INDIAN STATISTICAL INSTITUTE

MASTERS THESIS

Some Contributions to Inference and Model Selection in High Dimensional Statistics

---

Author: Navonil Deb

Supervisor: Arijit Chakrabarti

A thesis submitted to Indian Statistical Institute
in partial fulfilment of the requirements for
the course of
M. Stat. Dissertation

Indian Statistical Institute Kolkata

June, 2021

Abstract

In high dimensional regime, the problem of shrinkage estimation plays an important role for inference and model selection purposes. When in a regression model the dimension p of parameters is larger than the sample size n, the two approaches that have drawn significant attention in recent times are penalized likelihood based methods and use of shrinkage priors. Among many variants of shrinkage priors available in the existing literature of statistics, a very useful class of shrinkage priors is the global-local scale mixtures where along with the global shrinkage parameter, one introduces the notion of a local parameter which is appropriately mixed such that the noise components of the regression coefficients are strongly shrunk yet the signal components are left almost unshrunk. The shrinkage estimation setting of our literature will concern the horseshoe prior of Carvalho et al. [2010], three-parameter Beta (TPB) prior of Armagan et al. [2011] and generalized double Pareto (GDP) prior of Armagan et al. [2013].

We consider the framework of Bhadra et al. [2019] and show that the horseshoe regression estimate outperforms the purely global estimates produced by ridge regression for a class of design matrices under some suitably posed restrictions on their singular values, depending on the global shrinkage parameters of both regression methods. To measure the relative efficacy of different shrinkage-based approaches, we use Stein's unbiased risk estimate (SURE) which unbiasedly estimates the predictive risk associated with the shrinkage method at hand. The hierarchy of global-local shrinkage method overcomes the sole dependence on only one global parameter controlling the component-wise shrinkage and the monotonicity of the relative shrinkage with the singular values of the design matrix. In addition, we aim to further generalize our findings on horseshoe prior to three parameter Beta and generalized double Pareto priors which are useful for inference and model selection in Bayesian paradigm.

Contents

1 Introduction

In high dimensional regime, where one often deals with "large p small n" setup, penalized likelihood-based approaches have been widely used and implemented for variable selection and prediction purpose. Another direction the statisticians have been exploring since past few decades is the use of shrinkage priors on the regression coefficient and inferences based on the posterior distribution, with applications in sparse problems. Although many methods have been proposed and the associated theory is now well developed, the relative efficacy of different approaches in finite-sample settings, as encountered in practice, is not well-understood in all cases. The problem of variable selection in high dimension has received notable interests in recent past for its wide range of applications in complex and high-throughput data sets arising in interdisciplinary fields, especially in biology and finance. However, the problem of variable selection in Bayesian paradigm is a challenging problem as, even though the methodological sophistication, the underlying computational issues make the task of variable selection difficult and daunting. This has motivated the use of continuous two-step (global-local) shrinkage priors which facilitate some nice statistical interpretations. Such two-step methodologies have been used in recent past by several authors viz. Polson and Scott [2010], Liang et al. [2008], Ghosh et al. [2016] and others. In the sparse setting of the regression parameters when its dimension is very large, exploiting such underlying prior structure often turns out to be useful for drawing meaningful inferences.

In high dimension, the regression model of our interest is following

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n) \tag{1}$$

where $y \in \mathbb{R}^n$, $X = [X_1 : \cdots : X_p] \in \mathbb{R}^{n\times p}$, $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$. $\hat{\beta}$ is the estimate of β based on observation y and fixed and non-stochastic design matrix X. In order to perform variable selection from the set of available predictors $X_1, \cdots, X_p$, a popular method is the use of two-component mixture priors, also referred to as the *spike and slab* priors (Mitchell and Beauchamp [1988], George and McCulloch [1993]). In usual settings, this kind of priors put a mass or a spike at zero along with a continuous density centered at zero (representing the signal density). Computational issues regarding Bayesian variable selection and the fact that many components of the parameter β might be zero has led to a wide variety of continuous shrinkage priors namely global-local (GL) scale mixture. As proposed by Polson and Scott [2010], the prior specification is given by

$$\beta_j \sim \mathcal{N}(0, \lambda_j^2\tau^2), \quad \tau \sim f, \quad \lambda_j \sim g \tag{2}$$

where f and g are two densities on the positive real line. The global parameter τ controls the global shrinkage towards the origin whereas the local parameter λ_j^2 allows variation in the degree of component-wise (or local) shrinkage. Some special cases of global-local shrinkage include Bayesian lasso by Park and Casella [2008], normal-gamma mixture by Griffin et al. [2010], horseshoe regression by Carvalho et al. [2009] etc. In subsequent sections of this report we will discuss more

on the predictive performance of horseshoe regression.

In this literature, we follow the model selection setup proposed by Bhadra et al. [2019] under a number of suitably chosen continuous shrinkage priors and we compare their predictive performances. The setup relies on a comparative study based on the predictive risk of different model selection criteria associated with the parameter β. For a future observation y^*, the predictive risk R is denoted as follows

$$R = \mathbb{E}_{[y^*,\ y|X,\beta]} \left(y^* - X\hat{\beta} \right)^2 \tag{3}$$

Several authors including Bhadra et al. [2019] focuses on comparing estimators $\hat{\beta}$ obtained from the underlying method according to the criteria (3) in a non asymptotic (fixed n, p) setting with $p > n$. Such criteria of comparing estimators and subsequent model selection process depends on the paradigm proposed by Stein [1981], namely *Stein's unbiased risk estimator* (SURE). Bhadra et al. [2019] identified the shortcomings of some commonly used global shrinkage priors where the amount of shrinkage is controlled by only one tuning (global) parameter. They provided a theoretical basis for the better predictive performance of horseshoe priors over ridge regression and selection based methods when the true β is sparse and robust. However, their theoretical finding on horseshoe regression outperforming ridge relies on some stringent conditions on the singular values of the design matrix X.

In this work, we propose some results which attempt to generalize the theoretical findings of Bhadra et al. [2019]. We find that under a more general set of conditions on the design matrix as well as the regression coefficients, the predictive performance of horseshoe can be guaranteed to be better than ridge. One of the many advantages of such results is that one needs not remain restricted to the case when all the singular values of the design matrix are 1 as assumed by Bhadra et al. [2019]. In our result, we get a set of ranges for the global shrinkage parameters of horseshoe as well ridge which, along with the restrictions on the design matrix, ensures that the Stein's risk for horseshoe is indeed less than that of ridge. In addition, we attempt to generalize our results from horseshoe regression to a setting incorporating slowly varying priors viz. generalized double Pareto Armagan et al. [2013] and three parameter Beta prior Armagan et al. [2011].

The rest of the article is organized as follows: in Section 2 we discuss the problem of our interest and review some relevant literature involving global-local shrinkage priors and horseshoe regression, Section 3 includes the main ideas and results we have derived. We conduct a comparative study of the global-local shrinkage estimates with some purely global estimates and compare their Stein's risk by empirical studies in Section 4. Conclusions then follow and the Appendix collects some of the figures and the mathematical proofs.

2 The problem setup and a review of relevant literature

We first describe the main setup and the problem at hand in the first few paragraphs in this section. For subsequent sections, we follow the notations followed by Bhadra et al. [2019]

The global shrinkage estimators such as ridge regression, principal component regression or PCR remain popular in prediction under the high-dimensional regression model (1). Shrinkage methods are more advantageous over simultaneous shrinkage and selection-based methods such as the lasso (Tibshirani [1996]) and outperform them in predictive performance in certain scenarios. On the theoretical side, Polson and Scott [2012], with the help of a representation proposed by Frank and Friedman (1993), showed that many popular high-dimensional shrinkage regression estimates, such as the estimates of ridge regression, PCR, regression with g-prior (Zellner [1986]) are essentially posterior means under a global shrinkage framework on the regression coefficients that are suitably orthogonalized. Polson and Scott [2012] also demonstrated two drawbacks of using purely global shrinkage estimators: (1) The amount of relative shrinkage is monotone in the singular values of the design matrix, (2) The shrinkage is determined by only a single tuning parameter hence component-wise freedom of shrinkage is not incorporated in prior specification.

Both of these factors can translate to poor out-of-sample prediction performance, which they demonstrated numerically. While they showed the outperforming of the horseshoe prior of Carvalho et al. [2010], the theoretical basis in support of such phenomena was provided by Bhadra et al. [2019]. We briefly discuss the underlying setup and the results of our interests.

2.1 Global shrinkage estimator as posterior mean

In the regression model (1), we let $X = UDW^\top$ be the singuar value decomposition of the design matrix X, with $U_{n\times n}$ and $W_{p\times n}$ satisfying $U^\top U = I$, and $W^\top W = I$. And D is a diagonal matrix containing the singular values corresponding to X with $D = \text{diag}(d_1, \cdots, d_n)$ with $d_1 \geq \cdots \geq d_n > 0$ and hence $\text{rank}(D) = n$. Let $Z = UD$ and $\alpha = W^\top \beta$ be the *orthogonalized* regression coefficient. The regression model (1) becomes

$$y = Z\alpha + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n) \tag{4}$$

The ordinary least square (OLS) estimate of α is $\hat{\alpha} = (Z^\top Z)^{-1} Z^\top y = (DU^\top UD)^{-1} DU^\top y = D^{-1}U^\top y$. As used by Frank and Friedman [1993], Polson and Scott [2012] and many other authors, many shrinkage estimators can be expressed in terms of the posterior mean of α under the following hierarchical model

$$(\hat{\alpha}_i |\, \alpha_i, \sigma^2) \sim \mathcal{N}(\alpha_i, \sigma^2 d_i^{-2}) \quad \text{independent} \tag{5}$$

$$(\alpha_i |\, \sigma^2, \tau^2, \lambda_i^2) \sim \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2) \quad \text{independent} \tag{6}$$

with $\sigma^2, \tau^2, \lambda_i^2 > 0$. τ^2 is the global shrinkage parameter and the local parameter λ_i^2 depends on the underlying shrinkage method being used. Using (5) and (6) the posterior mean of α is

$$\tilde{\alpha}_i = \frac{\tau^2\lambda_i^2 d_i^2}{1 + \tau^2\lambda_i^2 d_i^2}\hat{\alpha}_i \tag{7}$$

The special cases follow when the λ_i's are chosen according to the method at hand. Some of

the popular special cases are:

1. **Ridge Regression:** For all i, $\lambda_i = 1$ and $\tilde{\alpha}_i = \dfrac{\tau^2 d_i^2}{1+\tau^2 d_i^2}\hat{\alpha}_i$

2. **Regression with g-prior:** For all i, $\lambda_i^2 = d_i^{-2}$ and $\tilde{\alpha}_i = \dfrac{\tau^2}{1+\tau^2}\hat{\alpha}_i$

Therefore the choice of λ_i's determines the shrinkage method being used. To compare the predictive performance of the estimates Bhadra et al. [2019] used the paradigm used by Stein namely **Stein's unbiased risk estimate** (SURE) for prediction risk. We discuss this estimator in the following subsection.

2.2 Stein's unbiased risk estimate for global shrinkage regression

For the regression model (1), let $\hat{y} = X\hat{\beta} = Z\hat{\alpha}$ denote the fitted value of the response variable where $\hat{\beta}$ an $\hat{\alpha}$ are estimated values of the parameter. As pointed out by Stein that the fitted risk is an underestimate of the true risk, **Stein's unbiased risk estimate** (SURE) is given by

$$\text{SURE} = -n\sigma^2 + ||y - \hat{y}||^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} \tag{8}$$

where the second term $\displaystyle\sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$ is known as the degree of freedom (Efron [2004]). One can easily check that $\mathbb{E}(\text{SURE}) = R$, where R is as defined in Equation (3).

This can be an extremely useful tool. Aside from plainly estimating the risk of an estimator, we could also use it for model selection purposes: if our estimator depended on a tuning parameter $\theta \in \Theta$, denoted by $\hat{y}_\theta$, then we could choose this parameter to minimize SURE:

$$\theta^* = \underset{\theta\in\Theta}{\arg\min} \left\{ ||y - \hat{y}_\theta||^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \hat{y}_{\theta,i}}{\partial y_i}(y) \right\}$$

In the expression of SURE, the term $n\sigma^2$ does not involve X and y. Therefore, comparing the SURE for different methods for same regression setup and sample size n is essentially same as comparing the corresponding values of T. For model selection purposes, we focus more on the following statistic

$$T = \text{SURE} + n\sigma^2 = ||y - \hat{y}||^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$$

In our case, several authors have used Tweedie's formula which associates the the posterior mean with the marginal distribution of $\hat{\alpha}$ (detailed calculations provided in Bhadra et al. [2019]) and yields

$$T = \sigma^4 \sum_{i=1}^{n} d_i^{-2} \left(\frac{\partial}{\partial \hat{\alpha}_i} \log m(\hat{\alpha}_i) \right)^2 + 2\sigma^2 \sum_{i=1}^{n} \left(1 + \sigma^2 d_i^{-2} \frac{\partial^2}{\partial \hat{\alpha}_i^2} \log m(\hat{\alpha}_i) \right) \tag{9}$$

The expression for Stein's unbiased risk estimate is given by $\text{SURE} = \sum_{i=1}^{n} \text{SURE}_i$ where

$$T_i = \frac{\hat{\alpha}_i^2 d_i^2}{(1+\tau^2\lambda_i^2 d_i^2)^2} + 2\sigma^2 \frac{\tau^2\lambda_i^2 d_i^2}{1+\tau^2\lambda_i^2 d_i^2} \tag{10}$$

Some of the key observations on the expression of T_i are following:

1. For large values of $\hat{\alpha}_i^2 d_i^2$, the value of T_i increases. Increasing the value of τ^2 manages this issue but it also incurs a cost of $2\sigma^2$ in T_i even for the components having small values $\hat{\alpha}_i^2 d_i^2$.

2. As seen from (7), the amount of relative shrinkage $\tilde{\alpha}_i/\hat{\alpha}_i$ is monotone in τ^2 as well as d_i^2. For smaller values of τ^2 and d_i^2 the components are shrunk more.

2.3 Stein's unbiased risk estimate for global local shrinkage regression

The class of global local shrinkage priors as described by Polson and Scott [2010] associates to the following general hierarchical structure:

$$(\hat{\alpha}_i |\, \alpha_i, \sigma^2) \sim \mathcal{N}(\alpha_i, \sigma^2 d_i^{-2}) \quad \text{independent} \tag{11}$$

$$(\alpha_i | \sigma^2, \tau^2, \lambda_i^2) \sim \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2) \quad \text{independent} \tag{12}$$

$$\lambda_i \sim \pi(\lambda_i) \quad \text{independent} \tag{13}$$

$$(\tau^2, \sigma^2) \sim \pi(\tau^2, \sigma^2) \tag{14}$$

The above hierarchical formulation of prior on the regression coefficient is generally referred to as the global-local scale mixture of normal. The parameter τ^2 is the global parameter an the local parameter is λ_i^2.

Define $\kappa_i = 1/(1+\lambda_i^2\tau^2)$, which is is a random shrinkage coefficient and is interpreted as the weight that the posterior mean for α_i places on 0 after the $\hat{\alpha}_i$ has been observed. Then

$$\mathbb{E}(\alpha_i | y, \tau^2, \lambda_i^2) = \frac{\lambda_i^2\tau^2}{1+\lambda_i^2\tau^2} \cdot \hat{\alpha}_i + \frac{1}{1+\lambda_i^2\tau^2} \cdot 0 = (1-\kappa_i)\hat{\alpha}_i$$

Therefore, using the finiteness of κ_i ($\in [0,1]$) we can take the condtional expectation of κ_i and write the following

$$\mathbb{E}(\alpha_i | y) = (1 - \mathbb{E}(\kappa_i | y))\hat{\alpha}_i \tag{15}$$

By assigning a prior $\pi(\lambda_i^2)$ on the local parameter λ_i^2 we can check the behaviour of the shrinkage coefficient κ_i and have an understanding of how the underlying Bayesian model attempts on discerning the signal components and the noise terms. Examples of some commonly used global-local shrinkage prior include horseshoe prior, horseshoe+, generalized double Pareto, generalize beta prior and others.

