Bayesian Variable Selection and Estimation Based on Global-Local Shrinkage Priors

Xueying Tang and Malay Ghosh
University of Florida, Gainesville, USA
Xiaofan Xu
Stubhub Inc., Boston, USA
Prasenjit Ghosh
Presidency University, Kolkata, India

Abstract

We consider in this paper simultaneous Bayesian variable selection and estimation for linear regression models with global-local shrinkage priors on the regression coefficients. We propose a variable selection procedure that selects a variable if the ratio of the posterior mean of its regression coefficient to the corresponding ordinary least square estimate is greater than a half. The regression coefficient is estimated by the posterior mean or zero depending on whether the corresponding variable is selected or not. Under the assumption of orthogonal designs, we prove that if the local parameters have polynomial-tailed priors, the proposed method enjoys the oracle property in the sense that it can achieve variable selection consistency and optimal estimation rate at the same time. However, if, instead, an exponential-tailed prior is used for the local parameters, the proposed method has variable selection consistency but not the optimal estimation rate. We show via simulation and real data examples that our proposed selection mechanism works for nonorthogonal designs as well.

AMS (2000) subject classification. Primary 62F15, 62J05; Secondary 62J07. Keywords and phrases. Half-thresholding, Optimal estimation rate, Variable selection consistency.

1 Introduction

The objective of this article is simultaneous variable selection and estimation in linear regression models with global-local shrinkage priors on the regression coefficients. Selection of the best available model among a set of candidate models is extremely useful for most statistical applications. The problem often reduces to the choice of a subset of variables from all predictive variables in a regression setting. Linear regression models continue to occupy a prominent place in variable selection problems due to their interpretability
as well as analytical tractability. Throughout this paper, we consider classical linear regression models with response vector \( \mathbf{Y} = (y_1, \ldots, y_n) \) and a set of predictors \( \mathbf{x}_1, \ldots, \mathbf{x}_p \). The target is to fit a model of the form

\[
\mathbf{Y} = \sum_{i=1}^{p} \mathbf{x}_i \beta_i + \mathbf{\epsilon} = \mathbf{X}\beta + \mathbf{\epsilon},
\]

(1.1)

where \( \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \), \( \beta = (\beta_1, \ldots, \beta_p)^T \), and \( \mathbf{\epsilon} \sim N(0_n, \sigma^2 \mathbf{I}_n) \). The goal of variable selection is to pick only a subset of predictors that are relevant for explaining or predicting a given response.

Historically, penalized regression methods have been very successful for variable selection. In its general form, the problem reduces to minimization of the objective function

\[
f(\beta) = \| \mathbf{Y} - \mathbf{X}\beta \|^2 + \lambda u(\beta),
\]

(1.2)

where \( \lambda \) is the penalty parameter. The choice \( u(\beta) = \sum_{j=1}^{p} \beta_j^2 \) leads to the ridge estimator (Hoerl and Kennard, 1970), while \( u(\beta) = \sum_{j=1}^{p} |\beta_j| \) leads to the lasso estimator of \( \beta \) (Tibshirani, 1996). One of the advantages of the lasso estimator is that it can produce exact zero estimates for some of the regression coefficients. Despite this distinctive feature, Zou (2006) showed that lasso estimators cannot achieve variable selection consistency and optimal estimation rate at the same time. He proposed instead the adaptive lasso which heavily penalized zero coefficients and moderately penalized large coefficients using data dependent weights for different coefficients. Specifically, the adaptive lasso estimates \( \hat{\beta}_{\text{adap}} \) are found as the solution of the minimization problem (1.2) with \( u(\beta) = \sum_{i=1}^{p} |\hat{\beta}_i|/|\hat{\beta}_i|^{\gamma} \) for some \( \gamma > 0 \), where \( \hat{\beta}_i \) is the least squares estimator of \( \beta_i \). The adaptive lasso enjoys the oracle property in the sense that it achieves simultaneously variable selection consistency and asymptotic normality with \( \sqrt{n} \) convergence rate. One important feature, implicit in the proof of Theorem 2 of Zou (2006), is that \( \hat{\beta}_{\text{adap}}/\hat{\beta}_i \overset{P}{\to} 0 \) or 1 according as the true coefficient value equals or is different from zero. This property is the main motivation for us to develop a new thresholding method in a Bayesian context.

In Bayesian context, approaches to variable selection for linear regressions fall in two categories. Approaches in the first category assume priors on the model space of \( 2^p \) possible submodels and on the parameters of each submodel. Then, the model with the highest posterior model probability is chosen as the best. Examples of recent work along this line are Bayarri et al. (2012) and Maruyama et al. (2011). The other category of approaches starts
with the full model and then selects variables based on the posterior distributions of the regression coefficients. The method we proposed in this article falls in the latter category, which is closely related to penalized regression. Literature along this line goes a long way back starting with Mitchell and Beauchamp (1988) and has thrived recently. Some more recent development includes Johnson and Rossell (2012), Liang et al. (2013), and Ročková and George (2016).

The most intuitive class of priors for variable selection is spike-and-slab priors, also called two-group priors. A prior in this family is essentially a mixture of two distributions, spike and slab. The spike part is usually a point mass at zero or a distribution that concentrates around zero. It models the regression coefficients of irrelevant variables, which are essentially zero. The slab part is a relatively flat distribution and describes the nonzero coefficients of relevant variables. In Mitchell and Beauchamp (1988), the spike part is a probability mass at zero and the slab part is a uniform distribution with a large symmetric range. The stochastic search variable selection proposed in George and McCulloch (1993) used a mixture of two normal distributions with different variances for the coefficients. Geweke (1996) used positive mass at zero for the spike part and a normal distribution for the slab part. Ročková and George (2016) considered Laplace priors for both spike and slab parts. Narisetty and He (2014) and Castillo et al. (2015) provide thorough theoretical examination on the performance of the spike-and-slab priors. Xu and Ghosh (2015) also considered spike-and-slab priors, but used median thresholding to select variables in the group lasso framework. Ishwaran and Rao (2005) considered rescaled spike-and-slab priors, which still adapted the idea of spike-and-slab priors, but rescaled the response to prevent the shrinkage effect vanishing as the sample size grows.

Another popular class of priors used for Bayesian variable selection is called one-group shrinkage priors. The Bayesian lasso introduced by Park and Casella (2008) is an early example. The authors performed a full Bayesian analysis analogous to lasso, interpreting $\|Y - X\beta\|^2$ in (1.2) as the negative of a multiple of the log-likelihood and the penalty function $\lambda \sum_{i=1}^{p} |\beta_i|$ as the negative of the logarithm of a double exponential prior for $\beta_i$. Following a similar idea, Li and Lin (2010) proposed the Bayesian elastic net.

Unlike the spike-and-slab priors, these shrinkage priors cannot naturally produce exact zero estimates of the regression coefficients with positive probability. Thus, a critical question to answer when using shrinkage priors for variable selection is how to select the relevant variables. Li and Lin (2010) presented the credible interval criterion which selects predictor
x_i if the credible interval of \( \beta_i \) does not cover zero. A criterion called scaled neighborhood criterion is also considered in Li and Lin (2010). It selects predictors with posterior probability of belonging to the interval \([-\sqrt{\text{Var}(\beta_j \mid Y)}, \sqrt{\text{Var}(\beta_j \mid Y)}]\) less than a certain threshold. Bhattacharya et al. (2015) proposed Dirichlet-Laplace priors and a heuristic variable selection procedure based on clustering results for each MCMC iteration. These authors did not address the issue of any oracle property of their procedures. Recently, Hahn and Carvalho (2015) proposed selection of variables by minimizing the decoupled shrinkage and selection loss function after finding the posterior mean of \( \beta \). However, a surrogate optimization problem has to be used since the original one is intractable in the presence of moderate to large number of predictors.

In this paper, we consider the problem of variable selection and estimation using global-local shrinkage priors, which includes the prior of Park and Casella (2008) as a special case. Specifically, we assume that the prior distribution of \( \beta_i \) is a scale mixture of normals:

\[
\beta_i \mid \gamma_i, \tau \sim \text{ind} \ N(0, \sigma^2 \gamma_i \tau), \\
\gamma_i \mid \tau \sim \text{ind} \ \pi(\gamma_i).
\]

These priors approximate the spike-and-slab priors, but instead are unimodal and absolutely continuous. They place significant probability mass around zero and have heavy tails to signify the inclusion of relevant variables. The local parameters \( \gamma_i \) control the degree of shrinkage of each individual \( \beta_i \) while the global parameter \( \tau \) causes an overall shrinkage. We give a list of such priors in a later section. The list includes not only the now famous horseshoe prior of Carvalho et al. (2010), but several other priors considered, for example, in Griffin and Brown (2010) & Griffin and Brown (2011), Polson and Scott (2010) & Polson and Scott (2012), and Armagan et al. (2011) & Armagan et al. (2013). We also find it convenient to classify the priors \( \pi(\gamma_i) \) into two subclasses: those with exponential tails and those with polynomial tails. We propose a thresholding procedure to select relevant variables in the model. It turns out that the theoretical properties of our proposed method are closely related to the tails of \( \pi(\gamma_i) \). As we will show in the subsequent sections, if polynomial-tailed priors are used, the proposed method attains the oracle property for certain choice of \( \tau \) in the same way as the adaptive lasso. In contrast, the exponential-tailed priors, while attaining variable consistency for some choice of \( \tau \), will fail to attain asymptotic normality at the \( \sqrt{n} \) rate.
The theoretical results in this paper are derived under the assumption of orthogonal designs. Although it looks restrictive, there are applications such as wavelet shrinkage problems, that satisfy this assumption. We want to emphasize that orthogonality is assumed only for the derivation of the theoretical results. The proposed variable selection and estimation procedure can be applied to regressions with general design matrices. Through simulations and real data analyses, we show that our proposed procedure have competitive performance even if the design matrices are not orthogonal.

The outline of the remaining sections is as follows. The general class of shrinkage priors and the proposed procedure are described in Section 2. In Section 3, we present the theoretical properties of the proposed method for orthogonal designs. Section 4 contains simulation results. Data analyses are included in Section 5. Some final remarks are made in Section 6. The technical proofs are included in the appendices.

2 Global-Local Shrinkage Priors and the Half-Thresholding Method

For clarity, we reiterate the model considered in this article:

\[ Y \sim N(X\beta, \sigma^2 I_n), \quad (2.1) \]
\[ \beta_i | \gamma_i, \tau \overset{\text{ind}}{\sim} N(0, \sigma^2 \gamma_i \tau), \quad i = 1, \ldots, p, \quad (2.2) \]
\[ \gamma_i \overset{\text{ind}}{\sim} \pi(\gamma_i), \quad i = 1, \ldots, p. \quad (2.3) \]

Throughout this article, we assume \( p \equiv p_n \leq n \).

