Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 2, 2020 (New Series), pp 317–332

Horseshoe and Strawderman-Berger Estimators for Non-negative Normal Means

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

Received: 22 June 2020; Revised: 25 August 2020; Accepted: 06 September 2020

Abstract

High dimensional inference problems are generating considerable interest due to the availability and accessibility of massive amount of data in several fields. Modern statistical problems, however, involve natural constraints on model parameters. For such estimation problems, it is not fitting to apply standard estimates designed for unrestricted alternatives and then to truncate it. Given a n dimensional independent normal observation with common variance, we consider the classical normal mean estimation problem where the mean vector lies in a non-negative orthant. We study the behavior and risk properties of Bayesian estimators under two popular priors, the horseshoe prior and Strawderman-Berger prior, originally developed in the unrestricted mean vector estimation regime and then restrict the distribution of prior to satisfy the parameter constraint. The performance of posterior mean based on the horseshoe prior and the posterior mean and posterior median based on Strawderman-Berger prior is compared with the maximum likelihood estimator, numerically under different sparsity configurations.

Key words: Constrained normal means; Shrinkage estimators; Mixture distribution; Sparsity.

1. Introduction

Traditional statistical theory has mostly focused on methods developed for large samples and a small number of features. The modern scientific world, however, is moving fast towards the regime of high dimensional data. In the 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 the mean of a normal distribution 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

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

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

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 are needed for meaningful estimation of the mean vector. Usually some level of sparsity is 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), thresholding estimators starting with Donoho and Johnston (1994), penalized estimators such as Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) and many other variants of them.

In the Bayesian setting, popular approaches include using the spike and slab priors and continuous shrinkage priors for sparse mean estimation. 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

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

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) 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 priors for normal mean estimation. Strawderman-Berger (SB) prior (Strawderman and Berger, 1996) 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 $g(\cdot)$ are allowed to change according to some prior for each μ_i. Specifically, they propose the following model

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

where $C[a, b]$ is the Cauchy density with location and scale equal to a and b, respectively.

A version of the spike-slab prior considered recently is the non-local prior recommended by Johnson and Rossell (2010, 2012) where the slab is well separated from the spike at zero. Being a single component prior, horseshoe type priors are computationally less demanding than the spike-slab priors.

Another class of priors considered for sparse estimation of mean are the shrinkage priors or the global-local priors. Park and Casella (2008) 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) proposed the *horseshoe* (HS) prior defined as

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

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 Strawderman-Berger and the single component shrinkage priors such as horseshoe. Empirical Bayes solutions for high-dimensional sparse mean estimation have been also looked at in the literature; see Johnston and Silverman (2004), Brown and Greenshtein (2009).

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). Danaher *et al.* (2012) provides 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

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

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, given its special importance in the applications, we will only consider $\mathcal{K}$ to be the positive orthant, $\mathbb{R}^d_+$. The geometry of the positive orthant is very similar to the unrestricted linear space, but there are subtle differences in estimation due to the constraint and that is what we explore via numerical investigation.

In Section 2 we discuss the Bayes estimators for the Strawderman-Berger and the horseshoe priors when they are extended to the convex cone case. In Section 3 we present results of a 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. Sparse Priors for the Non-negative Orthant

In this paper we consider the restriction $\mu_i \geq 0$ for all i and hence $\boldsymbol{\mu} \in \mathcal{K} = \mathbb{R}^n_+$. For the horseshoe prior and the Strawderman-Berger prior for the non-negative orthant, we simply replace the normal prior for μ_i with normal truncated to the positive half. To judge the performance of the estimators under different priors in the constrained case, we set forth a list of desirable properties. These are analogous to desirable properties in a sparse mean estimator in the unrestricted case, except adapted to the constrained mean case. For example, one would want the estimators for μ_i to provide considerable shrinkage for small to moderate y_i whereas to leave y_i nearly unperturbed for large positive y_i. In the constrained case, for negative y_i one would expect the estimated mean to be nearly zero, if not exactly zero. The maximum likelihood estimator for μ_i is exactly zero whenever y_i is negative.

Horseshoe prior

The extension for the horseshoe to the positive orthant considered here is then

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

where $N(\theta, v)_+$ represent a $N(\theta, v)$ truncated from below at 0 and $C(0,1)_+$ represent a standard half-Cauchy distribution on the positive reals. We use a Jeffrey's prior on σ and standard half-Cauchy prior with scale equal to σ on τ.