The literature review of this report underscores on the available literature of horseshoe regression and makes a further attempt to generalize the existing theories of Stein's unbiased risk estimation

in case of horseshoe for slowly varying priors. Here we use the notation $f(x) \sim g(x)$ which means $\lim_{x\to\infty} f(x)/g(x) = 1$

Definition 1 (Slowly varying function)**.** *A real-valued function $L(\cdot)$ is called slowly varying if $\lim_{|x|\to\infty} L(tx)/L(x) = 1$ for all real $t > 0$*

Definition 2 (Generalized slowly varying prior)**.** *We call a prior distribution π on the the lrisk ocal parameter λ_i to be slowly varying prior if $\pi(\lambda_i) \sim (\lambda_i^2)^{a-1} L(\lambda_i^2)$ as $\lambda_i^2 \to \infty$ for slowly varying function $L(\cdot)$*

2.3.1 Stein's risk for horseshoe regression

For horseshoe priors, the prior on λ_i is chosen to be the standard half-Cauchy prior $\mathcal{C}^+(0, 1)$. That is, the prior specification on λ_i is

$$\pi(\lambda_i) = \frac{2}{\pi} \cdot \frac{1}{(1 + \lambda_i^2)} \mathbb{1}\{\lambda_i > 0\}$$

The posterior mean is then obtained from (15). Polson and Scott [2012] has observed that use of horseshoe prior in regression incurs less error than other competing approaches. An intuitive explanation in support of this phenomenon is that the large values of α_i remain un-shrunk in the posterior due to the heavy-tailed nature of the half Cauchy distribution while the global parameter τ^2 ensures a same amount of towards-zero shrinkage for all components of α. Also in this literature, τ^2 and σ^2 will be treated as constants and positive real numbers.

Figure 1 gives a comparative view of horseshoe prior along with the Laplace and Student-t prior. To use Laplace prior on the regression parameter is the Bayesian analogue of lasso method. We note that near zero the mass of horseshoe prior is less than Laplace and Student-t distribution, which results into heavier tails and larger signal components of α is left almost un-shrunk.

It is easy to check as well as done in Bhadra et al. [2019] that for horseshoe regression, $\pi(\lambda_i)$ is a slowly varying prior. Also, the horseshoe+ prior suggested by Bhadra et al. [2017] falls in the slowly varying framework and in this case $\pi(\lambda_i) \propto \log \lambda_i/(\lambda_i^2 - 1)$. Ghosh et al. [2016] showed that the generalized double Pareto prior and the three-parameter Beta prior fall in this framework too.

The calculation of T for horseshoe regression is not as straightforward as that of purely global shrinkage setting. The marginal of $\hat{\alpha}$ involve integrals without closed form expressions and involve moments of compound confluent hypergeometric (CCH) distribution. The detailed calculations are provided in Bhadra et al. [2019].

Definition 3 (Compound confluent hypergeometric distribution)**.** *The compound confluent hypergeometric (CCH) density is given by*

$$f_{CCH}(x; p, q, r, s, \nu\theta) = \frac{x^{p-1}(1 - \nu x)^{q-1}\{\theta + (1 - \theta)\nu x\}^{-r} \exp(-sx)}{B(p, q)H(p, q, r, s, \nu, \theta)}$$

for $0 < x < 1/\nu$, and the parameters satisfying the following: $p > 0,\ q > 0,\ r \in \mathbb{R}, s \in \mathbb{R}, 0 \le \nu \le 1$ and $\theta > 0$. $B(\cdot, \cdot)$ denotes the beta function and H is given by

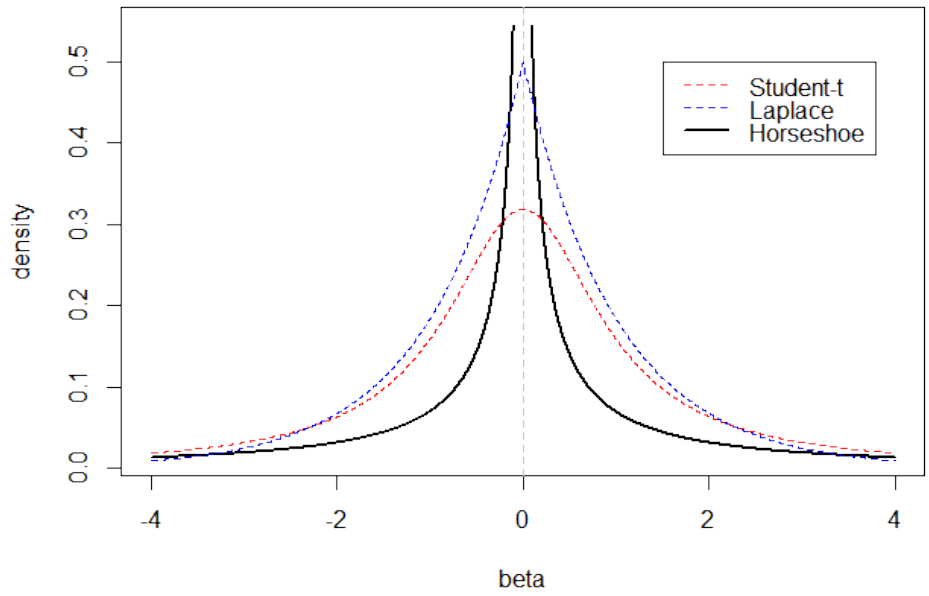

Figure 1: Horseshoe prior along with the student-t prior and Laplace prior on a regression parameter β

$$H(p,q,r,s,\nu,\theta) = \nu^{-1}\exp(-s/\nu)\Phi_1(q,r,p+q,s/\nu,1-\theta)$$

where Φ_1 *is the confluent hypergeometric function of two variables, given by*

$$\Phi_1(\alpha,\beta,\gamma,x_1,x_2) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(\alpha)_m(\beta)_n}{(\gamma)_{m+n}m!n!}x_1^m x_2^n$$

Here $(a)_k$ *denotes the rising factorial, i.e.* $(a)_0 = 1,\ (a)_k = (a+k-1)(a)_{k-1}$.

Using the moments of CCH distribution, the component-wise T_i's for horseshoe regression can be expressed. The theorem in Bhadra et al. [2019] for this expression in definition 3 states the following:

Theorem 1. *Denote* $m'(\hat{\alpha}_i) = (\partial/\partial\hat{\alpha}_i)m(\hat{\alpha}_i)$ *and* $m''(\hat{\alpha}_i) = (\partial^2/\partial\hat{\alpha}_i^2)m(\hat{\alpha}_i)$.

(a) The expression of T *for the Equations* (11)–(13) *is* $T = \sum_{i=1}^{n} T_i$, *where the contribution of* i^{th} *component is given by*

$$T_i = 2\sigma^2 - \sigma^4 d_i^{-2}\left[\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)}\right]^2 + 2\sigma^4 d_i^{-2}\frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} \tag{16}$$

(b) In horseshoe regression, under independent half-Cauchy prior on λ_i, the second and the third terms can be expressed as following

$$\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{\hat{\alpha}_i d_i^2}{\sigma^2}\mathbb{E}(Z_i), \quad \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{d_i^2}{\sigma^2}\mathbb{E}(Z_i) + \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}\mathbb{E}(Z_i^2)$$

where $(Z_i|\hat{\alpha}_i, \sigma, \tau) \sim CCH(p = 1, q = 1, r = 1/2, s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2, v = 1, \theta_i = (\tau^2 d_i^2)^{-1})$

Therefore, part (b) of Theorem 1 provides an explicit expression in terms of the moment of CCH distributions to calculate the T for horseshoe regression. Using the aforementioned expression Bhadra et al. [2019] suggests upper and lower bounds on the ratio $T_i/2\sigma^2$ when $|\hat{\alpha}_i|$ is large. In such case, Tweedie's formula used by Pericchi and Smith [1992] gives $\tilde{\alpha}_i = \hat{\alpha}_i + \sigma_2 d_i^{-2} m'(\hat{\alpha}_i) \approx \hat{\alpha}_i$. The theorem is as follows:

Theorem 2. *Define* $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2, \theta_i = (\tau^2 d_i^2)^{-1}$. *Then for* $s_i \geq 1$ *and* $\theta_i \geq 1$ *the following holds for the expression of* T *in horseshoe regression*

$$\left[1 - \theta_i(\tilde{C}_1 + \tilde{C}_2) \cdot \frac{1+s_i}{s_i^2} - \theta_i^2(\tilde{C}_1 + \tilde{C}_2)^2 \cdot \frac{(1+s_i)^2}{s_i^3}\right] \leq \frac{T_i}{2\sigma^2} \leq \left[1 + 2\theta_i(1+s_i)\left(\frac{C_1}{s_i^2} + \frac{C_2}{s_i^{3/2}}\right)\right]$$

where $C_1 \approx 3.53$, $C_2 = 16/15$, $\tilde{C}_1 \approx 1.95$ *and* $\tilde{C}_2 = 4/3$.

Remark 1. In theorem 2, if $|\hat{\alpha}_i| \to \infty$ that is $s_i \to \infty$, then for $\tau^2 \leq d_i^{-2}$, $T_i \to 2\sigma^2$. One interpretation of this result is that the heavy-tailed nature of λ_i's, which restricts the signal components of $\hat{\alpha}$ from being too shrunk, contributes a large bias in its T. This is an improvement over the purely global shrinkage method. The same result is in fact true for a general slowly varying priors as stated (and proved) in Theorem 4.2 of Bhadra et al. [2019].

Theorem 3. *Consider the hierarchy Equations* (11)–(13) *and* $\pi(\lambda_i) \sim (\lambda_i^2)^{a-1} L(\lambda_i^2)$ *as* $\lambda_i^2 \to \infty$ *where* $L(\cdot)$ *is a slowly varying function and* $a \leq 0$. *Then* $SURE_i \to 2\sigma^2$ *almost surely as* $s_i \to \infty$.

Theorem 3 ensures an almost sure convergence of the Stein's risk for slowly varying priors as $|\hat{\alpha}_i|$ are large. Thus when α is dense i.e. most of its components have large signals, intuitively the value of T should not have a large deviation from $2n\sigma^2$. Also, it is easy to show that the expected predictive risk for ordinary least square solution of the regression model (4) is exactly $2n\sigma^2$. The natural questions arise- (a) What restrictions should we impose on the regression model such that T is strictly smaller than $2n\sigma^2$ and accordingly horseshoe performs better than the least square estimate? and (b) How should the shrinkage parameter of a global method should be set such that the method at hand performs at least as the least square estimates in terms of Stein's risk?

The next subsection attempts to discuss the aforementioned questions through a comparative study of the prediction risk for purely methods and horseshoe regression.

2.4 Comparison of prediction risk for global and horseshoe regression

As pointed out in the theorems 4.2 and 4.4 of Bhadra et al. [2019] and subsequent discussions, horseshoe regression is very effective for handling the prediction risk when $\hat{\alpha}_i^2 d_i^2$ is either very small or very large. They provide a theoretical proof of the better performance of horseshoe regression over global shrinkage methods. The next result is pertinent to the prediction risk comparison and also stated and proved in Bhadra et al. [2019].

Theorem 4. *Let $X^\top X = I$, that is $D = I$ and the global parameter of horseshoe regression $\tau^2 = 1$. In a sparse situation i.e. true $\alpha_i = 0$, an upper bound on the component-wise prediction risk of horseshoe regression is approximately $1.75\sigma^2 < 2\sigma^2$.*

The above theorem 4 has the following indication: the prediction risk of horseshoe regression with τ^2 outperforms the optimal ridge regression, i.e. the global parameter of ridge is $\tau = \tau^*$. Therefore, clearly optimal horseshoe estimate will outperform the optimal ridge estimate. Bhadra et al. [2019], using simulations and comparing SURE (accordingly T) and SSE, have shown that for sparse and robust α horseshoe regression performs better than others while in case of dense α the ridge regression has a better record of predictive performance and selection-based methods are preferred in case of null α.

3 Main ideas and results

In this section, we state and motivate the main results of this article. As of now, our works are based on the predictive performance of horseshoe regression for a general design matrix X and the scopes of using a general slowly varying prior when α is sparse and robust. We first try to generalize the Theorem 4 when d_i's are more general than all being equal to 1. Using that we prove a theorem that states under certain assumptions, horseshoe can outperform ridge and least square estimate when the global shrinkage parameters of horseshoe and ridge are suitably tuned. Such results can be very helpful when one needs recommendations for the range of both shrinkage parameters in order to ensure a better performance of horseshoe. For other slowly varying priors including three-parameter Beta and generalized double Pareto for global-local shrinkage, we give a background review of the Stein's risk calculations and some simulation studies in Section 4.

3.1 Predictive performance of horseshoe regression for general design

We make an attempt to loose the condition on singular values of the design matrix X as in theorem 4. Our goal is to find theoretical evidence of the better performance of horseshoe estimate over the optimal ridge estimate for cases with true $\alpha_i = 0$. First we will propose an upper bound on the SURE of horseshoe estimate for a general design X and later, for this case, we will investigate whether the component-wise limiting risk of the ridge estimate is $2\sigma^2$.

Theorem 5. *Assume that $D = diag(d_1^2, \cdots, d_n^2)$ with $d_i > 0$ for all i and $d_i^2 \leq \tau^{-2}$. Denote $\theta_i = (\tau^2 d_i^2)^{-1}$. Then in a sparse situation i.e. true $\alpha_i = 0$, the following holds:*

(a) There exists constants $C_0, C_1, C_2, C_3 > 0$ *such that*

$$\mathbb{E}_{\alpha_i}(T_i) \leq \sigma^2 \left(C_0 + C_1\theta_i - \frac{C_2}{\theta_i} - \frac{C_3}{\theta_i^2} \right)$$

where $C_0 \approx 1.53,\ C_1 \approx 0.39,\ C_2 \approx 0.11,\ C_3 \approx 0.067.$

(b) For $\eta^{-1}\tau^{-2} < d_i^2 \leq \tau^{-2}$ *where* $\eta \approx 1.475$, $Risk_i \approx 1.743\sigma^2 < 2\sigma^2$.

Remark 2. If we put $\theta_i = 1$, the right hand side of the inequality in part (a) approximately becomes $1.743\sigma^2$ which means we almost retrieve the upper bound proposed in Theorem 5.1 of Bhadra et al. [2019]. And in the Theorem 5 we get an upper bound on the risk of horseshoe regression only in terms of the singular values of X and the global shrinkage parameter τ^2. As τ^2 increases, θ_i is lowered and the upper bound on $\mathbb{E}_{\alpha_i}(T_i)$ decreases accordingly. The proof is given in the appendix at A.2.

The condition on the part (b) of Theorem 5 regarding eigen values of X can alternatively written as $\eta^{-1}d_i^{-2} < \tau^2 \leq d_i^{-2}$. Thus depending on X, we get a range for the global shrinkage parameter for which the risk of horseshoe regression for a sparse component is strictly less than $2\sigma^2$.

3.2 Comparison with predictive risk of ridge regression

Bhadra et al. [2019] have shown that the optimal horseshoe estimator beats the optimal ridge estimator in terms of predictive performance when $X^\top X = I$. In order to check the same for the case of general X, we investigate the asymptotic behaviour of the risk of ridge regression along the same line of Bhadra et al. [2019].