Many priors in Bayesian literature can be expressed in the form of scale mixture of normals, as in (2.2) and (2.3). Table 1, given later in Section 3, presents a list of such priors (of \( \beta_i \)) and the corresponding form of \( \pi(\gamma_i) \). By employing two levels of parameters to express the variances in (2.2), the global-local shrinkage priors assign large probabilities around zero while assigning non-trivial probabilities to values far from zero. The global parameter \( \tau \) tends to shrink all \( \beta_i \)'s towards zero. At the same time, the local parameters \( \gamma_i \) control the degree of shrinkage of each individual \( \beta_i \). If \( \pi(\gamma_i) \) is appropriately heavy-tailed, the coefficients of important variables are left almost unshrunk.

In the same spirit as Park and Casella (2008), placing a prior on \( \beta_i \) is closely related with adding a penalty term of \( \beta_i \) to the ordinary least square objective function, so the properties of penalized regression estimators can
Table 1: A list of global-local shrinkage priors of $\beta_i$. The third column indicates whether the priors have a polynomial tail (P) or an exponential tail (E).

| Prior          | $\pi(\gamma_i)$                                             | Class | Reference                                    |
|----------------|-------------------------------------------------------------|-------|----------------------------------------------|
| Laplace        | $\exp\{-b\gamma_i\}$                                       | E     | Park and Casella (2008)                      |
| Student’s $T$  | $\gamma_i^{-a} \exp(-a/\gamma_i)$                         | P     | Carvalho et al. (2010)                       |
| Horseshoe      | $\gamma_i^{-1/2}(1+\gamma_i)^{-1}$                         | P     | Bhadra et al. (2017)                         |
| Horseshoe+     | $\gamma_i^{-1/2}(\gamma_i - 1)^{-1}\log(\gamma_i)$       | P     | Bhadra et al. (2017)                         |
| NEG            | $(1 + \gamma_i)^{-1-a}$                                    | P     | Armagan et al. (2005)                        |
| TPBN           | $\gamma_i^{u-1}(1 + \gamma_i)^{-a-u}$                     | P     | Armagan et al. (2011)                        |
| GDP            | $\int_0^\infty \lambda^{2a-1}\exp\left(-\frac{\lambda^2\gamma_i}{2}\right)d\lambda$ | P     | Armagan et al. (2013)                        |
| HIB            | $\gamma_i^{u-1}(1 + \gamma_i)^{-(a+u)}$                 | P     | Polson and Scott (2012)                      |
|                | $\exp\left\{-\frac{s}{1+\gamma_i}\right\}\left\{\phi^2 + \frac{1-\phi^2}{1+\gamma_i}\right\}^{-1}$ |       |                                              |

This sheds light on the features of Bayesian estimator of $\beta_i$. The proof of Theorem 2 in Zou (2006) implies that under mild conditions,

$$\frac{\hat{\beta}_{\text{adap}}^i}{\hat{\beta}_i} \xrightarrow{p} \begin{cases} 0, & \text{when } \beta_0^i = 0, \\ 1, & \text{when } \beta_0^i \neq 0, \end{cases} \quad (2.4)$$

where $\beta_0^i$ is the true value of $\beta_i$ and $\hat{\beta}_i$ is the ordinary least square estimator of $\beta_i$. This indicates that the adaptive lasso estimator for the coefficient of an irrelevant variable converges to zero faster than the least square estimator. In fact, (2.4) holds by replacing the adaptive lasso estimator with any penalized regression estimator that has the oracle property described in Zou (2006). Many of these estimators can also be interpreted as posterior modes of priors specified in (2.1)–(2.3). Due to the asymptotic closeness of posterior means and posterior modes under such priors, one can threshold the ratio of posterior mean and least square estimator to obtain an oracle variable selection procedure even though the posterior mean is not sparse. Motivated by this, we propose to select predictor $x_i$ if

$$\left|\frac{\hat{\beta}_{PM}^i}{\hat{\beta}_i}\right| > 0.5, \quad (2.5)$$

where $\hat{\beta}_{PM}^i$ is the posterior mean of $\beta_i$ under certain shrinkage prior. We refer to this procedure as half-thresholding (HT) and define the HT estimator of $\beta_i$ as

$$\hat{\beta}_i^{\text{HT}} = \hat{\beta}_{PM}^i I\left(\left|\frac{\hat{\beta}_{PM}^i}{\hat{\beta}_i}\right| > 0.5\right). \quad (2.6)$$
Our proposed HT procedure is simple and easy to implement. Once the posterior mean and the ordinary least square estimate of $\beta$ are obtained, variable selection can be performed without any extra optimization step as required for example in Hahn and Carvalho (2015). Besides its simplicity, as we will show in the next section, the HT procedure enjoys oracle properties for orthogonal designs if the global parameter $\tau$ and the prior of $\gamma_i$ are chosen appropriately.

Replacing 0.5 by a positive fraction in (2.5) and (2.6) will not change the theoretical results in Section 3, but it is not clear whether there is any particular advantage in an alternative choice. The idea of using 0.5 is to maintain some neutrality regarding inclusion and exclusion of a certain predictor. We will examine the effects of using different thresholds in Section 4.

### 3 Theoretical Results

We consider two general types of priors $\pi(\gamma_i)$ given by

$$
\pi(\gamma_i) = \gamma_i^{-a-1}L(\gamma_i), \quad a > 0,
$$

(3.1)

$$
\pi(\gamma_i) = \exp(-b\gamma_i)L(\gamma_i), \quad b > 0,
$$

(3.2)

where $L(\cdot)$ is a nonnegative slowly varying function in Karamata’s sense (Bingham et al., 1987, p. 6) defined on $(0, \infty)$. We call the priors in the form of (3.1) and (3.2) polynomial-tailed priors and exponential-tailed priors, respectively. Later in this section, we will show that the theoretical performances of the HT method is closely related to the tails of the prior of $\gamma_i$. Table 1 provides a list of commonly used scale mixture priors of $\beta_i$ and the corresponding form of $\pi(\gamma_i)$.

In this section, we assume the design matrix is orthogonal, that is $X^T X = nI_p$. With this assumption, $E(\beta_i | \gamma_i, \tau, Y) = (1 - s_i)\hat{\beta}_i$, where $s_i = 1/(1 + n\tau\gamma_i)$, is the shrinkage factor. By law of iterated expectations, $\hat{\beta}_i^{PM} = E(\beta_i | Y) = (1 - E(s_i | Y))\hat{\beta}_i$. Therefore, with the orthogonal design matrix assumption, the selection criterion (2.5) of the proposed method simplifies to

$$
1 - E(s_i | Y) > 0.5.
$$

(3.3)

A similar procedure was considered by Carvalho et al. (2010) and Ghosh et al. (2016) in other contexts.

Following Fan and Li (2001) and Zou (2006), we say a variable selection procedure is oracle if it results in both variable selection consistency and
optimal estimation rate. Let \( \mathcal{A} = \{ j : \beta_j^0 \neq 0 \} \) and \( \mathcal{A}_n = \{ j : \hat{\beta}_j^{\text{HT}} \neq 0 \} \). The variable selection consistency means
\[
\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) = 1, \quad \text{as } n \to \infty,
\]
while the optimal estimation rate means
\[
\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{\text{HT}} - \beta_0^\mathcal{A}) \overset{d}{\to} N(0, \sigma^2 I_{p_0}), \quad \text{as } n \to \infty,
\]
where \( p_0 \) is the cardinality of \( \mathcal{A} \) and it does not depend on \( n \).

Another point to clarify before the presentation of our theoretical results is the treatment of the global parameter \( \tau \). In Datta and Ghosh (2013) and part of the results of Ghosh et al. (2016), \( \tau \) was treated as a tuning parameter. Carvalho et al. (2010) considered a full Bayesian treatment and a half-Cauchy prior for the global parameter. Ghosh et al. (2016) also provided some results when an empirical Bayes estimate of the global parameter is used. In this article, we treat \( \tau \) as a tuning parameter or assume a hyperprior for it. To distinguish the two treatments, we write \( \tau \) as \( \tau_n \) when it is used as a tuning parameter.

### 3.1. Properties of Shrinkage Factors

By (3.3), the HT procedure is closely related to the shrinkage factor \( s_i \), so we present its properties first.

**Proposition 1.** Suppose the prior of \( \gamma_i \) is proper. For \( i \notin \mathcal{A} \), if \( n\tau_n \to 0 \), as \( n \to \infty \), then \( E(1 - s_i | \tau_n, Y) \overset{p}{\to} 0 \) as \( n \to \infty \). For \( i \in \mathcal{A} \),

1. if \( \gamma_i \) has a polynomial-tailed prior described in (3.1) and \( n\tau_n \to 0 \), \( \log(\tau_n)/n \to 0 \) as \( n \to \infty \), then \( E(1 - s_i | \tau_n, Y) \overset{p}{\to} 1 \), as \( n \to \infty \).

2. if \( \gamma_i \) has an exponential-tailed prior described in (3.2) and \( n\tau_n \to 0 \) and \( n^2\tau_n \to \infty \) as \( n \to \infty \), then \( E(1 - s_i | \tau_n, Y) \overset{p}{\to} 1 \), as \( n \to \infty \).

Proposition 1 shows that, regardless of the choice of the prior of \( \gamma_i \) in the given class, the HT procedure can identify an irrelevant variable correctly if \( \tau_n \) goes to zero at a rate faster than \( n^{-1} \). On the other hand, \( \tau_n \) should not converge to zero too fast in order to avoid overshrinkage and to correctly identify relevant variables. The conditions \( \log(\tau_n)/n \to 0 \) and \( n^2\tau_n \to \infty \) serve this idea for polynomial-tailed priors and exponential-tailed priors, respectively. Given \( n\tau_n \to 0 \), the condition \( n^2\tau_n \to \infty \) is more stringent than \( \log(\tau_n)/n \to 0 \). Intuitively, this has to be the case since exponential tails are lighter. To guarantee that the coefficients of relevant variables are not overly shrunk, the global parameter should decay at a slower rate than that for polynomial-tailed priors, and compensate the amount of shrinkage brought by exponential-tailed local parameters.
3.2. Polynomial-Tailed Priors We say a sequence of positive real numbers \( \{t_n\}_{n=1}^{\infty} \) satisfies poly-a condition if there exists \( \epsilon \in (0, a) \) such that

\[
p_n(nt_n)^{\epsilon} \to 0 \quad \text{and} \quad \log(t_n)/\sqrt{n} \to 0, \quad \text{as} \quad n \to \infty.
\]

The choice \( t_n = n^{-1-2a} \) satisfies the poly-a condition since \( p_n \leq n \).

If \( p \) does not vary with \( n \), the condition can be simplified to \( nt_n \to 0 \) and \( \log(t_n)/\sqrt{n} \to 0 \).

**Theorem 1.** Suppose a proper polynomial-tailed prior of the form (3.1) is assumed for \( \gamma_i, i = 1, \ldots, p \) with \( 0 < a < 1 \). If \( \{\tau_n\}_{n=1}^{\infty} \) satisfies the poly-a condition, then the HT procedure attains the oracle property.

As Theorem 1 demonstrates, if \( \tau_n \) is chosen to decay to zero at an appropriate rate, the HT procedure has the oracle property. This suggests if a hyperprior \( \pi_{\tau_n} \) of \( \tau \) concentrates most of its probability mass in an interval with its end points satisfying the poly-a condition, then the HT threshold should still enjoy the oracle property. With this observation, we have the following result.

