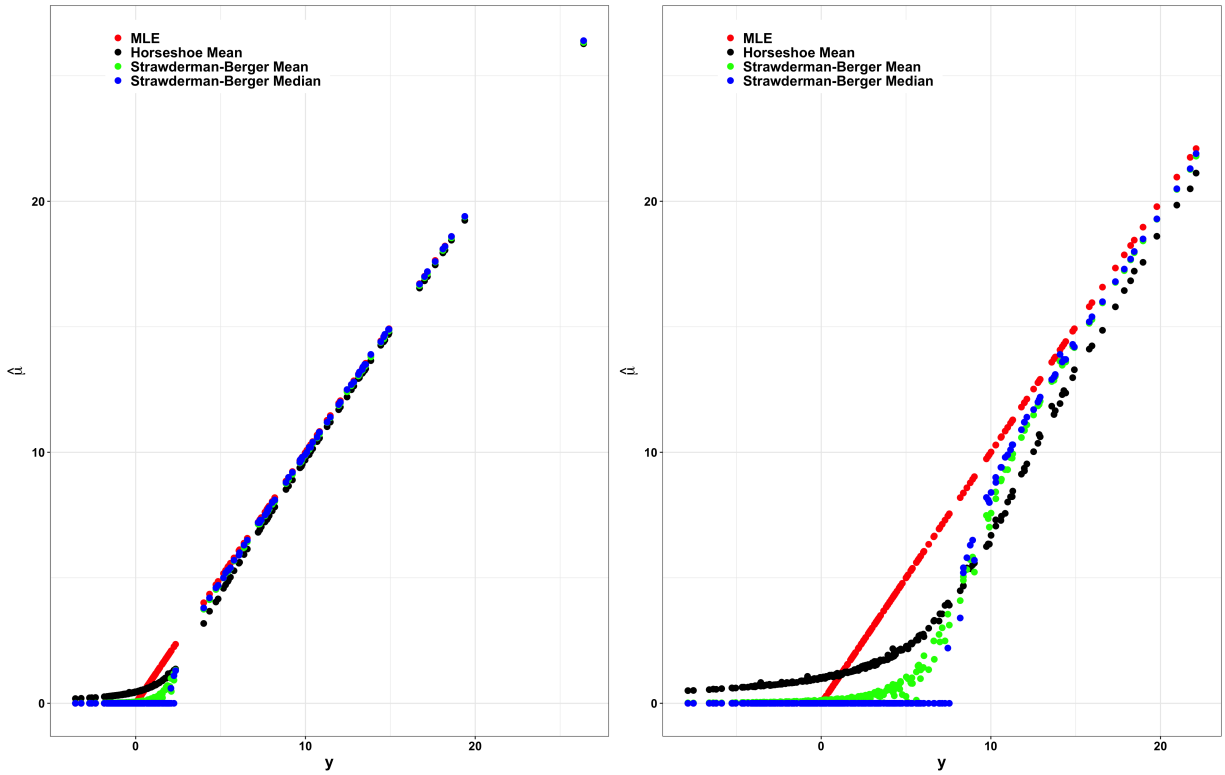

Figure 1: Plots of $\hat{\mu}$ versus y under strong sparsity with $\sigma = 1$ (left) and $\sigma = 3$ (right)

		$\sigma = 1$				$\sigma = 3$			
		MLE	HS	SB Mean	SB Median	MLE	HS	SB Mean	SB Median
Square Error Loss	MLE	171	1.39	2.23	2.33	1598	1.19	0.98	0.77
	HS		131	1.6	1.67		1361	0.81	0.64
	SB Mean			82	1.04			1636	0.78
	SB Median				78				2129
Absolute Error Loss	MLE	143	0.92	1.97	2.6	428	0.95	1.43	1.42
	HS		156	2.13	2.8		452	1.5	1.49
	SB Mean			73	1.32			295	0.98
	SB Median				56				299

Table 1: Risk under squared error loss and absolute error loss for strongly sparse signals in two scenarios: $\sigma = 1$ and $\sigma = 3$. The diagonal components are median sum of squared error and absolute error. The off diagonal components are average error ratios of estimator in row by estimator in column.