To generalize our result for optimal ridge regression and horseshoe regression, we attempt to follow the line of works in Bhadra et al. [2019]. For ridge regression $\lambda_i = 1$ for all i, and we have

$$T_i = \frac{\hat{\alpha}_i^2 d_i^2}{1 + \tau^2 d_i^2} + 2\sigma^2 \frac{\tau^2 d_i^2}{1 + \tau^2 d_i^2}$$

Therefore,

$$T = \sum_{i=1}^{n} T_i = \sum_{i=1}^{n} \left[\frac{\hat{\alpha}_i^2 d_i^2}{1 + \tau^2 d_i^2} + 2\sigma^2 \frac{\tau^2 d_i^2}{1 + \tau^2 d_i^2} \right]$$

For the case $X^\top X = I$ we have $d_i = 1$ for all i. In that case the expression of SURE can be minimized with respect to τ^2 and the optimal ridge estimator is following:

$$\alpha_i^* = \min_{\tau} \sum_{i=1}^{n} T_i = \left(1 - \frac{n\sigma^2}{\sum_{i=1}^{n} \hat{\alpha}_i^2} \right) \hat{\alpha}_i$$

One interesting observation is that the optimal ridge estimator is no longer linear in $\hat{\alpha}_i$'s but is similar in expression to the James-Stein estimate of α. The shrinkage of $\hat{\alpha}_i$ is not only dependent on the individual components, it depends on $\sum_{i=1}^{n} \hat{\alpha}_i^2$. From Casella and Hwang [1982] one gets

the following:

$$1 - \frac{n-2}{n+||\alpha^2||} \leq \frac{R(\alpha, \alpha^*)}{R(\alpha, \hat{\alpha})} \leq 1 - \frac{(n-2)^2}{n}\left(\frac{1}{n-2+||\alpha||^2}\right)$$

And if $||\alpha||^2/n \to K$ for some positive constant K as $n \to \infty$, then

$$\lim_{n\to\infty} \frac{R(\alpha, \alpha^*)}{R(\alpha, \hat{\alpha})} = \frac{K}{K+1}$$

Therefore, for a 'large' value of K the risk associated with ridge regression and ordinary least square becomes asymptotically same. The natural question that we can pose in this regard is- can we do a similar calculation for general d_i's and obtain suitable conditions on X and shrinkage parameters such that horseshoe outperforms ridge?

We first try to follow the similar approach by looking for a minimizer $\tau^{*2} = \arg\min_{\tau^2} \sum_i T_i$. The first order condition is following

$$\frac{\partial T_i}{\partial(\tau^2)}(\tau^{*2}) = \sum_{i=1}^{n}\left[\frac{-2\hat{\alpha}_i^2 d_i^2}{(1+\tau^{*2}d_i^2)^3} + \frac{2\sigma^2}{(1+\tau^{*2}d_i^2)^2}\right]$$

Apart from numerical methods, there is no immediate way of finding the optimal τ^2 from the above equation unless all d_i's are equal to a constant. To this end, we state a set of theoretical results which ensure that it is possible to at least get a range of local shrinkage parameters for both horseshoe and ridge, and also restrictions on d_i's such that the average risk of horseshoe is strictly smaller than that of ridge. Such results attempt to generalize the theoretical foundations of Bhadra et al. [2019] for a slightly unrestricted setup than the very stringent case where $d_i = 1$ for all i. We first state the following lemma:

Lemma 1. *For horseshoe regression,* $\mathbb{E}_\alpha(T_i) \to 2\sigma^2$ *as* $|\alpha_i| \to \infty$.

Remark 3. The detailed proof is given in A.3. The significance of this lemma is that for components having larger signals, the expected value of T_i is guaranteed to converge to $2\sigma^2$. For sparse components, by virtue of Theorem 5, the expected value of T_i is strictly smaller than $2\sigma^2$. In this way the average of T_i's can be made strictly smaller than $2\sigma^2$.

The next theorem stands for a generalization of Theorem 4 where for a range of the shrinkage parameters of horseshoe and ridge we can guarantee than average risk of horseshoe will be strictly smaller than that of ridge.

Theorem 6. *Let α be such that finitely many of its components are non-zero and rests are equal to zero. In addition, for the non-zero components of α, $|\alpha_i| \to \infty$ as $n \to \infty$ in such a way, that $||\alpha||^2/n \to K\sigma^2$ as $n \to \infty$ for a sufficiently large positive constant K.*

Let τ_H^2 and τ_R^2 be the global shrinkage parameters associated with horseshoe and ridge regression respectively.

For $1.743 < A < 2$, define $\eta(A) = \sup\left\{\theta : C_0 + C_1\theta - C_2\theta^{-1} - C_3\theta^{-2} \leq A\right\}$ where the constants C_0, C_1, C_2 and C_3 are as in Theorem 1. Also τ_H satisfies the condition:

$$\frac{1}{\eta(A)\min_i d_i^2} \leq \tau_H^2 \leq \frac{1}{\max_i d_i^2}$$

Then for any $\delta \in (0, 2-A)$ and sufficiently large n

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(T_i^H)) < (A+\delta)\sigma^2 < 2\sigma^2 < \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(T_i^R)$$

$$\text{if } 0 < \tau_R^2 \leq \tau_H^2\left[\frac{1+\dfrac{K}{2\eta(A)\tau_H^2}}{B(A)+\dfrac{1}{\eta(A)}} - 1\right] \text{ with } B(A) = \sqrt{\frac{1}{\eta^2(A)} + 2\left(1-\frac{1}{\eta(A)}\right)\left(1+\frac{K}{2\eta(A)\tau_H^2}\right)}$$

Remark 4. The proof of the theorem is available in A.4 of Appendix. The result we have proven assumes a restriction of α being sparse and robust that in a sense, α is highly sparse for large n, however a finitely many of them are growing to infinity in such a way that $||\alpha||^2/n$ is close to $K\sigma^2$ for a large K. That is, either α_i is sparse or is growing to infinity. If A is very close to 2, we shall require a sufficiently large n such that the average risk of horseshoe and that of ridge bounded by $2\sigma^2$ from above and below respectively. In other words, to ensure that horseshoe regression outperforms ridge in this case we can approximate $\eta(A)$ by $\eta(2)$ for practical purposes when A is very close to 2. The rationality for this approximation is that $\eta(A)$ is a continuous and increasing function of A. The upper bound for τ^R can also be approximated by replacing $\eta(A)$ by $\eta(2) \approx 1.475$.

We note that the first term inside the bracketed expression of the upper bound of τ_R^2 involves K. If K is sufficiently large, the first term can be ensured to be strictly larger than 1 as the numerator of the first term is linear in K. However, the denominator is $O(\sqrt{K})$ and grows slower with K than the numerator.

3.3 Stein's risk for three-parameter Beta regression

3.3.1 Background

Three-parameter Beta prior is another very useful slowly varying prior meant for Bayesian mode selection. In short, it is a generalization of the horseshoe prior specification. For horseshoe, we consider the prior specification Equations (11)–(13) with $\pi(\lambda_i)$ being a standard haf-Cauchy density.

We reparametrize λ_i's as $\kappa_i := \dfrac{1}{1+\tau^2\lambda_i^2}$, that is the shrinkage coefficient defined in Section 2.3. Then it is easy to check

$$\alpha_i|\kappa_i \sim \mathcal{N}\left(0, \frac{1}{\kappa_i} - 1\right) \quad \pi(\kappa_i|\tau) \propto \alpha_i^{-1/2}(1-\alpha_i)^{-1/2}\frac{1}{1+(\tau^2-1)\alpha_i}$$

When $\tau^2 = 1$, the standard horseshoe prior induces a Beta $(1/2, 1/2)$ distribution on the shrinkage coefficient κ_i. The distribution of κ_i essentially determines how strong a noise component of α is pulled towards zero or the signal components remain un-shrunk. The natural question that arises here is- what happens if we suitably vary the speed of shrinkage by changing the parameters a, b from (1/2, 1/2) to any other values in the range (0, 1)? A larger value of a or b would suggest a lower mass near 0 or 1 respectively, indicating a lower speed of shrinkage towards 0 or retreat towards 1. Armagan et al. [2011] have implemented this idea in form of a class of prior called three-parameter Beta distribution.

3.3.2 Prior specification

As specified in Armagan et al. [2011], the hierarchy is as follows:

$$\hat{\alpha}_i|\alpha_i, \tau^2, \lambda_i^2 \sim \mathcal{N}(0, \sigma^2 d_i^{-2}) \text{ independent} \tag{17}$$

$$\alpha_i|\tau^2, \lambda_i^2 \sim \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2) \tag{18}$$

$$\lambda_i^2 \sim \pi(\lambda_i^2) \text{ independent} \tag{19}$$

$$\text{such that} \quad \pi(\lambda_i^2) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(\lambda_i^2)^{a-1}(1+\lambda_i^2)^{-(a+b)} \quad (a, b > 0)$$

Here, the distribution $\pi(\lambda_i^2)$ is an inverted Beta distribution with parameters a and b. The marginal distribution of α_i is called a *three parameter Beta-Normal (TPBN) mixture*. A detailed description of the other forms of the prior specification and an MCMC-based estimation approach is provided in Armagan et al. [2011]. It is easy to calculate and verify that the density of the shrinkage coefficient κ_i is proportional to a Beta(b, a) density.

Figure 2 shows a comparative view of three-parameter Beta prior for different combination of the hyper-parameters. The red curve essentially corresponds to the horseshoe prior. As a increases, the mass near 1 is higher and near zero, it requires higher value of b for a speedy shrinkage.

The next result supplies an explicit formula which we can use for risk computation of three-parameter Beta regression. By virtue of being a slowly varying prior, the results we have derived for horseshoe can be mimicked for TPB regression also. As we shall observe in Section 4, the optimal values of T_i's in each cases have very less significant difference for most of the practical purposes.

Theorem 7. *Let the hierarchy associated with the three-parameter Beta distribution be specified by Equations* (17)–(19). *Also define* $s_i = \hat{\alpha}_i^2 d_i^2/\sigma^2$. *Then the formula for* T *is given by* $T = \sum_{i=1}^n T_i$ *where*

$$T_i = 2\sigma^2[1 - \mathbb{E}(Z_i) + 2s_i\mathbb{E}(Z_i^2) - s_i\{\mathbb{E}(Z_i)\}^2]$$

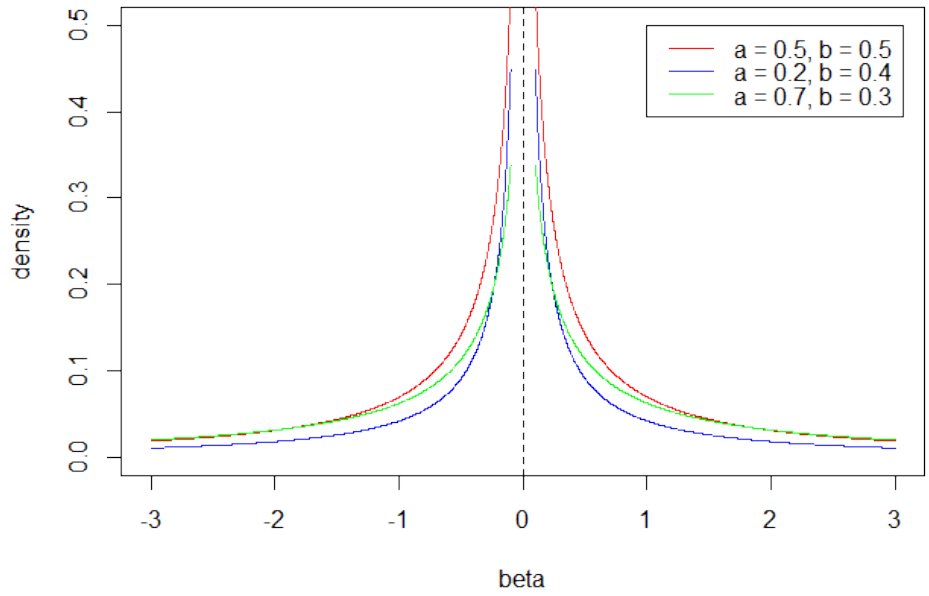

Figure 2: Three parameter beta prior family distributions for different values of a and b

$$\textit{where } Z_i \mid \hat{\alpha}_i, \sigma, \tau \sim CCH\left(b+\frac{1}{2}, a, a+b, \frac{\hat{\alpha}_i^2 d_i^2}{\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)$$

Remark 5. The result is a generalization of Theorem 1 and we use the formula in Theorem 7 in our simulation studies. A special case is horseshoe where $a = 1/2,\ b = 1/2$. The proof in Appendix A.5 seeks some of the ideas from the proof of Theorem 1 of Bhadra et al. [2019]. However, the fact that we arrive in an expression of T_i which again involves CCH distributions upon generalizing from horseshoe to three-parameter Beta regression, cannot be observed immediately and requires a proof. Theorem 7 is able to provide an answer in that direction. We also note in this context that the formula of Stein's risk for TPB regression has a very close connection with the horseshoe setup. This gives us a hint that similar calculations as Theorem 5 and 6 can possibly be carried out if we use three-parameter Beta prior instead of horseshoe.

3.4 Stein's risk for generalized double Pareto regression

3.4.1 Prior specification

Another variant from the class of general slowly varying priors is the *generalized double Pareto (GDP) prior*. According to Armagan et al. [2013], the GDP density is given by

$$f(\theta|\xi,\gamma) = \frac{1}{2\xi}\left(1+\frac{|\theta|}{\gamma\xi}\right)^{-(\gamma+1)} \tag{20}$$

where $\xi > 0$ is a scale parameter and $\gamma > 0$ is a shape parameter. Now let us consider the following hierarchy of distributions:

$$\hat{\alpha}_i|\alpha_i, \sigma^2 \sim \mathcal{N}(\alpha_i, \sigma^2 d_i^{-2}) \qquad \text{independent} \tag{21}$$

$$\alpha_i|\sigma^2, \tau^2, \lambda_i \sim \mathcal{N}(0, \sigma^2\tau^2\lambda_i) \qquad \text{independent} \tag{22}$$

$$\lambda_i \sim \text{Exponential}(\delta_i^2/2) \tag{23}$$

$$\delta_i \sim \text{Gamma}(a, b) \tag{24}$$

If one calculates the marginal of α_i from the set of equations Equations (21)–(24) then the prior density turns out to be

$$\pi(\alpha_i) = \frac{a}{2b\sigma\tau}\left(1 + \frac{|\alpha_i|}{b\sigma\tau}\right)^{-(a+1)} \tag{25}$$

That is $\alpha_i \sim \text{GDP}(b\sigma\tau/a, a)$. With this prior specification at hand, our next objective is to compute the Stein's risk estimate or accordingly T. The formula we have extensively followed so far for such computations comes from the part (a) of Theorem 1. To handle the terms involving the marginal of $\hat{\alpha}_i$ in the formula, we need to calculate the following integral

$$\begin{aligned} m(\hat{\alpha}_i) &= \int_{-\infty}^{\infty} \mathcal{N}(\hat{\alpha}_i; \alpha_i, \sigma^2 d_i^{-2})\ \pi(\alpha_i)\ \mathrm{d}\alpha_i \\ &= \int_{-\infty}^{\infty} \frac{d_i}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{d_i^2}{2\sigma^2}(\hat{\alpha}_i - \alpha_i)^2\right\}\ \frac{a}{2b\sigma\tau}\left(1 + \frac{|\alpha_i|}{b\sigma\tau}\right)^{-(a+1)}\ \mathrm{d}\alpha_i \end{aligned} \tag{26}$$

As the last integral does not have a closed form expression, we need some assistance of numerical computational methods which we discuss next.