**Corollary 1.** Suppose that a proper polynomial-tailed prior of the form (3.1) is assumed for \( \gamma_i, i = 1, \ldots, p \) with \( 0 < a < 1 \). We also place a prior \( \pi_{\tau_n} \) with support \( (\xi_n, \psi_n) \) on \( \tau \). If both \( \{\xi_n\}_{n=1}^{\infty} \) and \( \{\psi_n\}_{n=1}^{\infty} \) satisfy the poly-a condition, then the HT procedure is oracle.

3.3. Exponential-Tailed Priors Now, we examine the properties of HT procedure when exponential-tailed priors are assumed for the local parameters \( \gamma_i \).

**Theorem 2.** Suppose a proper exponential-tailed prior of the form (3.2) is assumed for \( \gamma_i, i = 1, \ldots, p \) and \( \int_{0}^{\infty} \gamma_i \pi(\gamma_i) d\gamma_i < \infty \). If \( nt_n \to 0, n^2 \tau_n \to \infty \) and \( (p_n n \tau_n)^2 / \log(n \tau_n) \to 0 \), as \( n \to \infty \), then the HT procedure achieves variable selection consistency.

**Remark.** The choice \( \tau_n = \log \log n/n^2 \) satisfies the conditions in Theorem 2. Similar to what we have mentioned in Section 3.1, these conditions are more stringent than the poly-a condition since exponential tails are lighter than polynomial ones. If \( p \) does not depend on \( n \), the conditions simplify to \( nt_n \to 0 \) and \( n^2 \tau_n \to \infty \), as \( n \to \infty \). If \( \pi(\gamma_i) = \exp(-\gamma_i) \), then the marginal prior of \( \beta_i \) is proportional to \( \exp(-|\beta_i|/\sqrt{\tau_n}) \). Thus, \( 1/\sqrt{\tau_n} \) corresponds to the penalty parameter \( \lambda_n \) in the lasso estimator. The two conditions on \( \tau_n \) assuming fixed \( p \) can be translated to \( \lambda_n/\sqrt{n} \to \infty \) and \( \lambda_n/n \to 0 \), which is a sufficient condition for the lasso estimator to be model selection consistent assuming the irrepresentable condition (Zou, 2006).
Theorem 3. Suppose a proper exponential-tailed prior of the form (3.2) and there exist 0 < m ≤ M < ∞ such that \( m < L(t) < M \) for all \( t \in (0, \infty) \). If \( n\tau_n \to 0 \) and \( n^2 \tau_n \to \infty \), then for \( i \in A \), with probability one,

\[
\frac{m}{M} S_n^{(i)} \leq n\sqrt{\tau_n} \left( \hat{\beta}_i^{HT} - \beta_i^0 \right) \leq \frac{M}{m} S_n^{(i)},
\]

where \( \{S_n^{(i)}, n \geq 1\} \) are sequences of random variables and

\[
S_n^{(i)} \overset{p}{\to} -\sqrt{2b}\sigma \text{sign}(\beta_i^0).
\]

Remark. Theorem 2 shows that, with exponential-tailed priors on local parameters, the HT procedure can achieve variable selection consistency when \( \tau_n \) vanishes at certain rate. However, Theorem 3 tells us that the procedure cannot achieve optimal estimation rate with \( \tau_n \) decaying at this rate. The boundedness condition in Theorem 3 looks restrictive, but all the three exponential-tailed distributions listed in Table 1 satisfy this condition. If \( \beta_i \) has a double exponential prior as in Park and Casella (2008), \( L(t) = 1 \), for all \( t > 0 \). In this case, we have

\[
n\sqrt{\tau_n} \left( \hat{\beta}_i^{HT} - \beta_i^0 \right) \overset{p}{\to} -\sqrt{2b}\sigma \text{sign}(\beta_i^0).
\]

Proposition 2. If the prior of \( \gamma_i \), \( \pi(\gamma_i) \), satisfies the condition

\[
\int_0^{\infty} \gamma_i^{-1/2} \pi(\gamma_i) d\gamma_i < \infty, \quad \text{for} \quad i = 1, \ldots, p,
\]

then the HT procedure cannot achieve variable selection consistency when \( n\tau_n \to c \in (0, \infty) \) as \( n \to \infty \).

Remark. Proposition 2 tells us that the HT procedure can not consistently select relevant variables if \( \tau_n \) does not converge to zero fast enough. This proposition holds for both polynomial-tailed and exponential-tailed priors. The finite integral condition is not very restrictive for exponential-tailed priors of \( \gamma_i \). In fact, the three exponential-tailed priors in Table 1 satisfy the condition. However, it excludes the horseshoe and some other polynomial-tailed priors.

Proposition 3. Suppose \( \pi(\gamma_i) \) is a proper exponential-tailed prior of the form (3.2) and there exist 0 < m ≤ M < ∞ such that

\[
m < L(t) < M \quad \text{for all} \quad t \in (0, \infty).
\]

If \( n\tau_n \to 0 \) as \( n \to \infty \), the HT procedure cannot achieve optimal estimation rate.
Remark. The proof of Proposition 3 implies that if $\tau_n$ goes to zero faster than $n^{-1}$, the HT procedure overshinks the nonzero coefficients and the convergence rate is slower than $n^{1/2}$.

According to Proposition 2, the condition $n\tau_n \to 0$ is necessary for variable selection consistency of exponential-tailed priors. As a result, Proposition 3 suggests that if $\tau_n$ is chosen in a way to ensure variable selection consistency, then optimal estimation rates cannot be achieved. Combining the two propositions, we have the following theorem:

**Theorem 4.** If $\pi(\gamma_i)$ is a proper exponential-tailed prior of the form (3.2) and it satisfies conditions (3.4) and (3.5), then the HT procedure does not have the oracle property for any choice of $\tau_n$.

Remark. As a special case, Theorem 4 implies that the HT procedure lacks oracle property with the prior introduced in Park and Casella (2008).

### 4 Simulation Results

In this section, we investigate the variable selection and parameter estimation performance of global-local shrinkage priors and the HT procedure. TPBN priors with $a = 0.5$, $u = 0.5$ and $a = 0.1$, $u = 0.5$ are used as representatives of polynomial-tailed priors. The former choice of parameters gives the horseshoe prior, while the latter leads to a prior with tail lighter than that of the horseshoe. The Laplace prior with $b = 1$ is used as a representative of exponential-tailed priors. The models with these three global-local priors are denoted by HS, TPBN, and LA, respectively. For each prior, HT procedure (2.5) was applied to perform variable selection. We also employed an half-thresholding method with the posterior mean $\hat{\beta}_{PM}^i$ in (2.5) replaced by the posterior median. This method is denoted by HTMed. For comparison, the spike-and-slab (SS) prior described in Castillo et al. (2015) is also used for the regression coefficients. Then, variable selection is achieved by either selecting a variable with posterior inclusion probability greater than 0.5 (Prob) or selecting the variables whose coefficients have nonzero posterior medians (Med). The Bayesian elastic net (BEN) with scaled neighborhood (SN) criterion in Li and Lin (2010) and the rescaled spike-and-slab (RSS) model with Zcut selection in Ishwaran and Rao (2005) were implemented as well. Two frequentist methods are also considered for comparison. One is Lasso (Tibshirani, 1996) and the other is SCAD (Fan and Li, 2001). The penalty parameters $\lambda$ in both frequentist methods are chosen by five-fold cross-validation.

We generated data from the linear regression model (1.1), where $n = 20$, $\sigma = 1, 3, 5$, and $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. The design matrix $X$ is generated
from multivariate normal distributions with mean $0$ and covariance matrix $\Sigma$ where the entry in the $i$th row and $j$th column of $\Sigma$ is $\rho^{|i-j|}$. We consider three choices of $\rho$, 0, 0.5, and 0.9, which represent uncorrelated, moderately correlated, and highly correlated covariates, respectively. For each combination of the parameters, 50 training and 50 test datasets were generated. Each model mentioned above was fit on the training data sets and variable selection is performed using corresponding procedures.

To measure the variable selection accuracy, we compared the misclassification probability (MP), false positive rate (FPR), and false negative rate (FNR) of each methods. The three criteria were calculated for each dataset and the averaged values over 50 datasets were reported in Tables 2, 3, and 4. Several observations can be made from these tables. First, TPBN+HTMed and RSS+Zcut produce the lowest MPs in general. The former works best when the correlation between covariates is low or moderate. Second, although the two methods using Laplace priors lead to FNR comparable to those produced by TPBN+HTMed and RSS+Zcut, they usually have much higher FPR, thus producing high MPs. Third, variable selection procedures utilizing posterior medians usually lead to lower FPRs but higher FNRs when compared to those procedures using posterior means. Last but not the least, for the two frequentist methods, SCAD usually gives lower MPs and FPRs while LASSO gives lower FNRs.

Table 2: Variable selection performance of various methods for $\rho = 0$. The methods with the lowest or the second lowest MP are marked by asterisks

|               | $\sigma = 1$   |               | $\sigma = 3$   |               | $\sigma = 5$   |
|---------------|---------------|---------------|---------------|---------------|---------------|
|               | MP | FPR | FNR | MP | FPR | FNR | MP | FPR | FNR |
| BEN + SN      | 0.28 | 0.12 | 0.54 | 0.32 | 0.03 | 0.79 | 0.35 | 0.03 | 0.89 |
| RSS + Zcut    | 0.08* | 0.12 | 0.00 | 0.18* | 0.12 | 0.27 | 0.28* | 0.10 | 0.59 |
| SS + Prob     | 0.23 | 0.38 | 0.00 | 0.28 | 0.35 | 0.17 | 0.28* | 0.18 | 0.47 |
| SS + Med      | 0.09 | 0.14 | 0.00 | 0.21 | 0.22 | 0.21 | 0.28* | 0.15 | 0.51 |
| HS + HT       | 0.35 | 0.56 | 0.00 | 0.37 | 0.39 | 0.34 | 0.36 | 0.22 | 0.58 |
| HS + HTMed    | 0.15 | 0.24 | 0.01 | 0.24 | 0.15 | 0.40 | 0.34 | 0.11 | 0.73 |
| TPBN + HT     | 0.18 | 0.28 | 0.01 | 0.28 | 0.30 | 0.25 | 0.32 | 0.25 | 0.44 |
| TPBN + HTMed  | 0.06* | 0.09 | 0.01 | 0.17* | 0.10 | 0.29 | 0.27* | 0.12 | 0.52 |
| LA + HT       | 0.56 | 0.89 | 0.00 | 0.48 | 0.58 | 0.32 | 0.41 | 0.32 | 0.56 |
| LA + HTMed    | 0.54 | 0.87 | 0.00 | 0.42 | 0.48 | 0.33 | 0.39 | 0.24 | 0.64 |
| LASSO         | 0.28 | 0.46 | 0.00 | 0.35 | 0.49 | 0.13 | 0.38 | 0.37 | 0.40 |
| SCAD          | 0.09 | 0.14 | 0.00 | 0.30 | 0.38 | 0.27 | 0.34 | 0.27 | 0.59 |
Table 3: Variable selection performance of various methods for $\rho = 0.5$. The methods with the lowest or the second lowest MP are marked by asterisks