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

One could estimate σ and τ using an Empirical Bayes approach. However, here we use a full Bayesian framework. Carvalho *et al* (2010) described λ_i as the local shrinkage parameter and τ the global shrinkage parameter. For the positive orthant, the horseshoe prior that we are considering is essentially 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, $\mu_i|\boldsymbol{y}$ are independently distributed as

$$\mu_i|\lambda_i, \tau, \sigma, \boldsymbol{y} \sim N(m_i, s_i^2)_+$$

where $m_i = s_i^2 \frac{y_i}{\sigma^2}$ and $s_i^2 = [\frac{1}{\sigma^2} + \frac{1}{\tau^2\lambda_i^2}]^{-1}$. Then, we have

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

The Bayes estimator of μ_i is given by

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

From the bounds on the Mill's ratio for the standard normal, we know that for $t > 0$,

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

This implies $E(\mu_i|\lambda_i, \tau, \sigma, \boldsymbol{y}) > 0$ for all y. Also, for $y_i < 0$, $E(\mu_i|\lambda_i, \tau, \sigma, \boldsymbol{y}) < \sigma^2|y_i|^{-1}$.

Moreover, for large positive y_i, $E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}) \approx [1-\frac{\tau^2\lambda_i^2}{\sigma^2+\tau^2\lambda_i^2}]y_i$. Hence, summarizing we have

Result 1: For the horseshoe prior for the constrained case when the true mean is restricted to the non-negative orthant:

1. $E(\mu_i|\boldsymbol{y}) > 0$ for all $\boldsymbol{y}$.
2. For $y_i < 0$, $E(\mu_i|\boldsymbol{y}) = O(|y_i|^{-1})$.
3. For large positive y_i, $E(\mu_i|\boldsymbol{y}) \approx \mathrm{E}([1-\frac{\tau^2\lambda_i^2}{\sigma^2+\tau^2\lambda_i^2}]|\lambda_i,\tau,\sigma,\boldsymbol{y})y_i$.

Thus, the posterior mean of μ_i acts as a shrinkage estimator and its behavior is similar to what observed in the unrestricted case.

Strawderman-Berger Prior

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

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

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

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

Conditional on λ_i,τ,p,σ, the posterior distribution of μ_i is a mixture distribution

$$\pi(\mu_i|\lambda_i,\tau,p,\sigma,\boldsymbol{y}) = c(\theta_i,y_i)\ \delta_o + \Big(1-c(\theta_i,y_i)\Big)\ N(m_i,s_i^2)_+ \tag{3}$$

where

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

is the posterior probability of $\mu_i = 0$ which acts as local shrinkage, $\theta_i = \{\lambda_i,\tau,\sigma,p\}$ and $l_i^2 = \sigma^2 + \lambda_i^2\tau^2$ for $i = 1,\ldots,n$.
Then, we have

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

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

Result 2: The following results hold for the posterior mean computed based on the Strawderman-Berger prior in the constrained case:

1. $\hat{\mu}_i > 0$ since $E(\mu_i|\lambda_i, \tau, p, \sigma, \boldsymbol{y}) > 0$ using the inequality in (2).
2. $\hat{\mu}_i$ is non-decreasing in y_i.
3. For large positive y_i, $E(\mu_i|\boldsymbol{y}) \approx \mathrm{E}\Big[\Big(1 - c(\theta_i, y_i)\Big)\Big(1 - \frac{\tau^2\lambda_i^2}{\sigma^2+\tau^2\lambda_i^2}\Big)|\lambda_i, \tau, \sigma, \boldsymbol{y}\Big] y_i$.

See the appendix for a proof of the above Result 2.

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 component-wise posterior median. For the posterior median, we use the estimator,

$$\hat{\mu}_{iM}(\mu_i|\boldsymbol{y}) = F_i^{-1}(1/2)$$

where

$$F_i(t) = E_{\boldsymbol{\theta}|\boldsymbol{y}}[P(\mu_i \leq t|\theta_i, y_i)]$$

and

$$P(\mu_i \leq t|\theta_i, y_i) = c(\theta_i, y_i) + \Big(1 - c(\theta_i, y_i)\Big)\ \Phi^{-1}\bigg(\frac{m_i}{s_i}\bigg)\ \bigg[\Phi\bigg(\frac{t - m_i}{s_i}\bigg) - \Phi\bigg(-\frac{m_i}{s_i}\bigg)\bigg].$$