3.4.2 Monte Carlo integration for Stein's risk computation

Consider the following function

$$f_{\alpha_i}(\hat{\alpha}_i) = \frac{d_i}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{d_i^2}{2\sigma^2}(\hat{\alpha}_i - \alpha_i)^2\right\} = \phi(\hat{\alpha}_i; \alpha_i, \sigma^2 d_i^{-2})$$

where $\phi(x; \mu, \sigma^2)$ denotes the density of a $\mathcal{N}(\mu, \sigma^2)$ random variable. Then the integral in Equation (26) can be reiterated as

$$m(\hat{\alpha}_i) = \int_{-\infty}^{\infty} f_{\alpha_i}(\hat{\alpha}_i)\pi(\alpha_i)\ \mathrm{d}\alpha_i = \mathbb{E}(\phi(\hat{\alpha}_i; Y, \sigma^2 d_i^{-2})) \tag{27}$$

where $Y \sim \text{GDP}(b\sigma\tau/a, a)$. To calculate the other derivative terms of $m(\hat{\alpha}_i)$ we will interchange the differentiation and integration. To check whether it can be indeed done, we differentiate f as follows:

$$f'_{\alpha_i}(\hat{\alpha}_i) = -\frac{d_i^3(\hat{\alpha}_i - \alpha_i)}{\sigma^3\sqrt{2\pi}} \exp\left\{-\frac{d_i^2}{2\sigma^2}(\hat{\alpha}_i - \alpha_i)^2\right\} = -(\hat{\alpha}_i - \alpha_i)\frac{d_i^2}{\sigma^2} f_{\alpha_i}(\hat{\alpha}_i)$$

We note that both $f_{\alpha_i}(\hat{\alpha}_i)$ and $f'_{\alpha_i}(\hat{\alpha}_i)$ are continuous in both α and $\hat{\alpha}_i$ in the whole $\mathbb{R}^2$. Therefore, we are allowed to change the derivative with integration to get the following expressions:

$$m'(\hat{\alpha}_i) = \int_{-\infty}^{\infty} f'_{\alpha_i}(\hat{\alpha}_i)\pi(\alpha_i)\ \mathrm{d}\alpha_i = -\mathbb{E}\left[(\hat{\alpha}_i - Y)\frac{d_i^2}{\sigma^2}\phi(\hat{\alpha}_i; Y, \sigma^2 d_i^{-2})\right] \tag{28}$$

Similarly, interchanging the derivative and integration again we have:

$$f''_{\alpha_i}(\hat{\alpha}_i) = \frac{d_i^2}{\sigma^2} f_{\alpha_i}(\hat{\alpha}_i)\left\{(\hat{\alpha}_i - \alpha_i)^2\frac{d_i^2}{\sigma^2} - 1\right\}$$

And, the 2^{nd} derivative will similarly be:

$$m''(\alpha_i) = \int_{-\infty}^{\infty} f''_{\alpha_i}(\hat{\alpha}_i)\pi(\alpha_i)\ \mathrm{d}\alpha_i = \mathbb{E}\left[\frac{d_i^2}{\sigma^2}\phi(\hat{\alpha}_i; Y, \sigma^2 d_i^{-2})\left\{(\hat{\alpha}_i - Y)^2\frac{d_i^2}{\sigma^2} - 1\right\}\right] \tag{29}$$

Therefore, the algorithm for Monte Carlo integration is following:

(a) Fix $a, b > 0$ and generate a sample of size N (large) from $\mathrm{GDP}(b\sigma\tau/a, a)$ distribution. Let the sample be $Y_1, \cdots, Y_N$.

(b) Compute the following expressions for each $1 \le i \le N$

$$\widehat{m}(\hat{\alpha}_i) = \frac{1}{N}\sum_{j=1}^{N} \phi(\hat{\alpha}_i; Y_j, \sigma^2 d_i^{-2})$$

$$\widehat{m}'(\hat{\alpha}_i) = -\frac{1}{N}\sum_{j=1}^{N} (\hat{\alpha}_i - Y_j)\frac{d_i^2}{\sigma^2}\phi(\hat{\alpha}_i; Y_j, \sigma^2 d_i^{-2})$$

$$\widehat{m}''(\hat{\alpha}_i) = \frac{1}{N}\sum_{j=1}^{N} \frac{d_i^2}{\sigma^2}\phi(\hat{\alpha}_i; Y_j, \sigma^2 d_i^{-2})\left\{(\hat{\alpha}_i - Y_j)^2\frac{d_i^2}{\sigma^2} - 1\right\}$$

(c) For each i, compute

$$\widehat{T}_i = 2\sigma^2 - \sigma^4 d_i^{-2}\left[\frac{\widehat{m}'(\hat{\alpha}_i)}{\widehat{m}(\hat{\alpha}_i)}\right]^2 + 2\sigma^4 d_i^{-2}\frac{\widehat{m}''(\hat{\alpha}_i)}{\widehat{m}(\hat{\alpha}_i)}$$

By virtue of *law of large number*, the $\widehat{T}_i$'s should be very close to T_i's if N is sufficiently large. For practical purposes, N can taken to be of order 10^5 (used for our simulation studies) or 10^6.

Table 1: The values of optimal T for ridge, g-prior, lasso, horseshoe and three-parameter Beta (TPB) regression for different sample size and generating models for the design matrix X. For each of the methods, the values of T is computed by optimizing over the tuning parameter (e.g. the shrinkage parameter). [The models are denoted by: F = factor model, U = unit singular values model, B = bounded singular values model. Descriptions are provided in Section 4.]

Model	n	Condition number	Ridge	g-prior	Lasso	Horseshoe	TPB
F	100	15727.79	395.4172	405.9583	331.7101	299.2207	302.1561
	200	38.65082	782.8027	814.8452	631.4341	560.3065	566.8637
	500	17.09405	1741.366	2034.969	1427.219	1252.948	1274.938
U	100	1.0	380.2061	380.2061	237.0804	218.6526	218.2934
	200	1.0	763.529	763.529	490.1231	466.3304	468.8405
	500	1.0	1764.501	1764.501	1020.35	858.9972	880.5438
B	100	1.453409	375.6853	375.7516	254.2067	227.1044	230.0465
	200	1.471466	748.6505	755.0349	496.5098	432.9631	432.0447
	500	1.471801	1860.617	1870.072	1305.833	1155.029	1159.544

4 Simulation studies

As our interests concern the inspection of the results in Bhadra et al. [2019] for general d_i's, we generate a variety of design matrices with different structures along with sparse and robust α's. We will not only investigate the optimal Stein's unbiased risk estimators for the methods we have dealt with in this literature, we would also attempt for a performance analysis of the two main shrinkage estimation methods viz. horseshoe and ridge regression for the range of global shrinkage parameters obtained in Theorem 6. A description of the design matrices concerning this simulation setup is provided.

4.1 Simulation setup

We consider a high dimensional setup with $p = 500$ and $n = 100, 200, 500$.

(a) **Factor model (F)**: This setup is provided in Section 7 of Bhadra et al. [2019]. Let B be a $p \times k$ factor loading matrix whose all entries are 1. Let F_i be $k \times 1$ matrix of factor values, with all entries drawn independently from $\mathcal{N}(0, 1)$. The i^{th} row of the $n \times p$ design matrix X is generated by a factor model, with number of factors $k = 8$, as follows:

$$X_i = BF_i + \xi_i, \quad \xi_i \sim \mathcal{N}(\mathbf{0}_p, (0.1)I_p)$$

In this way we get a design matrix X whose columns are highly correlated and the matrix itself ill conditioned, i.e. the ration of maximum to minimum singular values of X is very large.

(b) **Unit singular values model (U)**: In this setup, we assume that d_i's are all 1 and hence $X = UDW^\top = UW^\top$, as $D = I$.

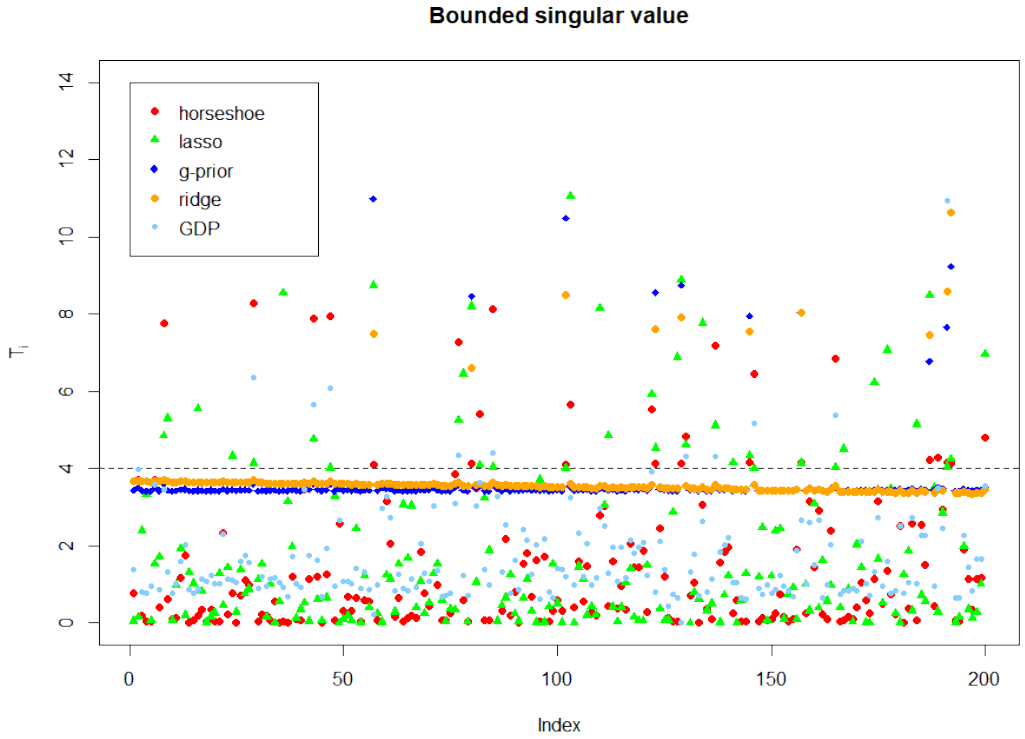

Figure 3: Component-wise T_i's for bounded d_i's

(c) **Bounded singular values model (B)**: As provided in Theorem 6, we generate two orthogonal matrices U and W, and a diagonal matrix D such that $D_{i,i} = d_i \sim \text{Uniform}(\eta^{-1}\tau^{-2}, \tau^{-2})$ as the range of d_i's given in Theorem 5. Then we construct X as $X = UDW^\top$. By our construction, it is clear that X is well-conditioned as the condition number is bounded by η.

For all of the aforementioned designs we consider the same structure of α for a fixed n. We assume that α has a sparse and robust structure, that is most of the α_i's are zero ensuring a sparse structure of α, and a few of the α_i's are large in a way that $||\alpha^2||/n$ is close to $K\sigma^2$ where K is close to 20. Hence, we work with the setup of Theorem 6. There are other possibilities for α, e.g. null α (when $\sum \alpha_i^2 = 0$), dense α (when all α_i's are non-zero) which have been explored by Bhadra et al. [2019] but we will focus on the sparse and robust setting in order to support our theoretical findings.

We generate i.i.d random errors ϵ_i's from $\mathcal{N}(0, \sigma^2)$ and calculate y_i's according to Equation (4). We compare the estimated predictive risk given by Stein's unbiased risk estimates associated with regression g-prior, ridge regression, lasso regression, horseshoe regression and three-parameter Beta regression. For each of these methods the formula of T is available in this literature. The ridge regression, lasso and regression with g-prior are purely global methods whereas the other two methods have a global-local setting.

For ridge and g-prior regression, the formula of component-wise T directly follows from Equation

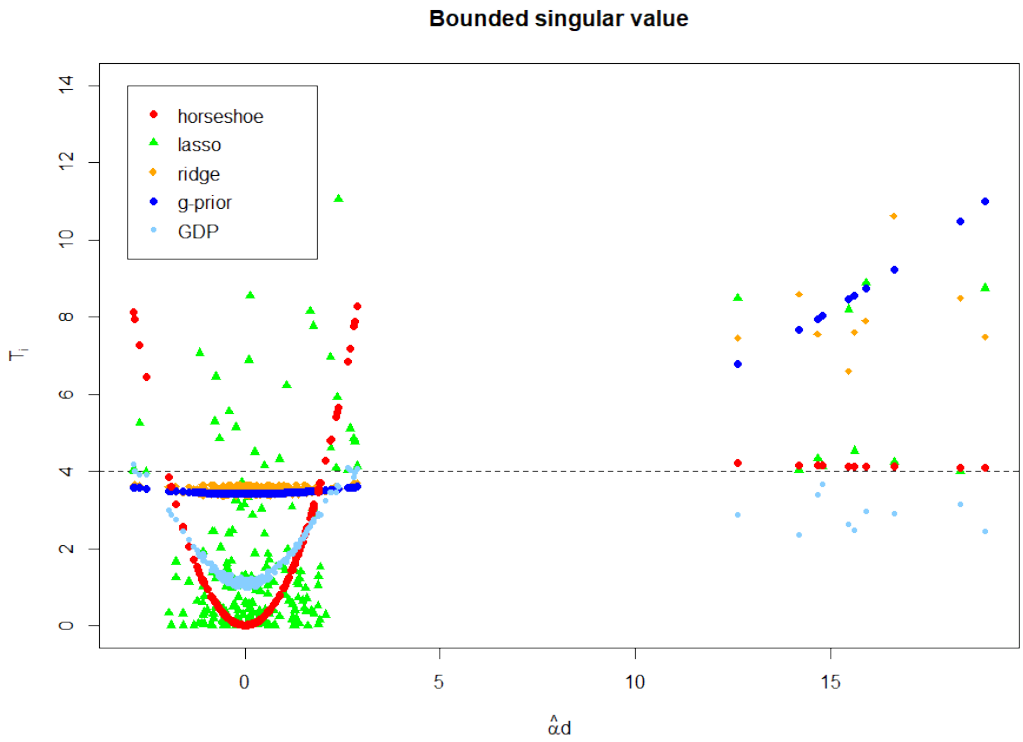

Figure 4: Component-wise T_i's vs $\hat{\alpha}d$ for bounded d_i's

(10) by putting $\lambda_i = d_i^{-2}$ and $d_i = 1$ respectively. In order to find the optimal estimates of risk we find $T^* = \min_{\tau^2} \sum T_i$ for each of these two methods. As the minimizer does not directly comes from the first order condition, we use numerical methods to solve for the optimal τ^2.

The formula of T for lasso can be computed using the dual of lasso, the details are provided in Tibshirani and Wasserman [2015] and Tibshirani and Taylor [2012]. The formula for optimal T in this case is

$$T^*_{\text{lasso}} = \underset{\lambda>0}{\arg\min} \left\{ ||y - Z\hat{\alpha}_\lambda||^2 + 2\sigma^2 \cdot (\#\{i : \hat{\alpha}_{\lambda,i} \neq 0\}) \right\} \tag{30}$$

where $\hat{\alpha}_\lambda$ is the estimate of lasso coefficients for penalty parameter being λ, that is

$$\hat{\alpha}_\lambda = \underset{\alpha \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2n} ||y - Z\alpha||^2 + \lambda ||\alpha||_1 \right\}$$

For horseshoe and three-parameter Beta regression, we calculate the optimal T by minimizing over τ^2 like we do in ridge as well as g-prior regression. However, as three-parameter Beta regression involves two other hyper-parameters viz. a and b, one can attempt to find the optimal risk estimate by jointly minimizing over (a, b, τ). For the ease of computation, we do not deal with it and simply assume a special case where $a = 0.7,\ b = 0.3$.