| Method         | $\sigma = 1$ |       |       | $\sigma = 3$ |       |       | $\sigma = 5$ |       |       |
|----------------|--------------|-------|-------|--------------|-------|-------|--------------|-------|-------|
|                | MP | FPR | FNR | MP | FPR | FNR | MP | FPR | FNR |
| BEN + SN       | 0.30 | 0.25 | 0.39 | 0.28 | 0.20 | 0.41 | 0.29 | 0.09 | 0.63 |
| RSS + Zcut     | 0.10* | 0.16 | 0.00 | 0.18* | 0.13 | 0.27 | 0.24* | 0.06 | 0.55 |
| SS + Prob      | 0.40 | 0.64 | 0.00 | 0.32 | 0.41 | 0.16 | 0.27 | 0.15 | 0.47 |
| SS + Med       | 0.12 | 0.20 | 0.00 | 0.21 | 0.20 | 0.25 | 0.24* | 0.08 | 0.51 |
| HS + HT        | 0.26 | 0.41 | 0.00 | 0.29 | 0.31 | 0.26 | 0.29 | 0.22 | 0.42 |
| HS + HTMed     | 0.14 | 0.22 | 0.00 | 0.23 | 0.14 | 0.37 | 0.28 | 0.11 | 0.57 |
| TPBN + HT      | 0.14 | 0.22 | 0.00 | 0.25 | 0.26 | 0.25 | 0.30 | 0.26 | 0.37 |
| TPBN + HTMed   | 0.04* | 0.07 | 0.00 | 0.18* | 0.10 | 0.33 | 0.27 | 0.10 | 0.54 |
| LA + HT        | 0.53 | 0.84 | 0.00 | 0.38 | 0.49 | 0.19 | 0.34 | 0.32 | 0.39 |
| LA + HTMed     | 0.51 | 0.82 | 0.00 | 0.34 | 0.40 | 0.23 | 0.33 | 0.22 | 0.51 |
| LASSO          | 0.29 | 0.47 | 0.00 | 0.30 | 0.39 | 0.14 | 0.32 | 0.38 | 0.21 |
| SCAD           | 0.11 | 0.17 | 0.00 | 0.25 | 0.29 | 0.27 | 0.31 | 0.21 | 0.55 |

To examine the performance in estimating the regression coefficients $\beta$, the squared $L_2$ norm of the difference between an estimate $\hat{\beta}$ and the true $\beta$ (BSE) is calculated and the averaged values are reported in Tables 5, 6, and 7. These tables also include the averaged relative prediction errors (RPE) $||Y_0 - X_0\hat{\beta}||^2/\sigma^2$ computed on each test dataset. As we can see

Table 4: Variable selection performance of various methods for $\rho = 0.9$. The methods with the lowest or the second lowest MP are marked by asterisks

| Method         | $\sigma = 1$ |       |       | $\sigma = 3$ |       |       | $\sigma = 5$ |       |       |
|----------------|--------------|-------|-------|--------------|-------|-------|--------------|-------|-------|
|                | MP | FPR | FNR | MP | FPR | FNR | MP | FPR | FNR |
| BEN + SN       | 0.38 | 0.21 | 0.66 | 0.35 | 0.41 | 0.25 | 0.34 | 0.37 | 0.29 |
| RSS + Zcut     | 0.26 | 0.41 | 0.01 | 0.22* | 0.16 | 0.32 | 0.33* | 0.16 | 0.61 |
| SS + Prob      | 0.62 | 1.00 | 0.00 | 0.42 | 0.60 | 0.12 | 0.38 | 0.37 | 0.40 |
| SS + Med       | 0.24 | 0.36 | 0.05 | 0.23* | 0.16 | 0.35 | 0.34 | 0.16 | 0.64 |
| HS + HT        | 0.28 | 0.39 | 0.10 | 0.30 | 0.21 | 0.45 | 0.34 | 0.20 | 0.59 |
| HS + HTMed     | 0.20* | 0.22 | 0.16 | 0.30 | 0.12 | 0.61 | 0.32* | 0.08 | 0.72 |
| TPBN + HT      | 0.23 | 0.28 | 0.14 | 0.27 | 0.17 | 0.45 | 0.37 | 0.22 | 0.61 |
| TPBN + HTMed   | 0.16* | 0.15 | 0.18 | 0.26 | 0.08 | 0.56 | 0.34 | 0.10 | 0.73 |
| LA + HT        | 0.41 | 0.64 | 0.02 | 0.35 | 0.31 | 0.41 | 0.36 | 0.26 | 0.54 |
| LA + HTMed     | 0.36 | 0.56 | 0.02 | 0.32 | 0.25 | 0.43 | 0.35 | 0.18 | 0.62 |
| LASSO          | 0.28 | 0.44 | 0.01 | 0.31 | 0.37 | 0.21 | 0.40 | 0.38 | 0.43 |
| SCAD           | 0.24 | 0.26 | 0.01 | 0.32 | 0.20 | 0.32 | 0.35 | 0.22 | 0.61 |
Table 5: Estimation and prediction performance of various methods for $\rho = 0$. The methods with the lowest or the second lowest BSE or RPE are marked by asterisks.

| Method          | $\sigma = 1$ | $\sigma = 3$ | $\sigma = 5$ |
|-----------------|--------------|--------------|--------------|
|                 | BSE | RPE  | BSE | RPE  | BSE | RPE  |
| BEN + SN        | 10.70 | 11.49  | 13.72 | 2.42 | 14.38 | 1.52  |
| RSS + Zcut      | 0.48 | 1.49   | 4.91  | 1.42 | 10.79 | 1.44  |
| SS + Prob       | 0.40 | 1.43   | 4.57* | 1.38*| 10.58 | 1.42  |
| SS + Med        | 0.37* | 1.41*  | 4.73* | 1.41*| 11.12 | 1.45  |
| HS + HT         | 0.40 | 1.42*  | 6.13  | 1.53 | 11.84 | 1.44  |
| HS + HTMed      | 0.39 | 1.41*  | 6.22  | 1.53 | 12.64 | 1.48  |
| TPBN + HT       | 0.40 | 1.42*  | 4.97  | 1.43 | 10.38*| 1.40* |
| TPBN + HTMed    | 0.37* | 1.42*  | 4.98  | 1.44 | 10.28*| 1.41* |
| LA + HT         | 0.54 | 1.50   | 7.05  | 1.62 | 12.16 | 1.47  |
| LA + HTMed      | 0.53 | 1.49   | 6.90  | 1.60 | 12.61 | 1.49  |
| LASSO           | 0.45 | 1.44   | 5.40  | 1.46 | 13.29 | 1.57  |
| SCAD            | 0.37 | 1.45   | 6.10  | 1.48 | 14.65 | 1.59  |

from the tables, the prediction performance is closely related to the performance in estimating $\beta$. Although the method that produces the smallest BSE and RPE varies as the error variance changes, the methods with polynomial-tailed global-local shrinkage priors consistently lead to good results, especially when the correlation between covariates is low or moderate. Procedures based on posterior medians and posterior means do not perform significantly different in terms of parameter estimation and prediction. Although SCAD usually yields smaller MPs than LASSO, in the sense of parameter estimation and prediction, it outperforms LASSO only when the correlation between covariates is low or moderate and the noise level is low.

In the HT procedure, the threshold 0.5 is chosen to maintain some neutrality regarding inclusion and exclusion of a certain predictor for finite sample studies. Theoretically, it could be changed to any constant between zero and one. For a series of thresholds chosen on a fine grid between zero and one, we applied the thresholding procedure on the ratio of posterior mean with the TPBN prior to the ordinary least square estimate. Based on the results shown in Fig. 1, using 0.5 as a threshold gives reasonable estimation, prediction, and selection results in general. Using a lower threshold will not cause notable changes in RPE and BSE, but it can lead to a significant increase in FPR without much improvement in true positive rate especially when the noise level is low. If the noise level is low, being conservative in
Table 6: Estimation and prediction performance of various methods for $\rho = 0.5$. The methods with the lowest or the second lowest BSE or RPE are marked by asterisks.

| Method       | $\sigma = 1$ |          | $\sigma = 3$ |          | $\sigma = 5$ |          |
|--------------|--------------|----------|--------------|----------|--------------|----------|
|              | BSE  | RPE | BSE  | RPE | BSE  | RPE |
| BEN + SN     | 8.33 | 12.09 | 9.58 | 2.45 | 11.90 | 1.62 |
| RSS + Zcut   | 0.52 | 1.44 | 5.04* | 1.60* | 9.30 | 1.32 |
| SS + Prob    | 0.44 | 1.36 | 4.95* | 1.55* | 9.35 | 1.31 |
| SS + Med     | 0.37 | 1.30 | 5.41 | 1.61 | 10.01 | 1.34 |
| HS + HT      | 0.46 | 1.32 | 5.63 | 1.70 | 8.84* | 1.30* |
| HS + HTMed   | 0.41 | 1.31 | 6.23 | 1.81 | 9.99 | 1.36 |
| TPBN + HT    | 0.35* | 1.28* | 5.34 | 1.60* | 9.30 | 1.27* |
| TPBN + HTMed | 0.32* | 1.26* | 5.71 | 1.66 | 10.48 | 1.31 |
| LA + HT      | 0.72 | 1.55 | 5.93 | 1.79 | 8.69* | 1.35 |
| LA + HTMed   | 0.69 | 1.53 | 5.99 | 1.84 | 9.70 | 1.39 |
| LASSO        | 0.70 | 1.57 | 6.97 | 1.74 | 9.71 | 1.33 |
| SCAD         | 0.55 | 1.38 | 8.45 | 1.85 | 14.43 | 1.49 |

Selection by choosing a threshold higher than 0.5, say 0.75, could produce similar prediction results and a lower FPR without sacrificing too much in the true positive rate. However, when the noise level is high, a conservative choice of threshold could lead to higher RPE and BSE.