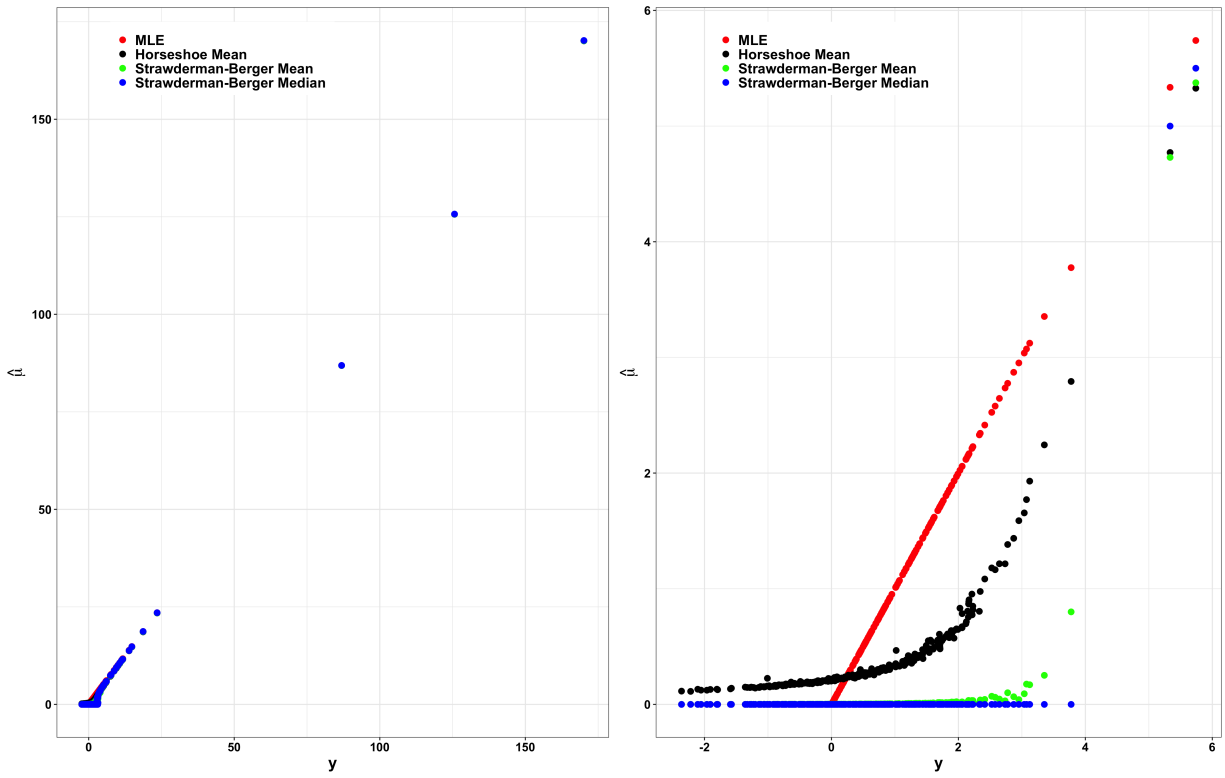

Figure 2: Plots of $\hat{\mu}$ versus y under weak sparsity where $\alpha \sim U(0.5, 1)$ (left) and $\alpha \sim U(1, 2)$ (right)

4. Discussion

In our simulation studies, we compared the performance of horseshoe posterior mean, Strawdermna-Berger posterior mean and posterior median for strongly sparse signals and weakly sparse signals. While the posterior mean for both horseshoe and Strawdermna-Berger prior are shrinkage estimators, MLE and SB posterior median are truncation based estimators with exact zeros for small signals.

In this paper, the numerical studies for non-negative orthant is restricted to horseshoe prior and Strawderman-Berger prior. It would be interesting to consider other scale mixture distributions, similar to lasso, with hard thresholding properties for non-negative mean vectors. Another interesting domain is the discrete mixture models where the positive means has a more flexible distribution like Gamma and the priors on scale and shape hyperparameters can be some heavy tailed distribution. While the scope of this paper is limited to non-negative orthant which has many popular applications, one can think of exploring some of these priors to a general closed convex polyhedral cones.

		$\alpha \sim U(0.5,1)$				$\alpha \sim U(1,2)$			
		MLE	HS	SB Mean	SB Median	MLE	HS	SB Mean	SB Median
Square Error Loss	MLE	200	2.68	2.91	2.65	179	15.7	128	400
	HS		122	0.73	0.6		63	1.63	5.11
	SB Mean			185	0.81			120	0.92
	SB Median				235				136
Absolute Error Loss	MLE	181	1.6	1.75	1.63	166	3.09	4.32	4.8
	HS		136	0.94	0.85		86	0.91	0.93
	SB Mean			162	0.89			128	0.95
	SB Median				186				134

Table 2: Risk under squared error loss and absolute error loss for weakly sparse signals in two scenarios: $\alpha \sim U(0.5,1)$ and $\alpha \sim U(1,2)$. The diagonal components are median sum of squared error and absolute error. The off diagonal components are average error ratios of estimator in row by estimator in column.