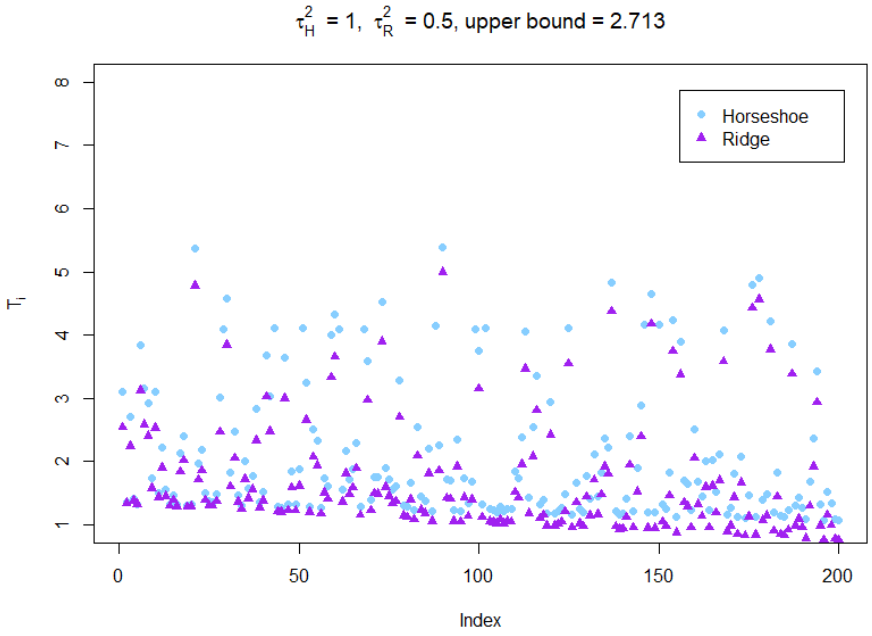

Figure 5: T_i's for Horseshoe and Ridge for $\tau_H^2 = 1,\ \tau_R^2 = 0.5$

4.2 Results

4.3 Comparison of different regression methods

The results are recorded in Table 1. The value of T for optimal horseshoe regression is the least for almost every sample size and design combination. It can be noted that for any design, as the sample size n increases the difference between T and $2n\sigma^2$ also gets larger. This phenomenon is natural since the number of sparse components in α increases more rapidly than the signal components and the contribution of those components in T is less than $2\sigma^2$, assuming that the bound of the inequality obtained in Theorem 5 is not tight.

The values of optimal T for for three-parameter Beta regression is extremely close to that of optimal horseshoe regression. Even if horseshoe is a special case of three-parameter Beta prior (with $a = 0.5,\ b = 0.5$), the change in optimal T is not very significant and hence for most practical purposes use of horseshoe prior should suffice for risk minimization as well as model selection.

The optimal T for lasso yields comparatively higher values than horseshoe and three-parameter Beta regression does. In addition, the difference between the risk estimates do not vary much with the sample size, giving us a hint that in terms of predictive risk horseshoe performs better than lasso. Similar computations may be carried out for other variants of lasso (adaptive lasso) as done in Bhadra et al. [2019] for all d_i's equal to 1.

The optimal ridge and optimal g-prior regression yields much larger values than optimal horseshoe and optimal three-parameter Beta regression. In addition, the variance of T_i's is also very less as the points are concentrated near a straight line. For the case $d_i = 1$ for all i, such phenomena

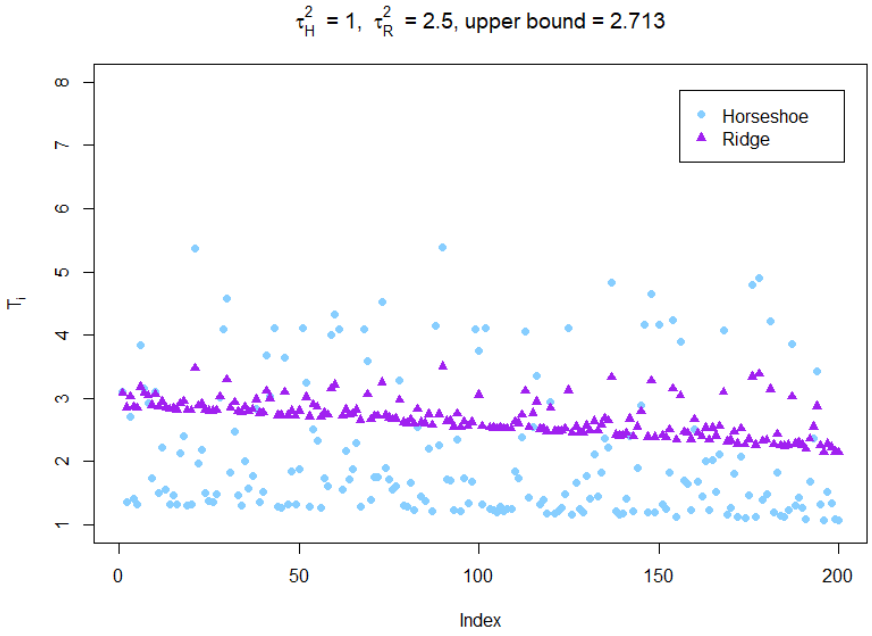

Figure 6: T_i's for Horseshoe and Ridge for $\tau_H^2 = 1,\ \tau_R^2 = 2.5$

directly follows from Corollary 5.1 of Bhadra et al. [2019]. However, for general d_i's we do not have any such theoretical evidence yet and theoretical investigations in this regard may be considered as a domain yet to be explored.

A similar pattern across different designs can be well-observed for any regression method. When T_i's are plotted according their component-wise indices, there is a high concentration of horseshoe-points in the bottom part and a lower concentration for higher values of T_i's (Figure 3). Same trend of concentration can be also observed in lasso, except for a relatively lower concentration in the bottom part and higher concentration in the upper part than horseshoe points, causing a slightly higher value of $T = \sum_i T_i$. The points from generalized double-Pareto regression concentrate on a relatively higher region than the cluster jointly created by the points corresponding to horseshoe and lasso. Put together, horseshoe is the best among these three methods. GDP regression has component-wise larger values of T_i's than lasso, but is more consistent in terms of the concentration of risk values.

In Figure 4, the red points (corresponding to horseshoe) converge to $2\sigma^2 = 4$ as $\hat{\alpha}_i d_i$ increases, validating the discussion of the remark after Theorem 2 as well as Theorem 3. For GDP regression the values do not fully converge because of the errors in Monte-Carlo integration. For ridge, lasso and g-prior regression, T_i's diverge away from $2\sigma^2$ as $\hat{\alpha}_i d_i$ grows. Hence for large signal components, horseshoe and TPB (as they are marginally close in values with those of horseshoe points) is the most effective in a sense of controlling the component-wise Stein's risk, followed by GDP regression.

The plots of component-wise T_i's as well as T_i vs $\hat{\alpha}_i d_i$'s are available in Appendix A.1 (Figures 7, 8, 9 & 10). When $d_i = 1$ for all i, the expression of T_i is same for ridge and g-prior regression

and hence the corresponding plots coincide. In the case when X is generated from a factor model, as $\hat{\alpha}_i$ increases, the rate of convergence of T_i to $2\sigma^2$ is slower than when the singular values of X are either all 1 or bounded.

4.4 Comparison of horseshoe and ridge regression for bounded singular values

To validate Theorem 6 of our main results, we fix $\tau_H^2 = 1$, and generate the singular values of X from $d_i \sim \text{Uniform}(\eta^{-1}, 1)$. By theorem 6, the upper bound on τ_R^2 turns out to be approximately 2.713 such that for $K \approx 20$, horseshoe can ensure less risk than ridge regression. The following table records the values of T for $\tau_H^2 = 1$ and τ_R^2 taking different values in $(0, 2.713)$.

Shrinkage	$\tau_R^2 = 0.5$	$\tau_R^2 = 1.0$	$\tau_R^2 = 1.5$	$\tau_R^2 = 2.0$	$\tau_R^2 = 2.5$	$\tau_R^{*2} = 8.64$
Ridge	1852.7092	1347.2999	1101.4982	967.1030	887.7046	741.2161

On the basis of the T values and Figures 5 and 6 following are a few observations:

(a) The value of T, and consequently Stein's risk, is strictly less for horseshoe than for ridge regression. Upon increasing the value of the shrinkage parameter τ_R^2s, the value of T rapidly decreases. When τ_R^2 is close to the upper bound (say $\tau_R^2 = 2.5$), the value of T is 887.7046 which is in a vicinity of the optimal value $T^* = 741.2161$ when $\tau_R^{*2} = 8.64$. Thus by increasing τ_R^2 by a relatively smaller amount we can obtain a value of T in close proximity of the optimal value. However, we need some further refinements in order to get a tighter bound towards the optimal situation.

(b) If we observe the component-wise T_i's for ridge, they are more dispersed away from the horizontal line (Figure 5) at $2\sigma^2$ if τ_R^2 is small. As the value τ_R^2 grows, the points corresponding to ridge start to concentrate on $2\sigma^2$ and eventually the points are aligned near a line slightly below the horizontal line at $2\sigma^2$ (Figure 6). As indicated by equation (10), if τ_R^2 is very large, the first summand in the expression of T_i turns out near-negligible and the second summand is close to $2\sigma^2$. Figure 5 and 6 captures two of the extreme cases; the plots for some other values of τ_R^2 is provided in the Appendix A.1 (Figures 11, 12 & 13).

5 Conclusion

In this article, we have studied the performance of a number of global-local shrinkage priors and analyzed their predictive performance with reference to a statistical criteria called Stein's unbiased risk estimate (SURE). Our theoretical results are mainly concerned with the very useful horseshoe regression. We make a gentle attempt to establish some theoretical results in support of the better performance of horseshoe over ridge. We obtain certain restrictions on the orthogonalized regression coefficient α and the global shrinkage parameters such that horseshoe outperforms ridge if the restrictions are satisfied. A possible direction yet to be explored may be to theoretically prove that

under certain conditions, optimal horseshoe regression will beat the optimal ridge estimate with respect to average risk. In addition, we give an introductory theoretical background for Stein's risk calculation of three-parameter Beta and generalized double Pareto regression. Further research in this direction can be an interesting direction to explore. We have also performed a detailed empirical study of the relative performance of a number of purely global shrinkage methods with horseshoe, three-parameter Beta and generalized double Pareto estimates.

In this article, we have theoretically as well empirically shown that horseshoe prior has a very impressive predictive performance than many popular global-shrinkage or selection based regression methods. The prediction is performed on future observations using the Bayes estimate of α_i which has a shrinkage coefficient $1 - \mathbb{E}(\kappa_i|y)$ (Equation (15)) multiplied with the least square estimate. The results in this paper and Bhadra et al. [2019] that the good prediction performance implies that the Bayes estimator works well in the sparse situation when most of the α_i's are zero and the signal components are large. As a result, the predicted value of the future observation is close to the true future observation. This means that the Bayes estimate is good in sparse situation when the true α_i is 0, and as a result the predicted $\hat{y}_i$ is close to y_i. This in turn indicates that when α_i is exactly zero, the shrinkage factor is perhaps very small and when α_i is very large the shrinkage factor is close to 1. Note in this context that the model selection rules suggested by Datta et al. [2013] and Ghosh et al. [2016] declare a parameter value to be 0 if the corresponding shrinkage factor is less than half and a signal (i.e. nonzero) if that factor is bigger than half. Also it had been shown that this model selection rule is optimal in a Decision Theoretic sense although the loss function used was different in their literature. There seems to be a connection of the good performance of the horseshoe in this apparently different context of model selection through the shrinkage factor. From a point of view concerning prediction risk results, one may think that it will perform well when the above model selection rule also shows us impressive performance. On the other hand, one may expect that the prediction risk for future observations might be small if the model selection is well-executed. Therefore, a plausible future direction of this work can be a predictive analysis based on Stein's risk, as done in Bhadra et al. [2019] and this literature, after using the model selection rule that one estimates α_i by zero if it is declared a noise and by the Bayes estimate if it is declared a signal. Another possible direction is explore risk associated with the estimation scheme with squared error loss function and whether such estimates indeed have a minimax rate of convergence.

6 Acknowledgement

I express my sincere gratitude to Professor Arijit Chakrabarti for advising my dissertation research. This opportunity marks inclusion of an important research work in my academic experience and it would definitely turn out helpful while I pursue my doctoral research in future. Without Professor Chakrabarti's constant assistance and valuable mentorship through the entire span of this work, the desired results would not have been accomplished. I also express my heartfelt thanks to Professors

Gopal K. Basak and Tapas Samanta for their insightful comments.

7 References

A. Armagan, D. B. Dunson, and M. Clyde. Generalized beta mixtures of gaussians. *Advances in neural information processing systems*, 24:523, 2011.

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

A. Bhadra, J. Datta, N. G. Polson, B. Willard, et al. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017.

A. Bhadra, J. Datta, Y. Li, N. G. Polson, and B. T. Willard. Prediction risk for the horseshoe regression. *J. Mach. Learn. Res.*, 20:78–1, 2019.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR, 2009.

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

G. Casella and J. T. Hwang. Limit expressions for the risk of james-stein estimators. *Canadian Journal of Statistics*, 10(4):305–309, 1982.

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

B. Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

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

M. B. Gordy et al. *A generalization of generalized beta distributions*, volume 18. Division of Research and Statistics, Division of Monetary Affairs, Federal . . . , 1998.

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

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

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

L. Pericchi and A. Smith. Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):793–804, 1992.

N. G. Polson and J. G. Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105, 2010.

N. G. Polson and J. G. Scott. Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311, 2012.

K. D. Schmidt. *On the covariance of monotone functions of a random variable*. Professoren des Inst. für Math. Stochastik, 2003.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

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

R. Tibshirani and L. Wasserman. Stein's unbiased risk estimate. *Course notes from "Statistical Machine Learning*, pages 1–12, 2015.

R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, pages 1198–1232, 2012.

A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.