Table 7: Estimation and prediction performance of various methods for $\rho = 0.9$. The methods with the lowest or the second lowest BSE or RPE are marked by asterisks.

| Method       | $\sigma = 1$ |          | $\sigma = 3$ |          | $\sigma = 5$ |          |
|--------------|--------------|----------|--------------|----------|--------------|----------|
|              | BSE  | RPE | BSE  | RPE | BSE  | RPE |
| BEN + SN     | 12.21 | 27.52 | 9.01 | 2.74 | 10.09* | 1.45 |
| RSS + Zcut   | 2.35* | 1.51 | 6.71* | 1.38 | 14.02 | 1.33 |
| SS + Prob    | 2.40* | 1.36* | 7.23* | 1.26* | 14.86 | 1.21* |
| SS + Med     | 2.48 | 1.65 | 8.42 | 1.50 | 15.80 | 1.39 |
| HS + HT      | 3.30 | 1.77 | 11.37 | 1.81 | 14.90 | 1.47 |
| HS + HTMed   | 3.50 | 2.09 | 12.79 | 2.13 | 16.13 | 1.62 |
| TPBN + HT    | 3.39 | 1.81 | 9.87 | 1.60 | 18.08 | 1.45 |
| TPBN + HTMed | 3.59 | 1.98 | 11.40 | 1.87 | 18.92 | 1.59 |
| LA + HT      | 3.14 | 1.60 | 9.31 | 1.94 | 11.55* | 1.54 |
| LA + HTMed   | 3.05 | 1.67 | 9.21 | 2.11 | 12.16 | 1.70 |
| LASSO        | 3.31 | 1.46* | 11.32 | 1.33* | 25.79 | 1.19* |
| SCAD         | 6.36 | 1.73 | 25.80 | 1.57 | 51.08 | 1.31 |
Figure 1: Performance of the thresholding procedure for the TPBN prior with different thresholds. Left: averaged RPE against threshold. Middle: averaged BSE against threshold. Right: ROC curve

5 Data Analyses

5.1. Wavelet Shrinkage Example Wavelets are widely used to analyze time series and image data. Wavelet shrinkage aims at reconstruction of functions from shrunk wavelet coefficients of discretely sampled observations. Because of the nice properties of wavelet functions, wavelet shrinkage has the ability to filter out noises while preserving local signals. For $i = 1, \ldots, n$, let $z_i$ be the observations of unknown function $f$ at $t_i$, where $t_1, \ldots, t_n$ are equally spaced points. A simple model for $Z = (z_1, \ldots, z_n)'$ is $Z = f + \epsilon$, where $f = (f(t_1), \ldots, f(t_n))'$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, is a vector of independent and identically distributed normal errors with mean 0 and variance $\sigma^2$. Given a set of orthogonal wavelet basis functions, the above model can be expressed in wavelet domain through discrete wavelet transformation as $Y = \beta + \epsilon$, where $Y = WZ$, $\beta = Wf$, and $W$ is an orthogonal matrix corresponding to the wavelet basis functions. Wavelet shrinkage filters out noises by producing a shrunk estimate of $\beta$ with elements corresponding to high frequency noises set to zeros and estimate $f$ by the inverse discrete wavelet transformation of the shrunk estimate. In this sense, a wavelet shrinkage problem can be treated as a variable selection problem with response $Y$, an identity design matrix and coefficients $\beta$.

We use an example in Tomassi et al. (2015) to illustrate how our half-thresholding procedure can be applied to remove noise from a signal. We took a segment of a real electromyogram (EMG) data of a healthy person from the Physionet data bank and add independent Gaussian noise with variance 0.04 to the standardized EMG. The clean standardized signal and the noisy signals are plotted in the first two panels of Fig. 2. We use Daubechies
wavelet with four vanishing moments to perform discrete wavelet transformation and assume independent HS priors for wavelet coefficients $\beta$ and an inverse gamma prior with both shape and rate parameters being $10^{-10}$ for the global parameter. The half-thresholding estimator is used as a shrunk estimate of $\beta$. The estimated signal is presented in the third panel of Fig. 2. As a comparison, we alternatively use a spike-and-slab prior proposed in Castillo et al. (2015) to model $\beta$ and set an element of the posterior mean to zero if its posterior probability of being zero is greater than 0.5. In Clyde et al. (1998), a prior with the same spirit was used for wavelet shrinkage. The mean squared errors are 0.024 and 0.089 for global-local HT estimate and
the spike-and-slab estimate, respectively. From Fig. 2, we can also see that
the global-local HT estimate recovers the signal better than the spike-and-
slab estimate. The latter overshrinks most of the large signals and misses
some, for example the one near 0.075, completely.

5.2. Diabetes Example

The dataset used in this example are from Efron et al. (2004), which contains ten baseline variables (predictors): age, sex, body mass index (BMI), average blood pressure (MAP), six blood serum measurements (TC, LDL, HDL, TCH, LTH, GLU), and a quantitative measure of disease progression one year after baseline (response) for 442 diabetes patients. The baseline variables are standardized to have zero mean and unit $L_2$ norm. The response variable is centered to have zero mean. The ten methods discussed in Section 4 were applied and the results are presented in Table 8.

Four variables (BMI, MAP, HDL, LTG) are selected by all the methods. These four variables are the first four variables that enter the regression equation in Efron et al. (2004). Among all the methods, RSS+Zcut selects least variables while SS+Prob selects all the variables. The results based on the polynomial-tailed global-local shrinkage priors HS, and TPBN are very similar. The only difference is that TPBN+HTMed does not select GLU. Notice that methods with the Laplace prior tend to select more variables than those with polynomial-tailed global-local shrinkage priors. A possible reason for this is that the Laplace prior has a slower rate of going to infinity near zero.

5.3. Prostate Cancer Example

The data used in this example is from Stamey et al. (1987) and has been analyzed by many authors. The response

Table 8: Variable selection results for the diabetes example. Variables that are selected by a given method are indicated by 1 in the corresponding entry.

| Method         | Age | Sex | BMI | MAP | TC  | LDL | HDL | TCH | LTG | GLU |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| RSS + Zcut     | 0   | 0   | 1   | 1   | 0   | 0   | 1   | 0   | 1   | 0   |
| BEN + SN       | 1   | 1   | 1   | 1   | 0   | 0   | 1   | 1   | 1   | 1   |
| SS + Prob      | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   |
| SS + Med       | 0   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   |
| HS + HT        | 0   | 1   | 1   | 1   | 0   | 0   | 1   | 0   | 1   | 1   |
| HS + HTMed     | 0   | 1   | 1   | 1   | 0   | 0   | 1   | 0   | 1   | 1   |
| TPBN + HT      | 0   | 1   | 1   | 1   | 0   | 0   | 1   | 0   | 1   | 1   |
| TPBN + HTMed   | 0   | 1   | 1   | 1   | 0   | 0   | 1   | 0   | 1   | 0   |
| LA +HT         | 0   | 1   | 1   | 1   | 0   | 0   | 1   | 1   | 1   | 1   |
| LA +HTMed      | 0   | 1   | 1   | 1   | 0   | 0   | 1   | 1   | 1   | 1   |
variable is the logarithm of prostate-specific antigen (LPSA). The predictors are eight clinical measures: log(cancer volume) (lcavol), log(prostate weight) (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason), and percentage Gleason score 4 or 5 (pgg45). The 97 observations in the dataset are divided into a training and a test part. The training part has 67 observations and the test part has 30 observations. The six methods used in the diabetes example were applied for the training data and mean squared prediction error (MSPE) is calculated based on the test data for each method. The results are presented in Table 9.

In this example, the four variables (lcavol, lweight, lbph, and svi) selected by all the ten methods are the first four variables that enter the linear equations according to lasso. BEN+SN and the two methods involving LA select the most variables (7) while TPBN+HTMed selects the least variables. In terms of prediction performance, except for BEN+SN having a larger error, the methods considered here are comparable. Although polynomial-tailed priors and exponential-tailed priors lead to similar prediction errors, the former tends to select fewer variables, thus producing a simpler model.

### 6 Discussion

In this paper, we consider Bayesian variable selection problem of linear regression model with global-local shrinkage priors on the regression coefficients. Our proposed variable selection procedure selects a variable if the ratio of the posterior mean to the ordinary least square estimate of the

| Table 9: Results for the prostate example. Selected variables are indicated as 1 |
|---------------------------------|--------|--------|--------|--------|--------|--------|----------------|--------|--------|
| RSS+Zcut                      | 1      | 1      | 0      | 1      | 1      | 0      | 0             | gleason| 1      | 0.35   |
| BEN+SN                        | 1      | 1      | 0      | 1      | 1      | 1      | 1             | 1      | 0.47   |
| SS+Prob                       | 1      | 1      | 0      | 1      | 1      | 1      | 0             | 1      | 0.35   |
| SS+Med                        | 1      | 1      | 0      | 1      | 1      | 0      | 0             | 1      | 0.33   |
| HS+HT                         | 1      | 1      | 0      | 1      | 1      | 0      | 1             | 0      | 0.33   |
| HS+HTMed                      | 1      | 1      | 0      | 1      | 1      | 0      | 0             | 0      | 0.33   |
| TPBN+HT                       | 1      | 1      | 0      | 1      | 1      | 0      | 1             | 0      | 0.33   |
| TPBN+HTMed                    | 1      | 1      | 0      | 1      | 1      | 0      | 0             | 0      | 0.33   |
| LA+HT                         | 1      | 1      | 1      | 1      | 1      | 0      | 1             | 1      | 0.33   |
| LA+HTMed                      | 1      | 1      | 1      | 1      | 1      | 0      | 0             | 1      | 0.33   |
corresponding coefficient is greater than 1/2. With orthogonal design matrices, we show that if the local parameters have polynomial-tailed priors, our proposed method is oracle in the sense that it can achieve variable selection consistency and optimal estimation rate at the same time. However, if, instead, an exponential-tailed prior is used for the local parameters, the proposed method does not have the oracle property.

Although the theoretical results are obtained only under the assumption of orthogonal designs, the simulation study shows our method still has competitive performance when the design matrices are moderately correlated.

Because of the use of the ordinary least square estimate in the proposed method, we only consider the situation when the sample size is greater than the number of predictors in the model. If $p > n$, the posterior mean estimator under the global-local shrinkage prior in (2.5) is still valid. In order to work with the half-thresholding method, the ordinary least square estimator should be replaced by a reasonable estimator that does not have adaptive shrinkage effects. A possible choice is the posterior mean estimator under a multivariate normal prior on $\beta$. A similar idea has been used in Wang (2012) in the context of covariance estimation.

In the article, we treat the global parameter $\tau$ as a tuning parameter most of the time in the theoretical results. In the context of normal mean estimation, van der Pas et al. (2017) showed that the polynomial-tailed priors maintain good theoretical results in an empirical Bayes or full Bayes framework. Since $n\tau$ in this article plays a similar role with the global parameter in van der Pas et al. (2017), one may consider to use

$$\hat{\tau}_1 = \frac{1}{n} \max \left\{ \frac{1}{p_n}, \frac{\# \{ i : \sqrt{n} |\hat{\beta}_i| > \sqrt{c_1 \log p_n} \} \} }{c_2 p_n} \right\} \quad (6.1)$$

or

$$\hat{\tau}_2 = \arg\max_{\tau \in [1/(np_n), 1]} \prod_{i=1}^{p_n} \int_{-\infty}^{\infty} \phi(\sqrt{n}(\hat{\beta}_i - \beta)) g_\tau(\beta) d\beta \quad (6.2)$$

as an estimator of $\tau$ in our context, where $c_1, c_2$ are positive constants, $g_\tau(\beta) = \int_0^{\infty} \phi(\frac{\beta}{\sqrt{\tau}}) \frac{1}{\sqrt{\tau}} \pi(\gamma) d\gamma$, and $\phi(\cdot)$ is the probability density function of the standard normal distribution. By examining the proof of Theorem 1 and making use of the fact that $E(1 - s_i | \tau, y)$ is monotonic in $\tau$, we can get that, in an empirical Bayes framework with $\tau$ estimated by $\hat{\tau}$, a sufficient condition for the HT method with polynomial-tailed priors to preserve oracle property is that there are two sequences, $\{\tau_n^{(1)}\}$ and $\{\tau_n^{(2)}\}$, satisfying the poly-a condition such that $p_n P(\hat{\tau} > \tau_n^{(1)}) \to 0$ and $P(\hat{\tau} < \tau_n^{(2)}) \to 0$ as
n \to \infty$. Unfortunately, neither of $\hat{\tau}_1$ and $\hat{\tau}_2$ satisfies the condition since $\hat{\tau}_j \geq 1/(np_n)$, $j = 1, 2$ with probability 1. However, it cannot be concluded that the corresponding HT method with polynomial-tailed priors cannot be oracle because the condition mentioned above is a sufficient condition. More careful examination of the empirical Bayes method in this context is needed. Similar arguments can be made for the full Bayes method.