5. Appendix

5.1 The Bayes estimator $\hat{\mu}$ is non-decreasing in y

$$\begin{aligned} y|\mu &\sim N(\mu,1), \quad \mu \sim g(\mu) \\ g(\mu) &= \pi\delta_o + (1-\pi)g_1(\mu) \end{aligned} \tag{19}$$

An estimator of μ is then given by

$$\begin{aligned} l(y) = E(\mu|y) &= \frac{\int \mu\phi(y-\mu)g(\mu)}{\int \phi(y-\mu)g(\mu)} \\ &= \frac{(1-\pi)\int \mu\phi(y-\mu)g_1(\mu)d\mu}{\pi\phi(y) + (1-\pi)\int \phi(y-\mu)g_1(\mu)d\mu} \\ &= \frac{\int \mu\phi(\mu)g_1(\mu)e^{\mu y}d\mu}{\frac{\pi}{1-\pi} + \int \phi(\mu)g_1(\mu)e^{\mu y}d\mu} \\ &= \frac{a(y)}{b(y)} \end{aligned} \tag{20}$$

where $a(y) = \int \mu\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu$ and $b(y) = \frac{\pi}{1-\pi} + \int \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu$.
Then $a'(y) = \int \mu^2\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu$ and $b'(y) = a(y)$

$$
\begin{aligned}
l'(y) &= \frac{b(y)a'(y) - a(y)b'(y)}{b^2(y)} \\
&= \frac{(\frac{\pi}{1-\pi} + \int \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu)(\int \mu^2\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu) - (\int \mu\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu)^2}{b^2(y)} \\
&= \frac{\frac{\pi}{1-\pi} \int \mu^2 f^*(\mu)\ d\mu + c(y) \int \mu^2 f^*(\mu)\ d\mu) - (\int \mu f^*(\mu)\ d\mu)^2}{\left(\frac{\pi}{1-\pi} + c(y)\right)^2}
\end{aligned}
\tag{21}
$$

where $f^*(\mu) = \phi(\mu)\ g_1(\mu)\ e^{\mu y}$ and $c(y) = \int f^*(\mu) d\mu$.
Therefore $l'(y)$ reduces to

$$
\begin{aligned}
l'(y) &= \frac{\frac{\pi}{1-\pi}\frac{1}{c(y)} \int \mu^2\ \frac{f^*(\mu)}{c(y)}\ d\mu + \int \mu^2\ \frac{f^*(\mu)}{c(y)}\ d\mu) - (\int \mu\ \frac{f^*(\mu)}{c(y)}\ d\mu)^2}{\left(\frac{\pi}{1-\pi}\frac{1}{c(y)} + 1\right)^2} \\
&= \frac{\frac{\pi}{(1-\pi)c(y)} E(\mu^2) + V(\mu)}{\left(\frac{\pi}{1-\pi}\frac{1}{c(y)} + 1\right)^2} \ \geq 0\ \forall y
\end{aligned}
\tag{22}
$$

Hence $l(y)$ is non-decreasing function of y for any $g_1(\mu)$ defined on positive μ.

The proof is similar for horseshoe prior.

References

[1] Reinaldo B. Arellano-Valle and Marc G. Genton. On fundamental skew distributions. *Journal of Multivariate Analysis*, 96(1):93–116, September 2005.

[2] Barry C. Arnold and Héctor W. Gómez. Hidden Truncation and Additive Components : Two Alternative Skewing Paradigms:. *Calcutta Statistical Association Bulletin*, March 2009. Publisher: SAGE PublicationsSage India: New Delhi, India.

[3] James O. Berger and William E. Strawderman. Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.*, 24(3):931–951, June 1996. Publisher: Institute of Mathematical Statistics.

[4] Lawrence D. Brown and Eitan Greenshtein. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.*, 37(4):1685–1704, August 2009. Publisher: Institute of Mathematical Statistics.

[5] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, June 2010. Publisher: Oxford Academic.

[6] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, September 1994. Publisher: Oxford Academic.

[7] 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, December 2001. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214501753382273.

[8] Graciela Gonzalez-Farias, Armando Dominguez-Molina, and Arjun K. Gupta. Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference*, 126(2):521–534, December 2004.

[9] W. James and Charles Stein. Estimation with Quadratic Loss. The Regents of the University of California, 1961. ISSN: 0097-0433.

[10] Valen E. Johnson and David Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B*, 72(2):143–170, 2010. Publisher: Royal Statistical Society.

[11] Valen E. JOHNSON and David ROSSELL. Bayesian Model Selection in High-Dimensional Settings. *J Am Stat Assoc*, 107(498), 2012.

[12] Iain M. Johnstone and Bernard W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649, August 2004. Publisher: Institute of Mathematical Statistics.

[13] 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. Publisher: Taylor & Francis.

[14] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214508000000337.

[15] Herbert Robbins. An Empirical Bayes Approach to Statistics. The Regents of the University of California, 1956. ISSN: 0097-0433.

[16] Mervyn J. Silvapulle and Pranab Kumar Sen. *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. John Wiley & Sons, September 2011. Google-Books-ID: TEh4fcDVfbQC.

[17] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. Publisher: [Royal Statistical Society, Wiley].