A more specific form of the posterior median is

$$\hat{\mu}_{iM}(\mu_i|\boldsymbol{y}) = \begin{cases} 0 & \text{if } E_{\boldsymbol{\theta}|\boldsymbol{y}}\Big(c(\theta_i, y_i)\Big) \geq 0.5, \\ \inf\{x \geq 0 : F_i(x) \geq 0.5\} & \text{if } E_{\boldsymbol{\theta}|\boldsymbol{y}}\Big(c(\theta_i, y_i)\Big) < 0.5. \end{cases} \tag{4}$$

Thus, for an additive loss

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

it makes sense to look at the component-wise posterior median, $\hat{\boldsymbol{\mu}}_M$.

One could also look at the Empirical Bayes estimator of the median which is the expectation of the posterior median expression with respect to $p(\boldsymbol{y}|\boldsymbol{\theta})$. Let

$$G_i(t|\theta_i, y_i) = P(\mu_i \leq t|\theta_i, y_i).$$

Then the expression for the median is

$$\tilde{\mu}_i(\theta_i, y_i) = 1[c(\theta_i, y_i) \leq 0.5]\ G^{-1}_{\theta_i, y_i}\bigg(\frac{\frac{1}{2} - c(\theta_i, y_i)}{1 - c(\theta_i, y_i)}\bigg). \tag{5}$$

One could show that the posterior median defined in (5) is a continuous shrinkage soft thresholding rule.

Result 3: For the Strawderman-Berger prior for the normal mean when the true mean is constrained to the non-negative orthant, the component-wise posterior median in (5) satisfies the following properties for a given value of the hyperparameter $\boldsymbol{\theta}$.

1. The posterior $p(\mu_i|\theta_i, y_i)$ is stochastically increasing in y_i and hence the posterior median of μ_i is a monotonically increasing in y_i for each value of the hyperparameter.

2. For each y_i, there exists $T(\theta_i)$ such that $\tilde{\mu}_i(\theta_i, y_i) = 0$ iff $y_i < T(\theta_i)$.

See the appendix for a proof of the above result.

3. Posterior Computation and Numerical Results

Let $\boldsymbol{\theta}$ denote the set of all hyper-parameters. We use $\boldsymbol{\theta}$ interchangeably for horseshoe and Strawderman-Berger prior where $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \tau, \sigma\}$ for the former and $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \tau, p, \sigma\}$ for the later. We use a Metropolis within Gibbs algorithm to generate random samples from the marginal posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and thus compute posterior summaries for the posterior $\pi(\mu_i|\boldsymbol{y})$ by averaging the value of the hyperparameters over the randomly generated sample of $\boldsymbol{\theta}$. For the posterior mean $E(\mu_i|\boldsymbol{y})$ we use the estimator

$$E(\mu_i|\boldsymbol{y}) = L^{-1}\sum_{l=1}^{L} \mathrm{E}(\mu_i|\boldsymbol{\theta}_l, \boldsymbol{y}),$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ are samples from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. For the posterior median, $Med(\mu_i|\boldsymbol{y})$, we use the estimator

$$Med(\mu_i|\boldsymbol{y}) = \hat{F}_i^{-1}(1/2),$$

where $\hat{F}_i(t) = L^{-1}\sum_{l=1}^{L} P(\mu_i \leq t|\boldsymbol{y}, \boldsymbol{\theta}_l)$.

The conditional marginal of $\boldsymbol{y}$ can be factorized as

$$\pi(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i|\lambda_i, \tau, \sigma, p),$$

where $\pi(y_i|\lambda_i, \tau, \sigma) = \frac{1}{l_i}\phi(y_i/l_i)\ \Phi(m_i/s_i)$ for the horseshoe prior and $\pi(y_i|\lambda_i, \tau, p, \sigma) = \frac{p}{\sigma}\phi(y_i/\sigma) + \frac{2(1-p)}{l_i}\ \phi(y_i/l_i)\ \Phi\left(\frac{m_i}{s_i}\right)$ for the Strawderman-Berger prior. The distribution of y_i conditional on the hyperparameters is Skew-Normal for the horseshoe prior and a mixture distribution of Normals for $\mu_i = 0$ and Skew-Normal for $\mu_i > 0$.