The proposed method uses posterior means as a summary of the posterior distribution of the regression coefficients and induces sparsity in the estimator by half-thresholding. Recently, there is an increasing trend to use posterior modes for Bayesian inference, especially in high-dimensional problems (Armagan et al., 2013; Ročková and George, 2016). The global-local shrinkage priors used in this article have a spike near zero, so it is likely that sparsity will appear in posterior modes and leads natural variable selection. We believe both the theoretical and empirical performance of this method deserves a detailed study in the future.

Acknowledgements. We would like to thank the editor, the associate editor, and two anonymous referees for their thoughtful comments and helpful suggestions.

References

ARMAGAN, A., CLYDE, M. and DUNSON, D.B. (2011). Generalized beta mixtures of Gaussians, p. 523–531.

ARMAGAN, A., DUNSON, D.B. and LEE, J. (2013). Generalized double Pareto shrinkage. Stat. Sin. 23, 1, 119–143.

BAYARRI, M.J., BERGER, J.O., FORTE, A. and GARCA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. Ann. Statist. 40, 3, 1550–1577. https://doi.org/10.1214/12-AOS1013.

BHADRA, A., DATTA, J., POLSON, N.G. and WILLARD, B. (2017). The horseshoe+ estimator of ultra-sparse signals. Bayesian Anal. https://doi.org/10.1214/16-BA1028, advance publication.

BHATTACHARYYA, A., PATI, D., PILLAI, N.S. and DUNSON, D.B. (2015). Dirichlet–Laplace priors for optimal shrinkage. J. Am. Stat. Assoc. 110, 512, 1479–1490.

BINGHAM, N.H., GOLDIE, C.M. and TEUGELS, J.L. (1987). Regular variation (Encyclopedia of Mathematics and its Applications). Cambridge University Press, Cambridge.

CARVALHO, C.M., POLSON, N.G. and SCOTT, J.G. (2010). The horseshoe estimator for sparse signals. Biometrika 97, 2, 465–480.

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. Ann. Stat. 43, 5, 1986–2018.

CLYDE, M., PARMIGIANI, G. and VIDAKOVIC, B. (1998). Multiple shrinkage and subset selection in wavelets. Biometrika 85, 2, 391–401.

DATTA, J. and GHOSH, J.K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. Bayesian Anal. 8, 1, 111–132.

EFRON, B., HASTIE, T., JOHNSTONE, I. and Tibshirani, R. (2004). Least angle regression. Ann. Stat. 32, 2, 407–499.
FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 456, 1348–1360.

GEORGE, E.I. and MCCULLOCH, R.E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**, 423, 881–889.

GEWEKE, J. (1996). Variable selection and model comparison in regression. *Bayesian statistics 5*, 609–620.

GHOSE, P., TANG, X., GHOSHEL, M. and CHAKRABARTI, A. (2016). Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Ann. Stat.* **44**, 2, 611–640.

GHOSE, P. and TANG, X. (2013). Bayesian variable selection in high-dimensional settings. *J. Am. Stat. Assoc.* **108**, 501, 681–686.

GEORGE, E.I., MARUYAMA, Y., and ZHANG, D. (2011). Fully Bayes factors with a generalized g-prior. *Ann. Stat.* **39**, 5, 2740–2765.

ISHWARAN, H. and RAO, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.* **33**, 2, 730–773.

JOHNSON, V.E. and ROSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Am. Stat. Assoc.* **107**, 498, 649–660.

LI, Q. and LIN, N. (2010). The Bayesian elastic net. *Ann. Stat.* **38**, 3, 1511–1541.

MARUYAMA, Y., GEORGE, E.I., and ZHANG, D. (2011). Fully Bayes factors with a generalized g-prior. *Ann. Stat.* **39**, 5, 2740–2765.

MITCHELL, T.J. and BEAUCHAMP, J.J. (1988). Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* **83**, 404, 1023–1032.

NARISETTY, N.N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Stat.* **42**, 2, 789–817.

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Am. Stat. Assoc.* **103**, 482, 681–686.

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A.A. (2017). Adaptive posterior contraction rates for the horseshoe. arXiv:170203698.

POLSON, N.G. and SCOTT, J.G. (2010). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics 9*, 501–538.

POLSON, N.G. and SCOTT, J.G. (2012). On the half–Cauchy prior for a global scale parameter. *Bayesian Anal.* **7**, 4, 887–902.

ROCKOVÁ, V. and GEORGE, E.I. (2016). The spike-and-slab lasso. Journal of the American Statistical Association (just-accepted). https://doi.org/10.1080/01621459.2016.1260469.

SHUSTER, J. (1968). On the inverse Gaussian distribution function. *J. Am. Stat. Assoc.* **63**, 324, 1514–1516.

STAMEY, T.A., YANG, N., HAY, A.R., MCNEAL, J.E., FREIHA, F.S. and REDWINE, E. (1987). Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. *N. Engl. J. Med.* **317**, 15, 909–916.
Bayesian Variable Selection

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288.

TOMASSI, D., MILONE, D. AND NELSON, J.D. (2015). Wavelet shrinkage using adaptive structured sparsity constraints. Signal Process. 106, 73–87.

WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. Bayesian Anal. 7, 4, 867–886.

XU, X. AND GHOSH, M. (2015). Bayesian variable selection and estimation for group lasso. Bayesian Anal. 10, 4, 909–936.

ZOU, H. (2006). The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101, 476, 1418–1429.

Appendix A: Properties of Slowly Varying Functions

Lemma 1. If $L$ is a slowly varying function, then

1. $L^\alpha$ is slowly varying for all $\alpha \in \mathbb{R}$;

2. $\log L(x)/\log x \to 0$, as $x \to \infty$;

3. $x^{-\alpha}L(x) \to 0$ and $x^\alpha L(x) \to \infty$, as $x \to \infty$ for all $\alpha > 0$;

4. for $\alpha < -1$, $-(\alpha + 1)x^{-\alpha - 1}\int_x^\infty t^{\alpha}L(t)dt \to 1$, as $x \to \infty$;

5. there exists $A_0 > 0$ such that for $\alpha > -1$, $(\alpha + 1)x^{-\alpha - 1}\int_{A_0}^x t^\alpha L(t)dt \to 1$, as $x \to \infty$.

Proof. See Propositions 1.3.6, 1.5.8, and 1.5.10 of Bingham et al. (1987).

Appendix B: Technical Proofs

Lemma 2. Suppose $n\tau_n \to 0$, as $n \to \infty$.

1. If $\gamma_i$ has a proper polynomial-tailed prior described in (3.1) with $0 < a < 1$, then there exist $A_0 > 1$ such that

$$E(1 - s_i \mid \tau_n, Y) \leq \frac{A_0(n\tau_n)^a}{a(1 - a)}L\left(\frac{1}{n\tau_n}\right)\exp\left(\frac{n\hat{\beta}_i^2}{2\sigma^2}\right)(1 + o(1))$$

2. If $\gamma_i$ has a proper exponential-tailed prior described in (3.2) and $C = \int_0^\infty \gamma_i \pi(\gamma_i)d\gamma_i < \infty$, then

$$E(1 - s_i \mid \tau_n, Y) \leq Cn\tau_n \exp\left(\frac{n\hat{\beta}_i^2}{2\sigma^2}\right)(1 + o(1)).$$

The $o(1)$ terms in both cases do not depend on $i$. 

PROOF. By Lebesgue dominated convergence Theorem,

\[
E(1 - s_i \mid \tau_n, Y) = \frac{\int_0^\infty \frac{n\tau_n\gamma_i}{1+n\tau_n\gamma_i} (1 + n\tau_n\gamma_i)^{-\frac{1}{2}} \exp \left\{ \frac{n\beta_i^2}{2\sigma^2} \frac{n\tau_n\gamma_i}{1+n\tau_n\gamma_i} \right\} \pi(\gamma_i) \, d\gamma_i}{\int_0^\infty (1 + n\tau_n\gamma_i)^{-\frac{1}{2}} \exp \left\{ \frac{n\beta_i^2}{2\sigma^2} \frac{n\tau_n\gamma_i}{1+n\tau_n\gamma_i} \right\} \pi(\gamma_i) \, d\gamma_i}
\]

\[
\leq \frac{\int_0^\infty (n\tau_n\gamma_i) (1 + n\tau_n\gamma_i)^{-\frac{3}{2}} \pi(\gamma_i) \, d\gamma_i}{\int_0^\infty (1 + n\tau_n\gamma_i)^{-\frac{1}{2}} \pi(\gamma_i) \, d\gamma_i}
\]

\[
= \int_0^\infty (n\tau_n\gamma_i) (1 + n\tau_n\gamma_i)^{-\frac{3}{2}} \pi(\gamma_i) \, d\gamma_i \exp \left\{ \frac{n\beta_i^2}{2\sigma^2} \right\} (1 + o(1)).
\]

We now consider the case that \(\gamma_i\) has a prior from the polynomial class. By property (5) in Lemma 1, there exists \(A_0 \geq 1\) such that \(\int_{A_0}^\infty \frac{t^{-a}L(t)dt}{x^{1-a}L(x)} \to \frac{1}{1-a}\) as \(x \to \infty\). Therefore,

\[
\int_{A_0}^{\frac{A_0}{n\tau_n}} (n\tau_n\gamma_i) (1 + n\tau_n\gamma_i)^{-\frac{3}{2}} \gamma_i^{-a-1}L(\gamma_i)d\gamma_i \leq n\tau_n \int_{A_0}^\infty \gamma_i^{-a}L(\gamma_i)d\gamma_i
\]

\[
= \frac{n\tau_n}{1-a} \left( \frac{A_0}{n\tau_n} \right)^{1-a} L \left( \frac{A_0}{n\tau_n} \right) (1 + o(1)) \leq \frac{A_0}{1-a} (n\tau_n)^a L \left( \frac{1}{n\tau_n} \right) (1 + o(1)).
\]