A Appendix

A.1 Figures

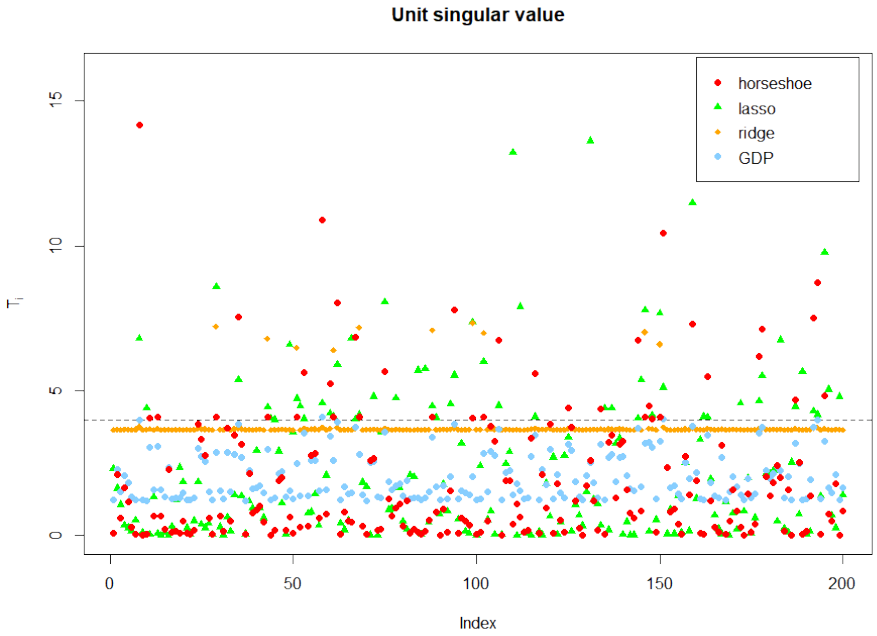

Figure 7: Component-wise T_i's for d_i's equal to 1

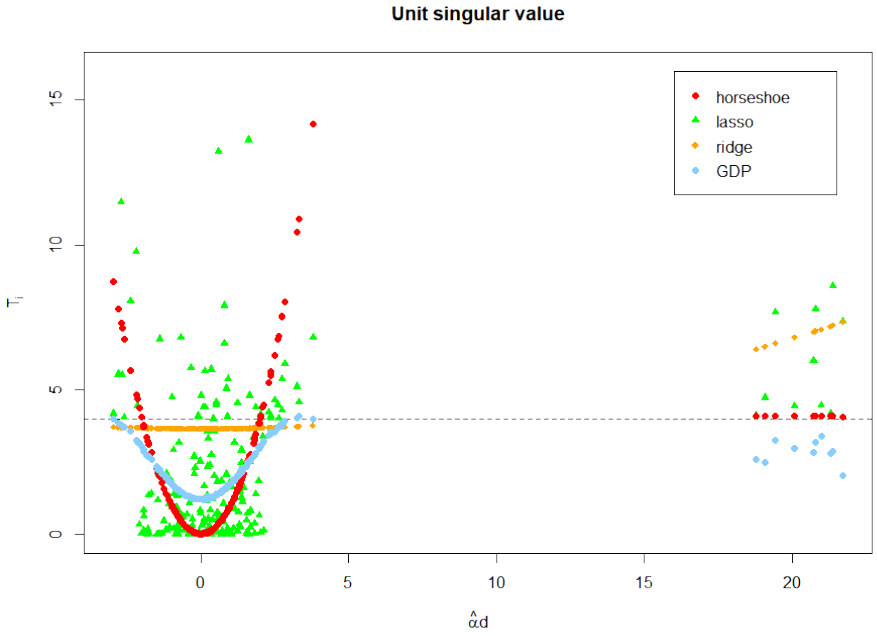

Figure 8: Component-wise T_i's vs $\hat{\alpha}d$ for d_i's equal to 1

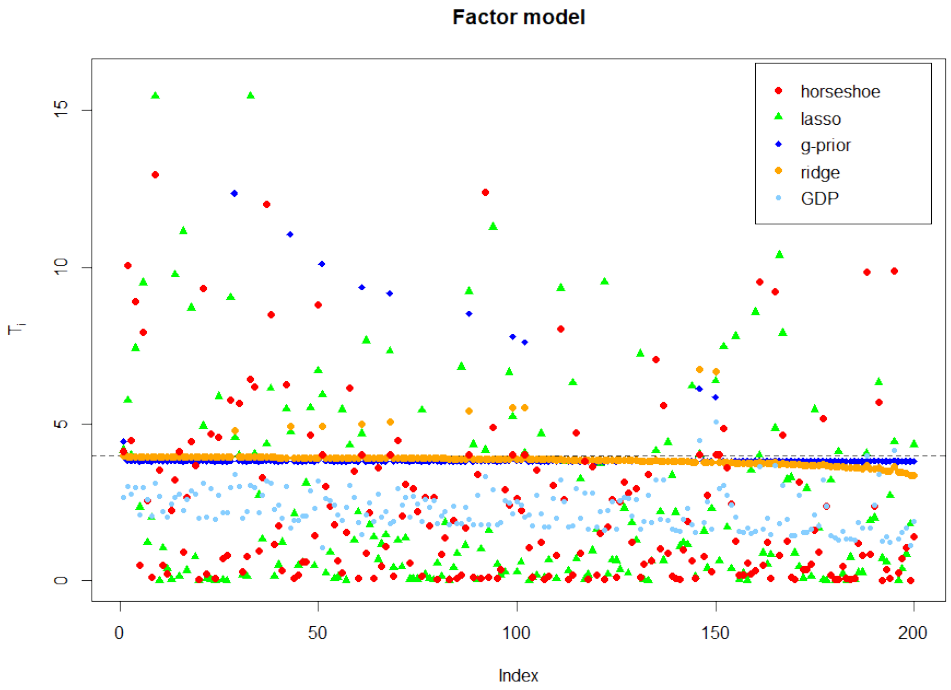

Figure 9: Component-wise T_i's when X is generated from an eight-factor model

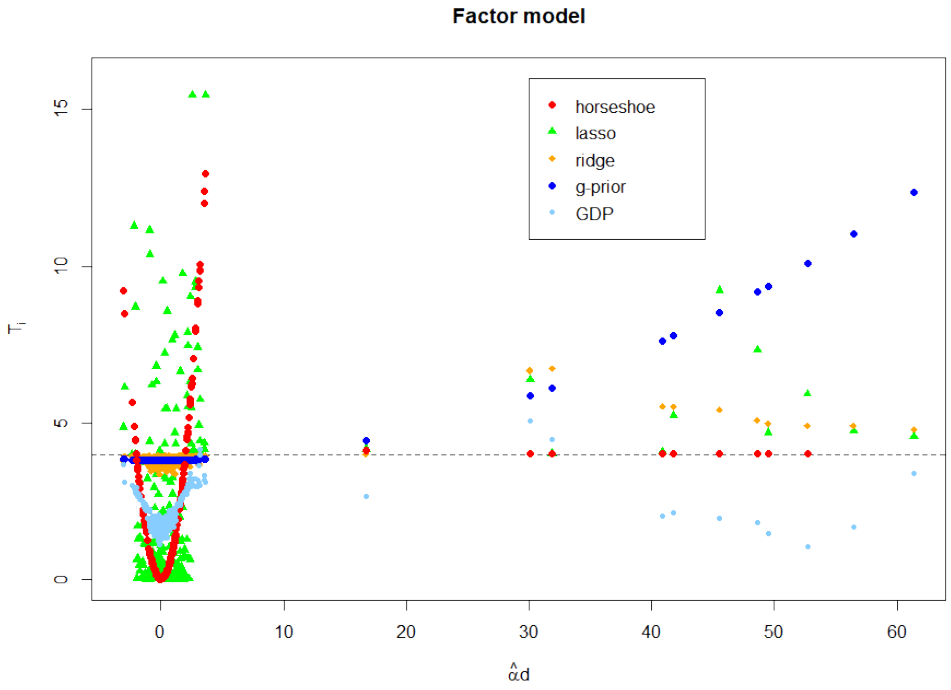

Figure 10: Component-wise T_i's vs $\hat{\alpha}d$ when X is generated from an eight-factor model

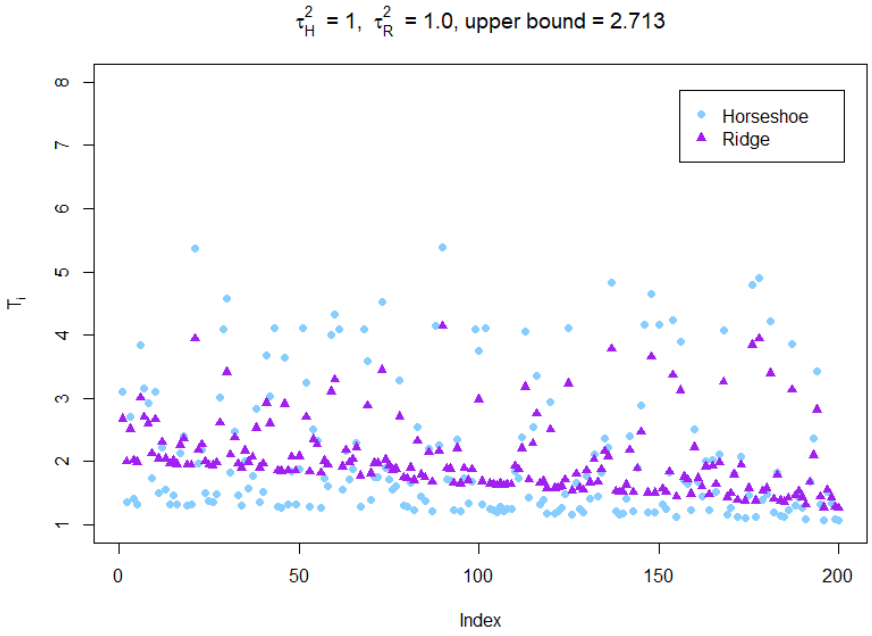

Figure 11: T_i's for Horseshoe and Ridge for $\tau_H^2 = 1,\ \tau_R^2 = 1.0$

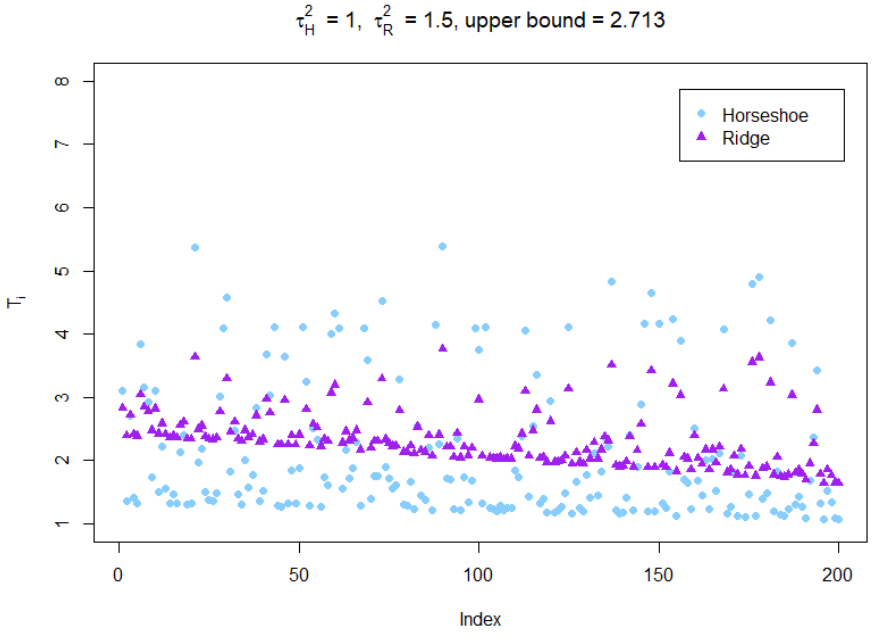

Figure 12: T_i's for Horseshoe and Ridge for $\tau_H^2 = 1,\ \tau_R^2 = 1.5$

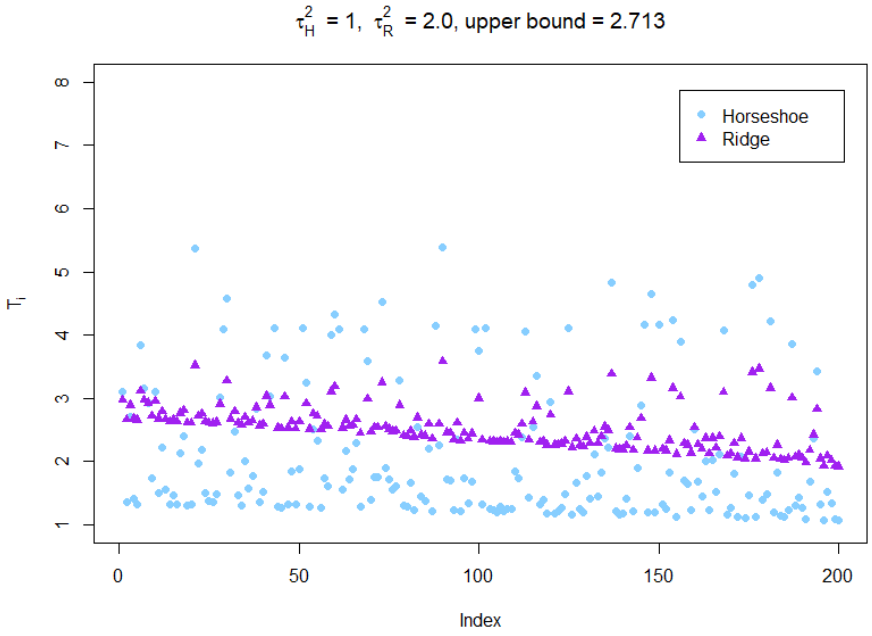

Figure 13: T_i's for Horseshoe and Ridge $\tau_H^2 = 1,\ \tau_R^2 = 2.0$

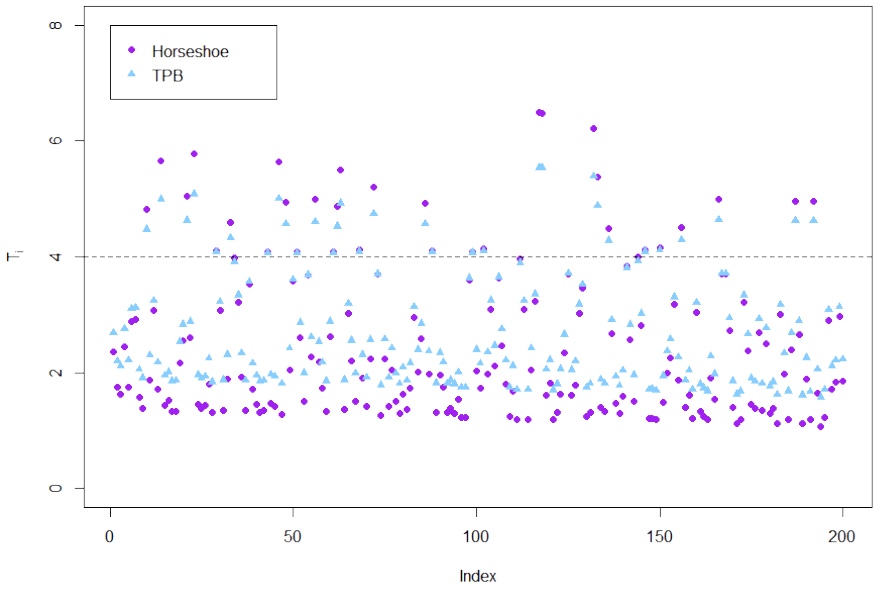

Figure 14: T_i's for optimal horseshoe and optimal TPB regression

A.2 Proof of Theorem 5

In this direction, we follow the architecture and the basic ideas of the proof of Theorem 4.1 of Bhadra et al. [2019].

Proof. Define $Z_i = 1/(1+\tau^2\lambda_i^2 d_i^2)$. Then the marginal of $\hat{\alpha}$ is

$$\begin{aligned} m(\hat{\alpha}) &= (2\pi\sigma^2)^{-n/2}\prod_{i=1}^{n}\int_0^1 \exp(-z_i\hat{\alpha}_i^2 d_i^2/2\sigma^2) d_i z_i^{1/2}\left(\frac{z_i\tau^2 d_i^2}{1-z_i+z_i\tau^2 d_i^2}\right)\frac{1}{\tau d_i}(1-z_i)^{-1/2} z_i^{-3/2}\,\mathrm{d}z_i \\ &= (2\pi\sigma^2)^{-n/2}\prod_{i=1}^{n}\int_0^1 \exp(-z_i\hat{\alpha}_i^2 d_i^2/2\sigma^2)(1-z_i)^{-1/2}\left(\frac{1}{\tau^2 d_i^2}+\left(1-\frac{1}{\tau^2 d_i^2}\right)z_i\right)^{-1}\,\mathrm{d}z_i \end{aligned}$$

The above integral is proportional to the normalizing constant of a CCH density. From this fact we have

$$(Z_i|\hat{\alpha}_i,\sigma,\tau) \sim \mathrm{CCH}\left(1,\frac{1}{2},1,\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2},1,\frac{1}{\tau^2 d_i^2}\right)$$

where $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$ and $\theta_i = (\tau^2 d_i^2)^{-1}$ with $\theta_i \geq 1$ (as $d_i^2 \leq \tau^{-2}$). From part (b) of theorem 1 we can write the risk estimate as $T = \sum_{i=1}^n T_i$ where

$$\begin{aligned} T_i &= 2\sigma^2[1-\mathbb{E}(Z_i)+s_i\mathbb{E}(Z_i^2)+s_i\mathbf{Var}(Z_i)] \\ &\leq 2\sigma^2[1-\mathbb{E}(Z_i)+s_i+s_i\mathbf{Var}(Z_i)] = R_i' \text{ (let)} \end{aligned}$$

Now consider the following cases:

(I) Consider $0 < s_i < 1$. We show that R_i' is convex in s_i for $0 < s_i < 1$. It suffices to show that $-\mathbb{E}(Z_i)$ and $s_i\mathbf{Var}(Z_i)$ are convex. In order to prove these we use the lemma in the appendix section of this literature.

$$\frac{\partial^2}{\partial s_i^2}\{-\mathbb{E}(Z_i)\} = -\mathbb{E}\{(Z_i-\mu)^3\} \geq 0$$

where the first equality follows from Lemma 2 and the inequality from Lemma 4 of Bhadra et al. [2019]. Further,

$$\begin{aligned} \frac{\partial^2}{\partial s_i^2}\{s_i\mathbf{Var}(Z_i)\} &= \frac{\partial}{\partial s_i}\left\{\mathbf{Var}(Z_i)+s_i\frac{\partial}{\partial s_i}\mathbf{Var}(Z_i)\right\} \\ &= 2\frac{\partial}{\partial s_i}\mathbf{Var}(Z_i)+s_i\frac{\partial^2}{\partial s_i^2}\mathbf{Var}(Z_i) \\ &= -2\mathbb{E}(Z_i-\mu)^3 - s_i\frac{\partial}{\partial s_i}\mathbb{E}(Z_i-\mu_i)^3 \quad \text{(lemma A.2)} \\ &= -2\mathbb{E}(Z_i-\mu)^3 + s_i\mathbb{E}(Z_i-\mu_i)^4 \geq 0 \quad \text{(lemma A.5)} \end{aligned}$$