Hence the for the Gibbs sampling algorithm, the full conditionals are

1. $\pi(\lambda_i|\tau, \sigma, \boldsymbol{y}) \propto \pi(y_i|\lambda_i, \tau, \sigma)\ \pi(\lambda_i), \quad i = 1, \ldots, n$
2. $\pi(\tau, \sigma|\boldsymbol{\lambda}, \boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda}, \tau, \sigma)\ \pi(\tau|\sigma)\ \pi(\sigma)$.

For the Strawderman-Berger prior we have in addition,

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

The one-dimensional conditionals can be sampled using a standard Metropolis step.

3.1. Simulation results

We compare the performances of the Strawderman-Berger estimators, horseshoe estimator and Maximum Likelihood Estimator (MLE) under different degrees 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 i.e. $\hat{\mu}_i = \max(y_i, 0)$. For a general polyhedral cones, with Σ other than $\sigma^2 I$, the MLE is not straightforward 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:

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

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 10 with variance 20. Two possible values of σ are considered: $\sigma = 1$ and $\sigma = 3$. The separation between y_i's at $\mu_i = 0$ and $\mu_i > 0$ is more prominent for $\sigma = 1$ than $\sigma = 3$.

Weak sparsity: For weakly sparse signals, we generate μ_i which decays according to the power law but none of its components are exactly zero. For this, we consider

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

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 $\alpha \sim \text{Unif}(0.5, 1)$ depending on 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. The convergence is assessed using the standard MCMC diagnostic checks and all chains seem to converge. 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 a single realization generated under strongly sparse signals with the variance set to $\sigma = 1$ or $\sigma = 3$. 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 $\alpha \sim \text{Unif}(1, 2)$.

From Figure 1 and Figure 2, we see that the posterior mean for the horseshoe provides shrinkage near zero, but it is still significantly positive even when the realized y is considerably negative. This is particularly undesirable in the constrained case when the