Also,

\[
\int_0^{\frac{A_0}{n\tau_n}} (n\tau_n\gamma_i) (1 + n\tau_n\gamma_i)^{-\frac{3}{2}} \gamma_i^{-a-1}L(\gamma_i)d\gamma_i \leq A_0n\tau_n \int_0^\infty \gamma_i^{-a-1}L(\gamma_i)d\gamma_i = A_0n\tau_n,
\]

and

\[
\int_{\frac{A_0}{n\tau_n}}^\infty (n\tau_n\gamma_i) (1 + n\tau_n\gamma_i)^{-\frac{3}{2}} \gamma_i^{-a-1}L(\gamma_i)d\gamma_i \leq \int_{\frac{A_0}{n\tau_n}}^\infty \gamma_i^{-a-1}L(\gamma_i)d\gamma_i
\]

\[
= \frac{1}{a} \left( \frac{A_0}{n\tau_n} \right)^{-a} L \left( \frac{1}{n\tau_n} \right) (1 + o(1)) \leq \frac{A_0}{a} (n\tau_n)^a L \left( \frac{1}{n\tau_n} \right) (1 + o(1)).
\]

Hence,

\[
\int_0^\infty (n\tau_n\gamma_i) (1 + n\tau_n\gamma_i)^{-\frac{3}{2}} \gamma_i^{-a-1}L(\gamma_i)d\gamma_i
\]

\[
\leq \frac{A_0(n\tau_n)^a}{a(1-a)} L \left( \frac{1}{n\tau_n} \right) \left[ (n\tau_n)^{1-a}a(1-a) + a(1 + o(1)) + (1-a)(1 + o(1)) \right]
\]

\[
= \frac{A_0(n\tau_n)^a}{a(1-a)} L \left( \frac{1}{n\tau_n} \right) (1 + o(1)).
\]
If $\gamma_i$ has a prior in the exponential-tailed class and $\int_0^\infty \gamma_i \pi(\gamma_i) d\gamma_i < \infty$, then
\[
\int_0^\infty n\tau_n \gamma_i (1 + n\tau_n \gamma_i)^{-3/2} \pi(\gamma_i) d\gamma_i \leq n\tau_n \int_0^\infty \gamma_i \pi(\gamma_i) d\gamma_i = Cn\tau_n.
\]

**Lemma 3.** Suppose $n\tau_n \to 0$ as $n \to \infty$ and $\eta, q$ are arbitrary constants in $(0,1)$.

1. If $\gamma_i$ has a proper polynomial-tailed prior described in (3.1), then
\[
P(s_i > \eta | \tau_n, \mathbf{Y}) \leq \frac{(a + \frac{1}{2})(\eta q)^{-\frac{a}{2}} (1 - \eta q)^a}{(n\tau_n)^a L \left( \frac{1}{n\tau_n} (\frac{1}{\eta q} - 1) \right)} \exp \left\{ -\frac{n\hat{\beta}^2}{2\sigma^2} \eta (1 - q) \right\} (1 + o(1)).
\]

2. If $\gamma_i$ has a proper exponential prior described in (3.2), then for sufficient large $n$ (not depending on $i$),
\[
P(s_i > \eta | \tau_n, \mathbf{Y}) \leq 2b \left( \frac{n\tau_n (1 - \eta q)}{1 - \eta q} \right) \exp \left\{ \frac{2b}{n\tau_n} \left( \frac{1}{\eta q} - 1 \right) \right\} \exp \left\{ -\frac{n\hat{\beta}^2}{2\sigma^2} \eta (1 - q) \right\}.
\]

**Proof.** For any $\eta, q \in (0, 1)$,
\[
P(s_i > \eta | \tau_n, \mathbf{Y}) = P \left( \gamma_i < \frac{1}{n\tau_n} (\frac{1}{\eta} - 1) \bigg| \tau_n, \mathbf{Y} \right)
\]
\[
\leq \frac{\int_0^{\frac{1}{n\tau_n} (\frac{1}{\eta} - 1)} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{1}{2}} \exp \left\{ -\frac{n\hat{\beta}^2}{2\sigma^2} \gamma_i \right\} \pi(\gamma_i) d\gamma_i}{\int_0^{\frac{1}{n\tau_n} (\frac{1}{\eta} - 1)} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{1}{2}} \pi(\gamma_i) d\gamma_i}
\]
\[
\leq \frac{\int_0^{\frac{1}{n\tau_n} (\frac{1}{\eta} - 1)} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{1}{2}} \pi(\gamma_i) d\gamma_i}{\int_0^{\frac{1}{n\tau_n} (\frac{1}{\eta} - 1)} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{1}{2}} \pi(\gamma_i) d\gamma_i} \exp \left\{ -\frac{n\hat{\beta}^2}{2\sigma^2} \eta (1 - q) \right\}.
\]

The numerator of the first factor in (6.3) is bounded by 1. For the denominator (denoted by $D$), we discuss the two types of priors separately.

First consider the case that $\gamma_i$ has a proper polynomial-tailed prior. By property (4) of Lemma 1,
\[
\int_0^{\frac{1}{n\tau_n} (\frac{1}{\eta} - 1)} \gamma_i^{-\frac{a}{2}} L(\gamma_i) d\gamma_i \to a + \frac{1}{2}, \text{ as } n \to \infty.
\]
Hence,

\[
D \geq \frac{(1 - \eta q)}{(n\tau_n)} \frac{L\left(\frac{1}{n\tau_n} \left( \frac{1}{\eta q} - 1 \right) \right)}{(a + \frac{1}{2}) \left( \frac{1}{n\tau_n} \left( \frac{1}{\eta q} - 1 \right) \right)^{a + \frac{1}{2}}} (1 + o(1))
\]

\[
= \frac{(n\tau_n)^a}{a + 1/2} (\eta q)^{a + \frac{1}{2}} (1 - \eta q)^{-a} L\left(\frac{1}{n\tau_n} \left( \frac{1}{\eta q} - 1 \right) \right) (1 + o(1)).
\]

If \(\gamma_i\) has a proper exponential-tailed prior,

\[
D = \int_{\gamma_i}^{\infty} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{1}{2}} \exp \left\{ -b\gamma_i \right\} L(\gamma_i) d\gamma_i
\]

\[
= \int_{\gamma_i}^{\infty} \left( \frac{n\tau_n \gamma_i}{1 + n\tau_n \gamma_i} \right)^{\frac{1}{2}} (n\tau_n)^{-\frac{1}{2}} \gamma_i^{-\frac{1}{2}} \exp \left\{ -b\gamma_i \right\} L(\gamma_i) d\gamma_i
\]

\[
\geq \int_{\gamma_i}^{\infty} \left( 1 - \eta q \right)^{\frac{1}{2}} (n\tau_n)^{-\frac{1}{2}} \gamma_i^{-\frac{1}{2}} \exp \left\{ -b\gamma_i \right\} L(\gamma_i) d\gamma_i
\]

\[
= \int_{\gamma_i}^{\infty} \left( 1 - \eta q \right)^{\frac{1}{2}} (n\tau_n)^{-\frac{1}{2}} \exp \left\{ -2b\gamma_i \right\} \left( \exp \{ b\gamma_i \} \gamma_i^{-1} \right) \left( \gamma_i^{1/2} L(\gamma_i) \right) d\gamma_i
\]

Since \(\exp(b\gamma_i)\gamma_i^{-1} \to \infty\) and \(\gamma_i^{1/2} L(\gamma_i) \to \infty\) as \(\gamma_i \to \infty\), for sufficiently large \(n\),

\[
D \geq \int_{\gamma_i}^{\infty} \left( \frac{1 - \eta q}{n\tau_n} \right)^{\frac{1}{2}} \exp \left\{ -2b\gamma_i \right\} d\gamma_i
\]

\[
= \frac{1}{2b} \left( \frac{1 - \eta q}{n\tau_n} \right)^{\frac{1}{2}} \exp \left\{ - \frac{2b}{n\tau_n} \left( \frac{1}{\eta q} - 1 \right) \right\}.
\]

**PROOF OF PROPOSITION 1.** It is clear that

\[
E \left( 1 - s_i \mid Y \right) = \frac{\int_{\gamma_i}^{\infty} \frac{n\tau_n \gamma_i}{1 + n\tau_n \gamma_i} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{1}{2}} \exp \left\{ \frac{n\beta_i^2}{2\sigma^2} \frac{n\tau_n \gamma_i}{1 + n\tau_n \gamma_i} \right\} \pi(\gamma_i) d\gamma_i}{\int_{\gamma_i}^{\infty} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{1}{2}} \exp \left\{ \frac{n\beta_i^2}{2\sigma^2} \frac{n\tau_n \gamma_i}{1 + n\tau_n \gamma_i} \right\} \pi(\gamma_i) d\gamma_i}
\]

\[
\leq \frac{\int_{\gamma_i}^{\infty} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{3}{2}} \pi(\gamma_i) d\gamma_i}{\int_{\gamma_i}^{\infty} \left( 1 + n\tau_n \gamma_i \right)^{-\frac{3}{2}} \pi(\gamma_i) d\gamma_i} \exp \left\{ \frac{n\beta_i^2}{2\sigma^2} \right\}.
\]

(6.4)

By Lebesgue dominated convergence Theorem, the numerator and denominator in (6.4) converge to 0 and 1, respectively, as \(n \to \infty\). If \(i \notin A\), \(n\beta_i^2 = O_p(1)\). Therefore, \(E(1 - s_i \mid \tau_n, Y) \overset{P}{\to} 0\), as \(n \to \infty\) by Slutsky’s theorem.
For any $0 < \epsilon \leq 1$, $E(s_i | \tau_n, \mathbf{Y}) = \int_{\epsilon/2}^{1} s_i p(s_i | \tau_n, \mathbf{Y}) ds_i + \int_{\epsilon/2}^{1} s_i p(s_i | \tau_n, \mathbf{Y}) ds_i \leq \epsilon/2 + P(s_i > \epsilon/2 | \tau_n, \mathbf{Y})$. Thus, $P(E(s_i | \tau_n, \mathbf{Y}) \geq \epsilon) \leq P(P(s_i > \epsilon/2 | \tau_n, \mathbf{Y}) \geq \epsilon/2)$. If $\gamma_i$ has a polynomial-tailed prior, using the first part of Lemma 3 with $\eta = \epsilon/2$, the above inequality yields

$$P(E(s_i | \tau_n, \mathbf{Y}) \geq \epsilon) \leq P\left( \frac{(a + \frac{1}{2})(\eta q)^{1-a}}{(n\tau)^a L\left(\frac{1}{n\tau} \left(\frac{1}{\eta q} - 1\right)\right)} \exp\left\{ -\frac{n\beta_i^2}{2\sigma^2} \eta(1 - q) \right\} > \epsilon/2 \right) (1 + o(1))$$

$$= P\left( \frac{\beta_i^2}{\eta q} < \frac{2\sigma^2}{\eta q} \left[ \frac{c_1}{n} - \frac{a \log(n\tau_n)}{n} \left\{ 1 + \frac{\log L\left(\frac{1}{n\tau} \left(\frac{1}{\eta q} - 1\right)\right)}{a \log(n\tau_n)} \right\} \right] \right) (1 + o(1)),$$

where $c_1$ is a constant that does not depend on $n$. By property (2) in Lemma 1 and our assumptions, the terms in the bracket converge to zero as $n \to \infty$. Since $\hat{\beta}_i \overset{\mathcal{D}}{\to} \beta_i^0 \neq 0$, we have $P(E(s_i | \tau_n, \mathbf{Y}) \geq \epsilon) \to 0$, as $n \to \infty$.