Therefore R_i' is convex in s_i. Also by the Equation A.8 of the proof of Theorem 4.4 of Bhadra et al. [2019], at $s_i = 0$, we have $\sup_{\theta_i \in \mathbb{R}, s_i=1} T_i = (2/3)\sigma^2$. Also,

$$R_i'\big|_{s_i=1} = [1 - \mathbb{E}(Z_i) + 1 + \mathbf{Var}(Z_i)] \leq 2\sigma^2(1 - 0.57 + 1 + 0.43 - (0.57)^2) = 3.07\sigma^2$$

Therefore by the fact that $T_i \leq R_i'$ and convexity of R_i', the following holds:

$$T_i \leq 0.67\sigma^2 + s_i(3.07 - 0.67)\sigma^2 = (0.67 + 2.4s_i)\sigma^2$$

(II) Consider $1 \leq s_i < 3$. Now, $\dfrac{\partial}{\partial s_i}\mathbb{E}(Z_i) = \mathbb{E}(Z_i)\mathbb{E}(Z_i) - \mathbb{E}(Z_i^2) = -\mathbf{Var}(Z_i) < 0$. Also, as shown in Schmidt [2003], for $g(z) = z^2$ being an increasing function of z, we have

$$\begin{aligned} &\mathbf{Cov}(Z_i, Z_i^2) = \mathbf{Cov}(Z_i, g(Z_i)) > 0 \\ \implies &\frac{\partial}{\partial s_i}\mathbb{E}(Z_i^2) = \mathbb{E}(Z_i)\mathbb{E}(Z_i^2) - \mathbb{E}(Z_i^3) = -\mathbf{Cov}(Z_i, g(Z_i)) < 0 \end{aligned}$$

Therefore we have,

$$T_i \leq 2\sigma^2 \left[1 - \mathbb{E}(Z_i)|_{s_i=3} + 2s_i\ \mathbb{E}(Z_i^2)|_{s_i=1} - s_i \left\{\mathbb{E}(Z_i)|_{s_i=3}\right\}^2\right]$$

Also

$$\mathbb{E}(Z_i)|_{s_i,\theta_i} = \frac{\int_0^1 z_i(1-z_i)^{-1/2}(\theta_i + (1-\theta_i)z_i)^{-1}\exp(-s_i z_i)\ \mathrm{d}z_i}{\int_0^1 (1-z_i)^{-1/2}(\theta_i + (1-\theta_i)z_i)^{-1}\exp(-s_i z_i)\ \mathrm{d}z_i}$$

For $d_i^2 \leq \tau^{-2}\ \forall i$, that is, for $\theta_i \geq 1\ \forall i$, we have $\theta_i^{-1} \leq (\theta_i + (1-\theta_i)z_i)^{-1} \leq 1$. Therefore,

$$\begin{aligned} \mathbb{E}(Z_i)|_{s_i=3} &= \frac{\int_0^1 z_i(1-z_i)^{-1/2}(\theta_i + (1-\theta_i)z_i)^{-1}\exp(-3s_i z_i)\ dz_i}{\int_0^1 (1-z_i)^{-1/2}(\theta_i + (1-\theta_i)z_i)^{-1}\exp(-3s_i z_i)\ \mathrm{d}z_i} \\ &\geq \frac{\theta_i^{-1}\int_0^1 z_i(1-z_i)^{-1/2}\exp(-3s_i z_i)\ \mathrm{d}z_i}{\int_0^1 (1-z_i)^{-1/2}\exp(-3s_i z_i)\ \mathrm{d}z_i} \\ &= \theta_i^{-1} I_1 \quad \text{(let. Also } I_1 \approx 0.3834362\text{)} \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{E}(Z_i^2)|_{s_i=1} &= \frac{\int_0^1 z_i^2(1-z_i)^{-1/2}(\theta_i + (1-\theta_i)z_i)^{-1}e^{-z_i}\ \mathrm{d}z_i}{\int_0^1 (1-z_i)^{-1/2}(\theta_i + (1-\theta_i)z_i)^{-1}e^{-z_i}\ \mathrm{d}z_i} \\ &\leq \frac{\int_0^1 z_i^2(1-z_i)^{-1/2}e^{-z_i}\ \mathrm{d}z_i}{\theta_i^{-1}\int_0^1 (1-z_i)^{-1/2}e^{-z_i}\ \mathrm{d}z_i} = \theta_i I_2 \quad \text{(let. Also } I_2 \approx 0.4297594\text{)} \end{aligned}$$

Therefore, for $1 \leq s_i < 3$ we have: $T_i \leq 2\sigma^2[1 - \theta_i^{-1} I_1 + 2\theta_i s_i I_2 - \theta_i^{-2} s_i I_1^2]$

(III) Consider $s_i \geq 3$. Therefore, by the upper bound given in Theorem 2,

$$T_i \leq 11.55\sigma^2 \quad \text{for } s_i \geq 3$$

When $\alpha_i = 0$, $\hat{\alpha}_i \sim \mathcal{N}(0, \sigma^2 d_i^{-2})$. Therefore, $\hat{\alpha}_i^2 d_i^2/\sigma^2 \sim \chi^2(1)$ i.e. Gamma(1/2, 1/2) and $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2 \sim$ Gamma$(1, 1/2)$ where we follow the convention that a Gamma(α, λ) random variable has shape parameter α and rate parameter λ. This means $p(s_i) = (\pi)^{-1/2} s_i^{-1/2} \exp(-s_i)$ for $s_i > 0$.

Also, denote F_1 and F_2 be the distribution functions of Gamma(α_i, λ_i) where $\alpha_1 = 1/2$, $\alpha_2 = 3/2$, and $\lambda_i = 1$ for $i = 1, 2$. Hence for true $\alpha_i = 0$ Combining the upper bounds on SURE in the three cases we had considered, the following holds:

$$\begin{aligned}
\mathbb{E}_{\alpha_i}(T_i) \leq & \int_0^1 \sigma^2(0.67 + 2.4 s_i)\pi^{-1/2} s_i^{-1/2} e^{-s_i}\, \mathrm{d}s_i \\
& + \int_1^3 2\sigma^2 \left[1 - \theta_i^{-1} I_1 + 2\theta_i s_i I_2 - \theta_i^{-2} s_i I_1^2\right] \pi^{-1/2} s_i^{-1/2} e^{-s_i}\, \mathrm{d}s_i \\
& + \int_3^\infty 11.55\sigma^2 \pi^{-1/2} s_i^{-1/2} e^{-s_i}\, \mathrm{d}s_i \\
= & \, 0.67\sigma^2 \left(F_1(1) - F_1(0)\right) + 2.4 \cdot (0.5)\sigma^2 \left(F_2(1) - F_2(0)\right) \\
& + 2\sigma^2 \left(F_1(3) - F_1(1)\right)\left(1 - \theta_i^{-1} I_1\right) \\
& + 2 \cdot (0.5)\sigma^2 \left(F_2(3) - F_2(1)\right)\left(2\theta_i I_2 - \theta_i^{-2} I_1^2\right) + 11.5\sigma^2 \left(1 - F_1(3)\right) \\
\approx & \, \sigma^2[1.53 - 0.11\ \theta_i^{-1} + 0.39\ \theta_i - 0.067\ \theta_i^{-2}]
\end{aligned}$$

In the last expression, we consider the the exact constant term as C_0 and the exact coefficient terms of θ_i, θ_i^{-1} and θ_i^{-2} respectively as C_1, C_2 and C_3, we have the proof for part (a) of Theorem 5.

To prove part (b), we let $f(\theta_i) = \sigma^2(C_0 + C_1\theta_i - C_2\theta_i^{-1} - C_3\theta_i^{-2})$ As, $f'(\theta_i) = C_1 + C_2\theta_i^{-2} + 2C_3\theta_i^{-3} > 0$ for $\theta_i \geq 1$, therefore f is monotonically increasing in θ_i for $\theta_i \geq 1$. Also, $f(\theta_i) < 2$ if and only if $\theta_i \in [1, \eta)$ with $\eta \approx 1.475$, which is same as $d_i^2 \in \left(\frac{1}{\eta\tau^2}, \frac{1}{\tau^2}\right]$. Hence part (b) is proved. □

A.3 Proof of Lemma 1

Proof. For any $1 \leq i \leq n$, let $|\alpha_i| \to \infty$. Hence there exists a sequence $\{\alpha_{i,k}\}$ such that $|\alpha_{i,k}| \to \infty$ as $k \to \infty$, and also there exists a natural number N such that $|\alpha_{i,k}| > 2$ for $k \geq N$.

For $k \geq N$,

$$\mathbb{E}_{\alpha_{i,k}}(T_i) = \mathbb{E}_{\alpha_{i,k}}(T_i \mathbb{1}\{s_i < 1\}) + \mathbb{E}_{\alpha_{i,k}}\left(T_i \mathbb{1}\{1 \leq s_i < |\alpha_{i,k}|/2\}\right) + \mathbb{E}_{\alpha_{i,k}}\left(T_i \mathbb{1}\{s_i \geq |\alpha_{i,k}|/2\}\right) \quad (31)$$

where $s_i = \hat{\alpha}_i^2 d_i^2/\sigma^2$ as it has been defined earlier, and the expectation is taken with α_i fixed.

From Theorem 4.4 of Bhadra et al. [2019], for $0 \leq s_i \leq 1$ and any fixed τ, T_i is an increasing function of s_i and for $s_i = 1$, $0 < T_i \leq 1.93\sigma^2$ when $0 < \tau^2 d_i^2 \leq 1$ i.e. $\theta_i \geq 1$. By this theorem we can write the following:

From Theorem 1, the following can be written:

$$\begin{aligned} T_i &= 2\sigma^2 - 2\sigma^2\mathbb{E}(Z_i) - 2\hat{\alpha}_i^2 d_i^2\{\mathbb{E}(Z_i)\}^2 + 2\hat{\alpha}_i^2 d_i^2 \mathbb{E}(Z_i^2) \\ &= 2\sigma^2[1 - \mathbb{E}(Z_i) + 2s_i\mathbb{E}(Z_i^2) - s_i\{\mathbb{E}(Z_i^2)\}] \end{aligned}$$

Thus,

$$2\sigma^2[1 - \mathbb{E}(Z_i) - s_i\{E(Z_i)\}^2] \leq T_i \leq 2\sigma^2[1 + 2s_i\mathbb{E}(Z_i^2)]$$

Also $0 \leq Z_i \leq 1$. For $0 \leq s_i \leq 1$, we have $-2\sigma^2 \leq T_i \leq 6\sigma^2$, i.e. $|T_i| \leq 6\sigma^2$

Therefore,

$$\left|\mathbb{E}_{\alpha_{i,k}}(T_i \mathbb{1}\{s_i < 1\})\right| \leq \mathbb{E}_{\alpha_{i,k}}(|T_i|\mathbb{1}\{s_i < 1\}) \leq \mathbb{E}(6\sigma^2\ \mathbb{1}\{s_i < 1\}) = 6\sigma^2\ \mathbb{P}_{\alpha_{i,k}}(s_i < 1) \tag{32}$$

For $s_i \geq 1$ and $\theta_i \geq 1$, we have an upper and a lower bound (from Theorem 2) on T_i which is

$$\left[1 - \theta_i(\tilde{C}_1 + \tilde{C}_2) \cdot \frac{1+s_i}{s_i^2} - \theta_i^2(\tilde{C}_1 + \tilde{C}_2)^2 \cdot \frac{(1+s_i)^2}{s_i^3}\right] \leq \frac{T_i}{2\sigma^2} \leq \left[1 + 2\theta_i(1+s_i)\left(\frac{C_1}{s_i^2} + \frac{C_2}{s_i^{3/2}}\right)\right]$$

$C_1, C_2, \tilde{C}_1$ and $\tilde{C}_2$ are positive constants. The lower bound on $T_i/2\sigma^2$ is increasing in s_i and the upper bound in decreasing in s_i. Hence we can put $s_i = 1$ in both sides and get:

$$1 - 2\theta_i(\tilde{C}_1 + \tilde{C}_2) - 4\theta_i^2(\tilde{C}_1 + \tilde{C}_2)^2 \leq \frac{T_i}{2\sigma^2} \leq 1 + 4\theta_i(C_1 + C_2)$$

Denote $M = \max\left\{|1 - 2\theta_i(\tilde{C}_1 + \tilde{C}_2) - 4\theta_i^2(\tilde{C}_1 + \tilde{C}_2)^2|\ ,\ |1 + 4\theta_i(C_1 + C_2)|\right\}$. Therefore, for $1 \leq s_i < \alpha_{i,k}/2$, we have $|T_i| \leq 2M\sigma^2$ and

$$\begin{aligned} \left|\mathbb{E}_{\alpha_{i,k}}(T_i \mathbb{1}\{1 \leq s_i < |\alpha_{i,k}|/2\})\right| &\leq \mathbb{E}_{\alpha_{i,k}}(|T_i|\mathbb{1}\{1 \leq s_i < |\alpha_{i,k}|/2\}) \\ &\leq \mathbb{E}(2M\sigma^2\ \mathbb{1}\{1 \leq s_i < |\alpha_{i,k}|/2\}) \\ &= 2M\sigma^2\ \mathbb{P}_{\alpha_{i,k}}(1 \leq s_i < |\alpha_{i,k}|/2) \end{aligned} \tag{33}$$

When $s_i \geq |\alpha_{i,k}|/2$, we again use Theorem 2 to write

$$\begin{aligned} &\left[1 - \theta_i(\tilde{C}_1 + \tilde{C}_2) \cdot \frac{(1+|\alpha_{i,k}|)}{|\alpha_{i,k}|^2} - \alpha_{i,k}^2(\tilde{C}_1 + \tilde{C}_2)^2 \cdot \frac{(1+|\alpha_{i,k}|)^2}{|\alpha_{i,k}|^3}\right] \\ &\qquad \leq \quad \frac{T_i}{2\sigma^2} \quad \leq \quad \left[1 + 2\theta_i(1+|\alpha_{i,k}|)\left(\frac{C_1}{|\alpha_{i,k}|^2} + \frac{C_2}{|\alpha_{i,k}|^{3/2}}\right)\right] \end{aligned} \tag{34}$$

Now we compute the following useful probability

$$\begin{aligned}\mathbb{P}_{\alpha_{i,k}}(s_i < |\alpha_{i,k}|/2) &= \mathbb{P}_{\alpha_{i,k}}(\hat{\alpha}_i^2 d_i^2/\sigma^2 < |\alpha_{i,k}|/2) \\ &= \mathbb{P}_{\alpha_{i,k}}\left(-\frac{\alpha_{i,k}d_i}{\sigma} - \frac{\sigma}{d_i}\sqrt{\frac{|\alpha_{i,k}|}{2}} < \frac{(\hat{\alpha}_i - \alpha_{i,k})d_i}{\sigma} < -\frac{\alpha_{i,k}d_i}{\sigma} + \frac{\sigma}{d_i}\sqrt{\frac{|\alpha_{i,k}|}{2}}\right) \\ &= \Phi\left(-\frac{\alpha_{i,k}d_i}{\sigma} + \frac{\sigma}{d_i}\sqrt{\frac{|\alpha_{i,k}|}{2}}\right) - \Phi\left(-\frac{\alpha_{i,k}d_i}{\sigma} - \frac{\sigma}{d_i}\sqrt{\frac{|\alpha_{i,k}|}{2}}\right)\end{aligned}$$