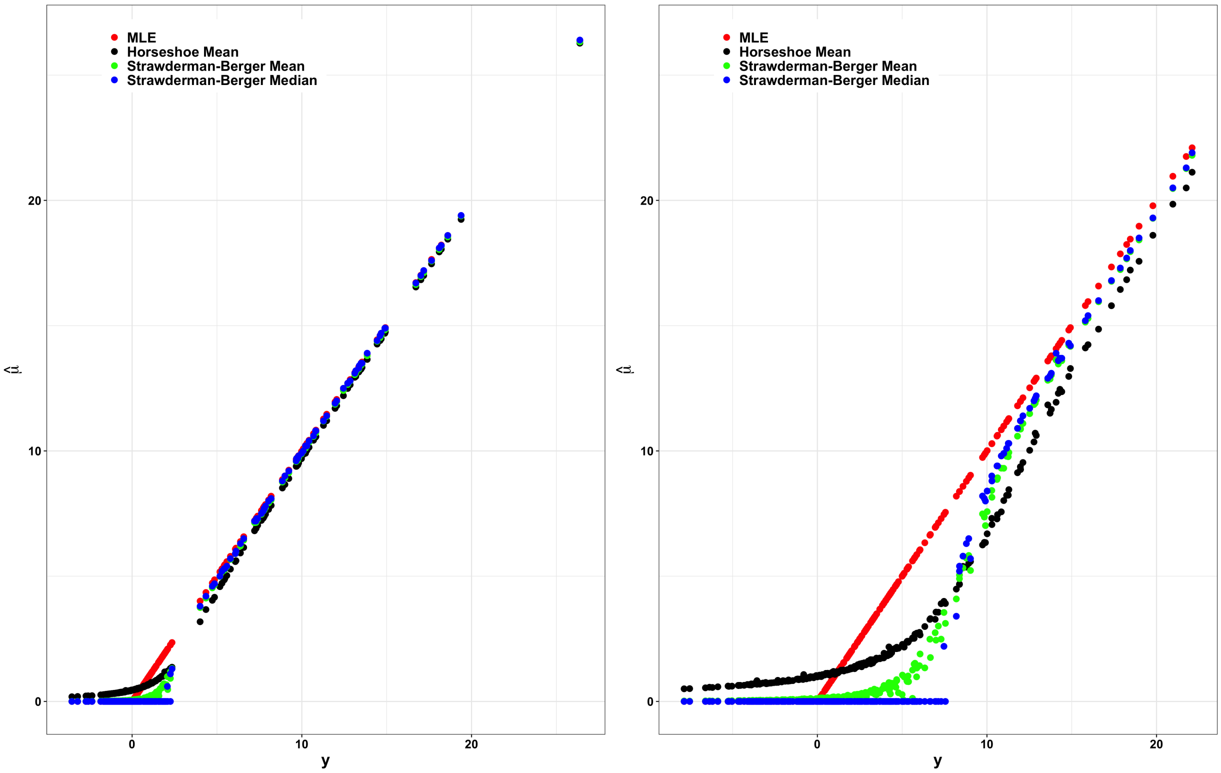

Figure 1: Plots of $\hat{\mu}$ versus y under strong sparsity with $\sigma = 1$ (left) and $\sigma = 3$ (right)

true mean is known to be non-negative. From Result 1, we know that for negative y, the horseshoe estimator decays as $O(|y|^{-1})$. This induces considerable bias. The posterior mean under the Strawderman-Berger prior shrinks more than the horseshoe posterior mean estimator. However, for large positive y the horseshoe estimator seems to perform better, and shrinks less than the posterior mean under the Strawderman-Berger prior.

The posterior median for Strawderman-Berger prior, as expected from the results in Result 3, provides a soft thresholding estimator that is truncated to zero below the truncation point $T(y, p)$ and provides continuous shrinkage for y above the truncation point. All estimators are monotonic in y and the shrinkage factor tends to one as y tends to infinity, thereby satisfying the requirement to not perturbing the big realized values of y.

Table 1 shows that the risk performance of Strawderman-Berger 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 Strawderman-Berger posterior mean and posterior median. Specifically, horseshoe posterior mean has of $6\% - 40\%$ more risk than the Strawderman-

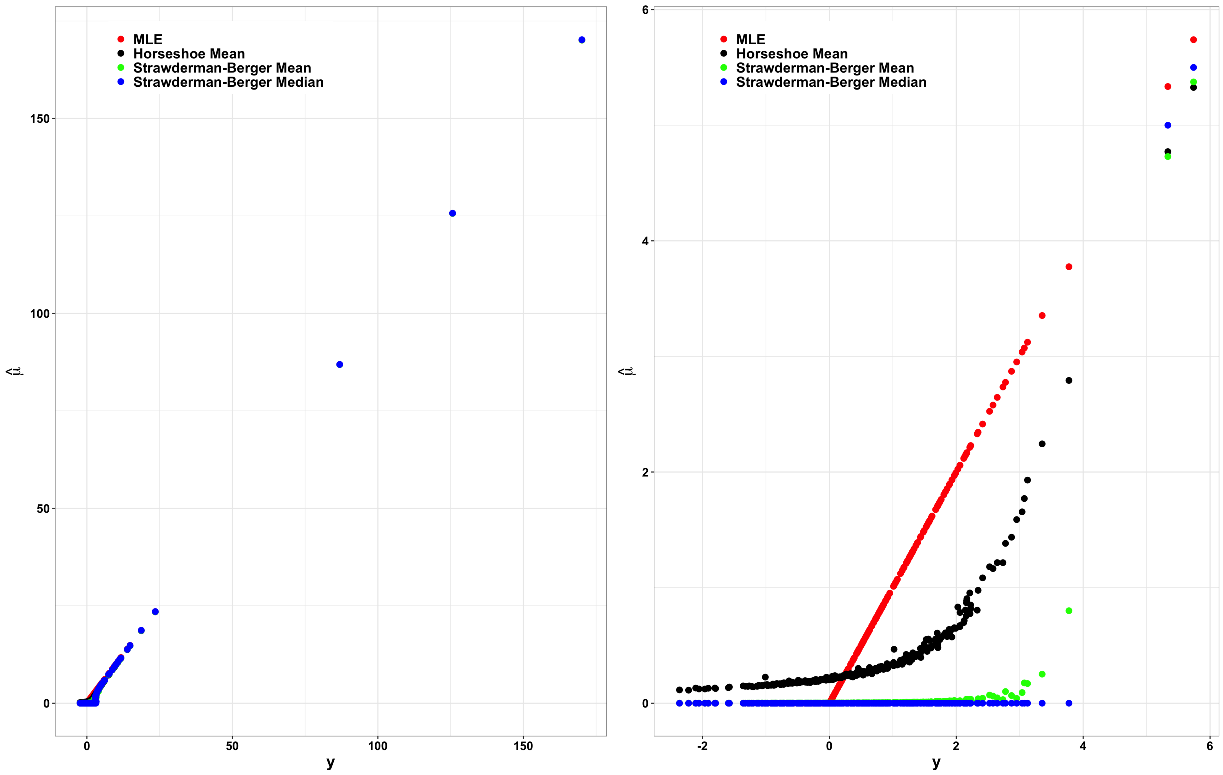

Figure 2: Plots of $\hat{\boldsymbol{\mu}}$ versus $\boldsymbol{y}$ under weak sparsity where $\alpha \sim U(0.5, 1)$ (left) and $\alpha \sim U(1, 2)$ (right)

Berger estimators. However, when $\alpha \sim U(1, 2)$, horseshoe estimator has 63% more risk than the Strawderman-Berger 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.

4. Real Data Analysis

We studied the performance of the estimators using the childhood acute lymphoblastic leukemia (ALL) data set (GSE412) which includes gene expression information for 110 childhood acute lymphoblastic leukemia samples before and after treatment. From the originally measured 12625 probe sets, genes that were not present in at least one sample were removed to obtain 8280 genes. After cleaning the data, we selected 250 genes for 50 pediatric newly diagnosed children for our analysis. Our goal is to estimate the standardized difference between post-treatment mean, $\boldsymbol{\theta_2}$ and pre-treatment mean, $\boldsymbol{\theta_1}$ regardless of the type of treatment used i.e. $\boldsymbol{\mu} = \frac{\boldsymbol{\theta_2} - \boldsymbol{\theta_1}}{\sigma}$. For illustration purposes, we assume up-regulation of gene expression level in ALL cells so that $\boldsymbol{\mu} \in \mathbb{R}^n_+$. We further assumed that the gene expression levels are uncorrelated and have same variance. The observed data is the standardized difference of the average post-therapy and pre-therapy gene expression.

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.

		$\sigma = 1$				$\sigma = 3$			
		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 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.

		$\alpha \sim U(0.5, 1)$				$\alpha \sim U(1, 2)$			
		MLE	HS	SB Mean	SB Median	MLE	HS	SB Mean	SB Median
Square Error Loss	MLE	200.24	2.68	2.91	2.65	179	15.7	128.67	400
	HS		122.52	0.73	0.6		63	1.63	5.11
	SB Mean			185.7	0.81			128	0.92
	SB Median				235.8				136
Absolute Error Loss	MLE	181.8	1.6	1.75	1.63	166	3.09	4.32	4.8
	HS		136.39	0.94	0.85		86	0.91	0.93
	SB Mean			162.67	0.89			128	0.95
	SB Median				186.75				134