Taking limit on $\alpha_{i,k}$,

$$\begin{aligned}&\lim_{|\alpha_{i,k}|\to\infty} \mathbb{P}_{\alpha_{i,k}}(s_i < |\alpha_{i,k}|/2) \\ &= \lim_{|\alpha_{i,k}|\to\infty} \left[\Phi\left(-\frac{\alpha_{i,k}d_i}{\sigma} + \frac{\sigma}{d_i}\sqrt{\frac{|\alpha_{i,k}|}{2}}\right) - \Phi\left(-\frac{\alpha_{i,k}d_i}{\sigma} - \frac{\sigma}{d_i}\sqrt{\frac{|\alpha_{i,k}|}{2}}\right)\right] \\ &= 0 \quad \text{[Both approaches either 0 or 1 simultaneously]}\end{aligned}$$

As a result we can write

$$\lim_{|\alpha_{i,k}|\to\infty} \mathbb{P}_{\alpha_{i,k}}(1 \le s_i < |\alpha_{i,k}|/2) = \lim_{|\alpha_{i,k}|\to\infty} \mathbb{P}_{\alpha_{i,k}}(s_i < 1) = 0$$

and

$$\lim_{|\alpha_{i,k}|\to\infty} \mathbb{P}_{\alpha_{i,k}}(s_i \ge |\alpha_{i,k}|/2) = 1$$

Taking limit in Equation (32) and (33)

$$\lim_{|\alpha_{i,k}|\to\infty} \mathbb{E}(T_i \mathbb{1}\{s_i < 1\}) = 0 \ , \quad \lim_{|\alpha_{i,k}|\to\infty} \mathbb{E}(T_i \mathbb{1}\{1 \le s_i < |\alpha_{i,k}|/2\}) = 0$$

In Equation (34), both the upper and lower bounds on $T_i/2\sigma^2$ approaches 1. Therefore,

$$\begin{aligned}&2\sigma^2\left[1 - \theta_i(\tilde{C}_1 + \tilde{C}_2)\cdot\frac{(1+|\alpha_{i,k}|)}{|\alpha_{i,k}|^2} - \alpha_{i,k}^2(\tilde{C}_1+\tilde{C}_2)^2\cdot\frac{(1+|\alpha_{i,k}|)^2}{|\alpha_{i,k}|^3}\right]\mathbb{P}_{\alpha_{i,k}}\left(s_i \ge \frac{|\alpha_{i,k}|}{2}\right) \\ &\le \mathbb{E}\left(T_i \mathbb{1}\left\{s_i \ge \frac{|\alpha_{i,k}|}{2}\right\}\right) \le 2\sigma^2\left[1 + 2\theta_i(1+|\alpha_{i,k}|)\left(\frac{C_1}{|\alpha_{i,k}|^2} + \frac{C_2}{|\alpha_{i,k}|^{3/2}}\right)\right]\mathbb{P}_{\alpha_{i,k}}\left(s_i \ge \frac{|\alpha_{i,k}|}{2}\right)\end{aligned}$$

By sandwich principle we have $\lim_{|\alpha_{i,k}|\to\infty} \mathbb{E}(T_i \mathbb{1}\{s_i \ge |\alpha_{i,k}|/2\}) = 2\sigma^2$
Combining the limits and from Equation (31)

$$\lim_{|\alpha_{i,k}|\to\infty} \mathbb{E}(T_i) = 2\sigma^2$$

□

A.4 Proof of Theorem 6

Proof. Without loss of generality, we assume that the signal components of α are $\alpha_1, \cdots, \alpha_m$ and the sparse components are $\alpha_i = 0$ for $m < i \leq n$. Here m is a fixed and finite number smaller than n.

For $i > m$, $\alpha_i = 0$ and Theorem 1 implies

$$T_i \leq \sigma^2[C_0 + C_1\theta_i - C_2\theta_i - C_3\theta_i^2] = f(\theta_i)\sigma^2 \text{ (let)}$$

We consider the case when for a positive constant $A \in (1.743, 2)$

$$T_i \leq f(\theta_i)\sigma^2 \leq A\sigma^2 \iff 1 \leq \theta_i \leq \eta(A)$$

The above holds true as f is increasing in θ_i.

For $i \leq m$, $|\alpha_i| \to \infty$ with $n \to \infty$ and we use Lemma 1 to have $\mathbb{E}(T_i) \to 2\sigma^2$.

That is, for $i \leq m$ and given $\epsilon > 0$, there exists a natural number N_i such that $\mathbb{E}(T_i)/2\sigma^2 < 1+\epsilon$ if $n \geq N_i$

Thus for $n \geq N' := \max\{N_1, \cdots, N_m\}$ the average of T_i's for horseshoe is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(T_i^H) &= \frac{1}{n}\sum_{i=1}^{m}\mathbb{E}(T_i^H) + \frac{1}{n}\sum_{i=m+1}^{n}\mathbb{E}(T_i^H) \\ &< \frac{1}{n}\left(m(2\sigma^2+\epsilon) + (n-m)A\sigma^2\right) \\ &= A\sigma^2 + (2+\epsilon-A)\sigma^2\frac{m}{n}\end{aligned}$$

We choose $N'' = \max\left\{N', \frac{(2+\epsilon-A)m}{\delta}\right\}$. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(T_i^H) < (A+\delta)\sigma^2 < 2\sigma^2 \quad \text{for } n \geq N''$$

We are left with showing that for the range of τ_H and τ_R, sufficiently large K and n

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(T_i^R) > 2\sigma^2$$

Also, $\hat{\alpha}_i^2 d_i^2/\sigma^2 \sim \chi^2$ with $d.f. = 1$ and $n.c.p = \alpha_i^2 d_i^2/\sigma^2$. Thus, $\mathbb{E}(\hat{\alpha}_i^2 d_i^2) = \alpha_i^2 d_i^2 + \sigma^2$.

taking expectation of the expression of T_i's given in Equation (10) with λ_i's equal to 1,

$$\sum_{i=1}^{n}\mathbb{E}(T_i^R) = \sum_{i=1}^{n}\left(\frac{\alpha_i^2 d_i^2 + \sigma^2}{(1+\tau^2 d_i^2)^2} + 2\sigma^2\frac{\tau^2 d_i^2}{1+\tau^2 d_i^2}\right)$$

From the bounds on τ_H^2 given in the conditions of the theorem,

$$\frac{1}{\eta(A)\min_i d_i^2} \leq \tau_H^2 \leq \frac{1}{\max_i d_i^2} \implies \frac{1}{\eta(A)\tau_H^2} \leq d_i^2 \leq \frac{1}{\tau_H^2} \quad \text{for all } i$$

Using the bounds on d_i's and the expression of $\sum_i \mathbb{E}(T_i^R)$ we can write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(T_i^R) &\geq \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\alpha_i^2\frac{1}{\eta(A)\tau_H^2}+\sigma^2}{\left(1+\frac{\tau_R^2}{\tau_H^2}\right)^2}+2\sigma^2\frac{\tau_R^2\frac{1}{\eta(A)\tau_H^2}}{1+\frac{\tau_R^2}{\tau_H^2}}\right\} \\ &= \frac{1}{\eta(A)\tau_H^2\left(1+\frac{\tau_R^2}{\tau_H^2}\right)^2}\cdot\frac{||\alpha||^2}{n}+\frac{\sigma^2}{\left(1+\frac{\tau_R^2}{\tau_H^2}\right)^2}+\frac{2\sigma^2}{\eta(A)}\cdot\frac{\tau_R^2/\tau_H^2}{1+\frac{\tau_R^2}{\tau_H^2}} \qquad (35)\end{aligned}$$

As $||\alpha||^2/n \to K\sigma^2$ as $n \to \infty$, hence there exists N_K such that

$$\frac{||\alpha||^2}{n} > \frac{K}{2}\sigma^2 \quad \text{if } n \geq N_K$$

Let $N = \max\{N'', N_K\}$. Then for $n \geq N$ we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(T_i^R) > \frac{1}{\eta(A)\tau_H^2\left(1+\frac{\tau_R^2}{\tau_H^2}\right)^2}\cdot\frac{K}{2}\sigma^2+\frac{\sigma^2}{\left(1+\frac{\tau_R^2}{\tau_H^2}\right)^2}+\frac{2\sigma^2}{\eta(A)}\cdot\frac{\tau_R^2/\tau_H^2}{1+\frac{\tau_R^2}{\tau_H^2}} \geq 2\sigma^2 \text{ (let)}$$

Let $x = \left(1+\frac{\tau_R^2}{\tau_H^2}\right)^{-1}$. Then we can write the condition as

$$x^2\left(1+\frac{K}{2\eta(A)\tau_H^2}\right)-\frac{2}{\eta(A)}x-2\left(1-\frac{1}{\eta(A)}\right) \geq 0$$

The discriminant for the above quadratic expression is

$$\left(\frac{2}{\eta(A)}\right)^2+8\left(1+\frac{K}{2\eta(A)\tau_H^2}\right)\left(1-\frac{1}{\eta(A)}\right) > 0 \text{ , because } A > 1.743 \text{ , and } \eta(A) > 1$$

The solution region for x is

$$\begin{aligned}x &\geq \frac{\frac{2}{\eta(A)}+\sqrt{\left(\frac{2}{\eta(A)}\right)^2+8\left(1+\frac{K}{2\eta(A)\tau_H^2}\right)\left(1-\frac{1}{\eta(A)}\right)}}{2\left(1+\frac{K}{2\eta(A)\tau_H^2}\right)} \\ &= \frac{\frac{1}{\eta(A)}+\sqrt{\left(\frac{1}{\eta(A)}\right)^2+2\left(1+\frac{K}{2\eta(A)\tau_H^2}\right)\left(1-\frac{1}{\eta(A)}\right)}}{1+\frac{K}{2\eta(A)\tau_H^2}}\end{aligned}$$

The other solution region is infeasible as x has to be positive. Define

$$B(A) = \sqrt{\left(\frac{1}{\eta(A)}\right)^2 + 2\left(1 + \frac{K}{2\eta(A)\tau_H^2}\right)\left(1 - \frac{1}{\eta(A)}\right)}$$

Therefore, from the expression of x in terms of τ_H and τ_R, we obtain the following relation

$$\frac{1}{1 + \frac{\tau_R^2}{\tau_H^2}} \geq \frac{\frac{1}{\eta(A)} + B(A)}{1 + \frac{K}{2\eta(A)\tau_H^2}} \implies \tau_R^2 \leq \tau_H^2 \left[\frac{1 + \frac{K}{2\eta(A)\tau_H^2}}{B(A) + \frac{1}{\eta(A)}} - 1\right]$$

Hence we have the range of τ_R in terms of τ_H stated in the theorem.

Now we choose $n \geq N$. Combining with the previous results we get

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(T_i^H)) < (A + \delta)\sigma^2 < 2\sigma^2 < \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(T_i^R)$$

□

A.5 Proof of Theorem 7

Proof. Let $\rho_i = \lambda_i^2$. The marginal of $\hat{\alpha}$ can be expressed as

$$\begin{aligned}
m(\hat{\alpha}) &= \prod_{i=1}^{n} \int_0^\infty \mathcal{N}(\hat{\alpha}_i; 0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2))\ \pi(\lambda_i^2)\ \mathrm{d}(\lambda_i^2) \\
&\propto \prod_{i=1}^{n} \int_0^\infty \exp\left\{-\frac{\hat{\alpha}_i^2 d_i^2/2}{\sigma^2(1 + \tau^2 d_i^2\lambda_i^2)}\right\}\ \frac{d_i}{(1 + \tau^2 d_i^2\lambda_i^2)^{1/2}}\ (\lambda_i^2)^{a-1}(1 + \lambda_i^2)^{-(a+b)}\ \mathrm{d}(\lambda_i^2) \\
&\propto \prod_{i=1}^{n} \int_0^\infty \exp\left\{-\frac{\hat{\alpha}_i^2 d_i^2/2}{\sigma^2(1 + \tau^2 d_i^2\rho_i)}\right\}\ \frac{d_i}{(1 + \tau^2 d_i^2\rho_i)^{1/2}}\ \rho_i^{a-1}(1 + \rho_i)^{-(a+b)}\ \mathrm{d}\rho_i \\
&\propto \prod_{i=1}^{n} \int_0^1 \exp\left\{-\frac{\hat{\alpha}_i^2 d_i^2 z_i}{2\sigma^2}\right\}\ d_i z_i^{1/2}\ \left\{\frac{1}{\tau^2 d_i^2}\left(\frac{1}{z_i} - 1\right)\right\}^{a-1}\ \left\{1 + \frac{1}{\tau^2 d_i^2}\left(\frac{1}{z_i} - 1\right)\right\}^{-(a+b)}\ \frac{1}{z_i^2}\mathrm{d}z_i \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{substitute} z_i = 1/(1 + \tau^2 d_i^2\rho_i) \\
&\propto \prod_{i=1}^{n} \int_0^1 \exp\left\{-\frac{\hat{\alpha}_i^2 d_i^2 z_i}{2\sigma^2}\right\} z_i^{b+\frac{1}{2}-1}(1 - z_i)^{a-1}\ \left\{\frac{1}{\tau^2 d_i^2} + \left(1 - \frac{1}{\tau^2 d_i^2}\right) z_i\right\}^{-(a+b)}\ \mathrm{d}z_i \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{The integral of the numerator of the density of CCH distribution}) \\
&\propto \prod_{i=1}^{n} H\left(b + \frac{1}{2}, a, a + b, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)
\end{aligned}$$

According to Lemma 3 of Gordy et al. [1998]

$$\frac{\partial}{\partial s} H(p, q, r, s, \nu, \theta) = H(p + 1, q, r, s, \nu, \theta)$$

Using this we have

$$m'(\hat{\alpha}_i) \propto -\frac{b+\frac{1}{2}}{a+b+\frac{1}{2}}\ H\left(b+\frac{3}{2}, a, a+b, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)\ \frac{\hat{\alpha}_i d_i^2}{\sigma^2}$$

$$\begin{aligned} m''(\hat{\alpha}_i) \propto & -\frac{b+\frac{1}{2}}{a+b+\frac{1}{2}}\ H\left(b+\frac{3}{2}, a, a+b, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)\ \frac{d_i^2}{\sigma^2} \\ & + \frac{\left(b+\frac{1}{2}\right)\left(b+\frac{3}{2}\right)}{\left(a+b+\frac{1}{2}\right)\left(a+b+\frac{3}{2}\right)}\ H\left(b+\frac{5}{2}, a, a+b, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)\ \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4} \end{aligned}$$

According to Gordy et al. [1998], for $X \sim \text{CCH}(p, q, r, s, \nu, \theta)$

$$\mathbb{E}(X^k) = \frac{(p)_k}{(p+q)_k}\ \frac{H(p+k, q, r, s, \nu\theta)}{H(p, q, r, s, \nu, \theta)}$$

where $(a)_k = (a+k-1)(a)_{k-1},$ and $(a)_0 = 1$. Therefore,

$$\begin{aligned} \frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} &= -\frac{\hat{\alpha}_i d_i^2}{\sigma^2}\mathbb{E}(Z_i) \\ \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} &= -\frac{d_i^2}{\sigma^2}\mathbb{E}(Z_i) + \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}\mathbb{E}(Z_i^2) \end{aligned}$$

where, $Z_i \sim \text{CCH}\left(b+\frac{1}{2}, a, a+b, \frac{\hat{\alpha}_i^2 d_i^2}{\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)$, and by part (a) of Theorem 1, we have $T = \sum_{i=1}^n T_i$, where the component-wise T_i values are

$$\begin{aligned} T_i &= 2\sigma^2 - \sigma^4 d_i^{-2}\left\{\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)}\right\}^2 + 2\sigma^4\frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} \\ &= 2\sigma^2 - 2\sigma^2\mathbb{E}(Z_i) - \hat{\alpha}_i^2 d_i^2\{\mathbb{E}(Z_i)\}^2 + 2\hat{\alpha}_i^2 d_i^2\mathbb{E}(Z_i^2) \\ &= 2\sigma^2[1 - \mathbb{E}(Z_i) + 2s_i\mathbb{E}(Z_i^2) - s_i\{\mathbb{E}(Z_i)\}^2] \end{aligned}$$

□