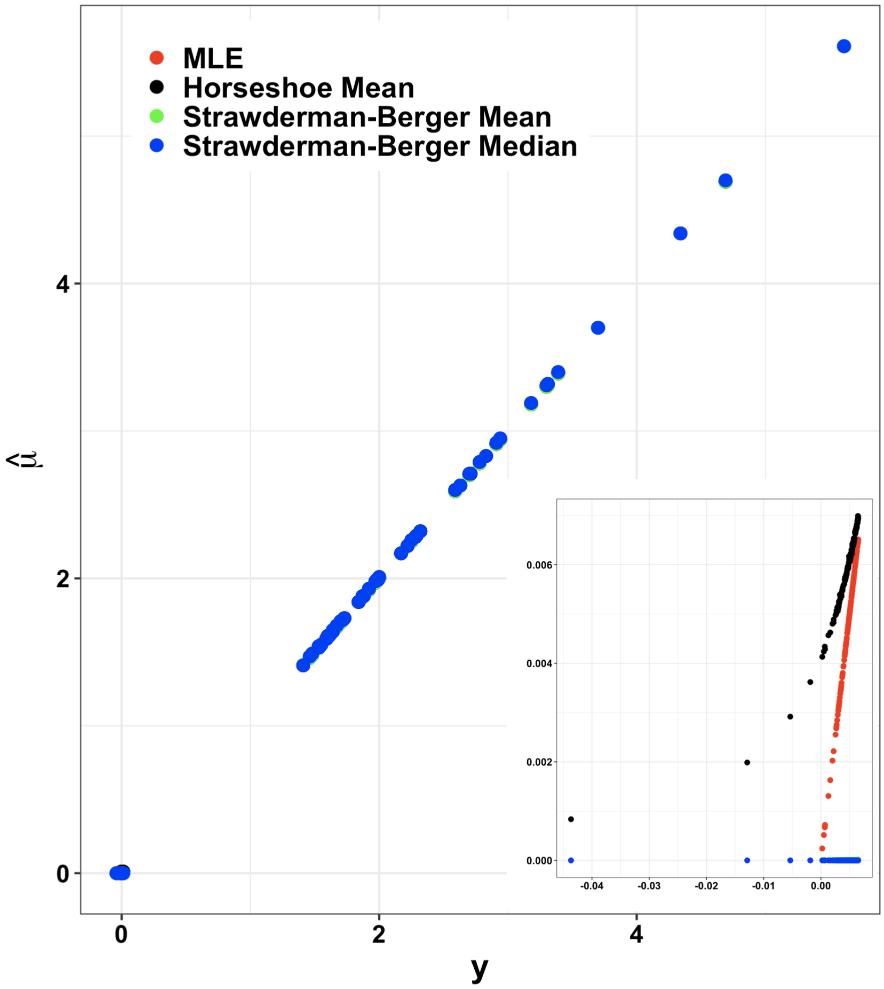

Figure 3: Plot of $\hat{\mu}$ versus y for the standardized difference in post-treatment and pre-treatment gene expression for 250 genes

The estimated MLE, posterior means and posterior median are shown in Figure 3. The summary results for the observed data and the estimates are presented in Table 3. We noticed that Horseshoe posterior mean is always positive whereas the MLE is 0 for negative y's. While the Strawderman-Berger posterior median is exactly 0 for $y < 1.4$, the SB posterior mean is close to 0 for these values of y. All four estimates perform similarly for larger values of y.

5. Discussion

In our simulation studies, we compared the performance of horseshoe posterior mean, Strawderman-Berger posterior mean and posterior median for strongly sparse signals and weakly sparse signals. While the posterior mean for both horseshoe and Strawderman-Berger prior are shrinkage estimators, MLE and Strawderman-Berger posterior median are truncation based estimators with exact zeros for small signals. When the true sparsity regime is strong sparsity, then truncation type estimators maybe preferred. The non-negative constraint does impact the relative performance of the mean and median estimators. It can be shown that the posterior mean under priors considered here are smooth differentiable functions of the observed value. Hence it cannot be

Table 3: Summary statistics for the estimates under Horseshoe (HS) and Strawderman-Berger prior

	$\boldsymbol{y}$	MLE	HS Posterior Mean	SB Posterior Mean	SB Posterior Median
Minimum	-0.0437	0.0000	0.0008	0.0000	0.0000
1st Quartile	0.0043	0.0043	0.0058	0.0000	0.0000
Median	0.0054	0.0054	0.0063	0.0000	0.0000
Mean	0.4618	0.4620	0.4630	0.4583	0.4592
3rd Quartile	0.0064	0.0064	0.0069	0.0000	0.0000
Maximum	5.6094	5.6094	5.6093	5.6094	5.6100

expected to capture the threshold like behavior present in the strongly sparse regime.

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 Bayesian lasso, with hard thresholding properties for non-negative mean vectors. Another interesting domain is the discrete mixture models where the mixing kernel for the positive means could be chosen in a more flexible manner, belonging to flexible families on the non-negative orthant, e.g. product of gamma densities where heavy tailed priors are used for the hyperparamters. 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. Moreover, the observations maybe allowed to be correlated with a known low-dimensional correlation structure. For example, one could use the model $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$ where the positive definite matrix $\boldsymbol{\Sigma}$ is completely known. The mean vector is again assumed to be in the non-negative orthant. A standard approach to dealing with general $\boldsymbol{\Sigma}$ matrix is to transform the observations to $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$ so that the problem reduced to the case considered here. However, the transformed mean $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$ need not remain in the positive orthant unless $\boldsymbol{\Sigma}$ is an M-matrix with an inverse that admits a positive square-root. Thus, for general $\boldsymbol{\Sigma}$ further investigation is required.

References

Berger, J. O. and Strawderman, W. E. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *Annals of Statistics*, **24(3)**, 931–951.

Brown, L. D. and Greenshtein, E. (2009). Non-parametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Annals of Statistics*, **37(4)**, 1685–1704.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97(2)**, 465–480.

Danaher, M. R., Roy, A., Chen, Z., Mumford, S. L. and Schisterman., E. F. (2012). Minkowski–Weyl Priors for models with parameter constraints: An analysis of the BioCycle study. *Journal of the American Statistical Association*, **107(500)**, 1395–1409.

Dharmadhikari, S. and Kumar, J. (1988). Choice of hierarchical priors: admissibility in estimation of normal means. *Unimodality, Convexity, and Applications*, Elsevier Science.

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

Fan, J. and Li, R. (2001). Variable Selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96(456)**, 1348–1360.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 361-379, University California Press, Berkeley.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B*, **72(2)**, 143–170.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high- dimensional settings. *Journal of the American Statistical Association*, **107(498)**.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, **32(4)**, 1594–1649.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, **83(404)**, 1023–1032.

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

Silvapulle, M. J. and Sen, P. K. (2011). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints.* John Wiley and Sons.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, **58(1)**, 267–288.

$https : //www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GPL8300$

APPENDIX

Here we prove the claims in Results 2 and 3 about the posterior mean and median of μ_i obtained using the Strawderman-Berger prior.

Result 2:

2. *$\hat{\mu}_i$ is non-decreasing in y_i.*

For notational simplicity, we denote μ_i by μ and y_i by y. Without loss of generality, let us assume $\sigma = 1$.

$$y|\mu \sim N(\mu, 1), \quad \mu \sim g(\mu)$$

$$g(\mu) = \pi\delta_o + (1-\pi)g_1(\mu)$$

An estimator of μ is then given by

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

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 $b'(y)=a(y)$

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

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

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

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

Result 3:

1. *The posterior $p(\mu_i|\theta_i, y_i)$ is stochastically increasing in y_i and hence the posterior median of μ_i is a monotonically increasing in y_i for a given value of the hyperparameter $\boldsymbol{\theta}$.*

For notational simplicity, we denote μ_i by μ and y_i by y. Without loss of generality, let us assume $\sigma=\tau=\lambda_i=1$ for $i=1,\ldots,n$. From the expression for the posterior of μ, we have,

$$\pi(\mu|y)=c(p,y)\delta_o(\mu)+(1-c(p,y))f(\mu|y),$$

where

$$\begin{aligned} f(\mu|y) &= h(y)^{-1}e^{\mu y}e^{-\frac{1}{2}\mu^2}g(\mu), \\ h(y) &= \int_0^\infty e^{\mu y}e^{-\frac{1}{2}\mu^2}g(\mu)d\mu, \\ c(p,y) &= [1+\frac{(1-p)}{p}h(y)]^{-1}. \end{aligned}$$

To show that $\pi(\mu|y)$ is stochastically increasing (SI) in y, it is enough to show $f(\mu|y)$ is SI in y since $c(p,y)$ decreases with decrease in y.

Let $\mu_1 < \mu_2$ and $y_1 < y_2$. Then,

$$\frac{f(\mu_1|y_1)f(\mu_2|y_2)}{f(\mu_2|y_1)f(\mu_1|y_2)} = e^{(\mu_2-\mu_1)(y_2-y_1)} \geq 1.$$

Thus,

$$f(\mu_1|y_1)f(\mu_2|y_2) \geq f(\mu_1|y_2)f(\mu_2|y_1).$$

Multiplying both sides by $\pi(y_1)\pi(y_2)$ where $\pi(y)$ is the marginal of y, we have,

$$f(\mu_1,y_1)f(\mu_2,y_2) \geq f(\mu_1,y_2)f(\mu_2,y_1).$$

Hence, $f(\mu,y)$ is Totally Positive of order 2 (TP_2). Hence, μ and y are SI in each other (Theorem 6.1, Dharmadhikari and Joag-Dev 1988).

2. *For each y_i, these exists $T(\theta_i)$ such that $\tilde{\mu}_i(\theta_i,y_i)=0$ iff $y_i < T(\theta_i)$.*

Since $c(\theta_i,y_i)$ is monotonically decreasing in y_i and

$$\begin{aligned} \lim_{y_i\to-\infty} c(\theta_i,y_i) &= 1, \\ \lim_{y_i\to\infty} c(\theta_i,y_i) &= 0. \end{aligned}$$

For each θ_i, $\exists \;\; \tilde{\mu}_i(\theta_i,y_i)=0 \iff y_i < T(\theta_i)$.