If $\gamma_i$ has an exponential-tailed prior, by the second part of Lemma 3, the assumption that $n\tau_n \to 0, n^2 \tau_n \to \infty$ as $n \to \infty$ implies that $P(s_i > \eta | \tau_n, \mathbf{Y}) \overset{\mathcal{D}}{\to} 0$ for any $\eta > 0$. Therefore, $P(P(s_i > \epsilon/2 | \tau_n, \mathbf{Y}) \geq \epsilon/2) \to 0$ and hence, $P(E(s_i | \tau_n, \mathbf{Y}) \geq \epsilon) \to 0$, as $n \to \infty$.

**Proof of Theorem 1.** We first prove the variable selection consistency part. It is clear that

$$P(A_n \neq \mathcal{A}) \leq \sum_{i \notin \mathcal{A}} P\left( E(1 - s_i | \tau_n, \mathbf{Y}) \geq \frac{1}{2} \right) + \sum_{i \in \mathcal{A}} P\left( E(1 - s_i | \tau_n, \mathbf{Y}) < \frac{1}{2} \right).$$

Since $p_0 = |\mathcal{A}|$ does not depend on $n$, by Proposition 1, the second term on the right hand side of the above inequality goes to zero as $n \to \infty$. If $i \notin \mathcal{A}$, by Lemma 2 and the fact that $\sqrt{n}\hat{\beta}_i$ has a standard normal distribution,

$$P\left( E(1 - s_i | \tau_n, \mathbf{Y}) > \frac{1}{2} \right) \leq P\left( \exp\left( \frac{n\beta_i^2}{2} \right) \frac{A_0(n\tau)^a}{a(1-a)L\left(\frac{1}{n\tau} \right)} \xi_n > 1/2 \right)$$

$$= 2 \left[1 - \Phi(\sqrt{M_n})\right],$$

where $\xi_n$, not depending on $i$, is a generic term that converges to 1 as $n \to \infty$, and $M_n = 2\log\left(\frac{C}{(n\tau_n)^a L(1/(n\tau_n))^{\xi_n}}\right)$ with $C$ being a generic constant. Noticing that the right-hand side of the above inequality does not depend on $i$, $\sum_{i \notin \mathcal{A}} P\left( E(1 - s_i | \tau_n, \mathbf{Y}) \geq \frac{1}{2} \right) \leq p_0 P\left( E(1 - s_i | \tau_n, \mathbf{Y}) > \frac{1}{2} \right)$. Therefore, the proof of the variable selection consistency part will be complete if we can
show $2p_n \left[ 1 - \Phi(\sqrt{M_n}) \right]$ converges to zero as $n \to \infty$. In fact, by property (3) in Lemma 1, $M_n \to \infty$, so

$$2p_n \left[ 1 - \Phi(\sqrt{M_n}) \right] \leq \frac{2\phi(\sqrt{M_n})}{\sqrt{M_n}} = C p_n (n \tau_n) \epsilon \frac{(n \tau_n)^{a-\epsilon} L(1/(n \tau_n))}{\sqrt{\log(1/(n \tau_n))}} (1 + o(1)).$$

Again by property (3) in Lemma 1, $(n \tau_n)^{a-\epsilon} L(1/(n \tau_n)) \to 0$ as $n \to \infty$. Therefore, $2p_n \left[ 1 - \Phi(\sqrt{M_n}) \right] \to 0$ as $n \to \infty$.

Now, we show the asymptotic normality part. For any $i \in A$, we have

$$\hat{\beta}_i \overset{p}{\to} \beta_i^0 \neq 0, \sqrt{n} \left( \hat{\beta}_i - \beta_i^0 \right) \overset{d}{\to} N(0, \sigma^2) \text{ and}$$

$$\sqrt{n} \left( \hat{\beta}_i^{HT} - \beta_i^0 \right) = \sqrt{n} \left( \hat{\beta}_i - \beta_i^0 \right) - \sqrt{n} E(s_i \mid \tau_n, Y) \hat{\beta}_i - \sqrt{n} \hat{\beta}_i^{PM} I (E(1 - s_i \mid \tau_n, Y) \leq 1/2).$$

Since the third term on the right-hand side converges to zero in probability by Proposition 1, the proof of the asymptotic normality part will be complete if we can show that $\sqrt{n} E(s_i \mid \tau_n, Y)$ converges to zero in probability. In fact, for any $\epsilon > 0$, by similar arguments as in the proof of Proposition 1,

$$P(\sqrt{n} E(s_i \mid \tau_n, Y) \geq \epsilon) \leq P(P(s_i \geq \epsilon/(2\sqrt{n}) \mid \tau_n, Y) > \epsilon/(2\sqrt{n})).$$

In Lemma 3, let $\eta = \eta_n = \epsilon/(2\sqrt{n})$. Then,

$$P \left( P \left( s_i \geq \frac{\epsilon}{2\sqrt{n}} \mid \tau_n, Y \right) > \frac{\epsilon}{2\sqrt{n}} \right) \leq P \left( \frac{(a + \frac{1}{2})(1 - \frac{\epsilon q}{2\sqrt{n}})^a \exp \left( \frac{n \beta_i^2 \epsilon (1 - q)}{4\sigma^2 \sqrt{n}} \right)}{(n \tau_n)^a \left( \frac{\epsilon q}{2\sqrt{n}} \right)^{a+\frac{1}{2}} L \left( \frac{1}{n \tau_n} \left( \frac{2\sqrt{n} \epsilon q}{\epsilon q - 1} \right) \right)} > \frac{\epsilon}{2\sqrt{n}} \right) (1 + o(1))$$

$$= P(\beta_i^2 < c_n)(1 + o(1)),$$

where $c_n = d_2 n^{-1/2} \left\{ \log (d_1 n^{3/4}) + a \log \left( \frac{1}{n \tau_n} \left( \frac{2\sqrt{n} \epsilon q}{\epsilon q - 1} \right) \right) \right\} - \log L \left( \frac{1}{n \tau_n} \left( \frac{2\sqrt{n} \epsilon q - 1}{\epsilon q} \right) \right) \right\} \right\}$. Since $c_n \to 0$ and $\hat{\beta}_i \overset{p}{\to} \beta_i^0 \neq 0$, we have $P(E(s_i \mid \tau_n, Y) \geq \epsilon/\sqrt{n}) \to 0$.

**Proof of Corollary 1.** Since $s_i = (1 + n \tau \gamma_i)^{-1}$ is a decreasing function in $\tau$,

$$E(1 - s_i \mid Y) = \int_{\xi_n}^{\psi_n} E(1 - s_i \mid \tau, Y) \pi_n^*(\tau) d\tau \leq E(1 - s_i \mid \tau = \psi_n, Y).$$
Bayesian Variable Selection

Similarly,

\[ P(s_i < \eta \mid \mathbf{Y}) \leq P(s_i < \eta \mid \tau = \psi_n, \mathbf{Y}), \]

and

\[ P(s_i > \eta \mid \mathbf{Y}) \leq P(s_i > \eta \mid \tau = \xi_n, \mathbf{Y}). \]

The rest of the proof follows the proof of Theorem 1.

**Proof of Theorem 2.** Similar to the proof of the variable selection consistency part of Theorem 1, the proof will be complete if we can show

\[ \sum_{i \in \bar{A}} P(E(1 - s_i \mid \tau_n, \mathbf{Y}) > 1/2) \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty. \]

By Lemma 2,

\[ P(E(1 - s_i \mid \tau_n, \mathbf{Y}) > 1/2) \leq P\left( n\hat{\beta}_i^2 > M'_n \right) = 2\left[ 1 - \Phi(\sqrt{M'_n}) \right], \]

where \( \xi'_n \), not depending on \( i \), is a generic term that converges to 1 as \( n \rightarrow \infty \) and \( M'_n = -2\log(2Cn\tau_n\xi'_n) \). If \( n\tau_n \rightarrow 0 \), then \( M'_n \rightarrow \infty \). Hence,

\[ \sum_{i \in \bar{A}} P(E(1 - s_i \mid \tau_n, \mathbf{Y}) > 1/2) \leq 2p_n \left[ 1 - \Phi(\sqrt{M'_n}) \right] \sim \frac{2p_n\phi(\sqrt{M'_n})}{\sqrt{-\pi \log(2Cn\tau_n\xi'_n)}} \rightarrow 0, \]

if \( (p_n n\tau_n)^2 / \log(n\tau_n) \rightarrow 0 \), as \( n \rightarrow \infty \).

**Proof of Proposition 2.** Notice that

\[
E(1 - s_i \mid \tau_n, \mathbf{Y}) = \frac{\int_0^\infty n\tau_n \gamma_i (1 + n\tau_n \gamma_i)^{-\frac{3}{2}} \exp\left\{ \frac{n\hat{\beta}_i^2}{2\sigma^2} \frac{n\tau_n \gamma_i}{1 + n\tau_n \gamma_i} \right\} \pi(\gamma_i) d\gamma_i}{\int_0^\infty (1 + n\tau_n \gamma_i)^{-\frac{3}{2}} \exp\left\{ \frac{n\hat{\beta}_i^2}{2\sigma^2} \frac{n\tau_n \gamma_i}{1 + n\tau_n \gamma_i} \right\} \pi(\gamma_i) d\gamma_i} \geq \frac{\int_0^\infty \gamma_i \left( \frac{1}{n\tau_n} + \gamma_i \right)^{-\frac{3}{2}} \pi(\gamma_i) d\gamma_i}{\int_0^\infty \left( \frac{1}{n\tau_n} + \gamma_i \right)^{-\frac{3}{2}} \pi(\gamma_i) d\gamma_i} \exp\left\{ -\frac{n\hat{\beta}_i^2}{2\sigma^2} \right\}. \]

Let \( h_n = \frac{\int_0^\infty \gamma_i \left( \frac{1}{n\tau_n} + \gamma_i \right)^{-\frac{3}{2}} \pi(\gamma_i) d\gamma_i}{\int_0^\infty \left( \frac{1}{n\tau_n} + \gamma_i \right)^{-\frac{3}{2}} \pi(\gamma_i) d\gamma_i} \). If \( n\tau_n \rightarrow c \in (0, \infty) \) and \( \int_0^\infty \gamma_i^{-\frac{3}{2}} \pi(\gamma_i) d\gamma_i < \infty \), by applying LDCTh to both the numerator and the denominator of \( h_n \),
we have $h_n$ that converges to some positive constant that depends on $c$ and $\pi(\cdot)$ as $n \to \infty$. Then, for any $i \notin A$,

$$P(A_n = A) \leq P(E(1 - s_i \mid \tau_n, Y) \leq 1/2) \leq P\left(h_n \exp \left\{ -\frac{n \hat{\beta}_i^2}{2\sigma^2} \right\} < \frac{1}{2} \right).$$

Note that $h_n \exp \left\{ -\frac{n \hat{\beta}_i^2}{2\sigma^2} \right\}$ converges in distribution to some distribution $Z$ with support on $(0, 1)$, so

$$P \left(h_n \exp \left\{ -\frac{n \hat{\beta}_i^2}{2\sigma^2} \right\} < \frac{1}{2} \right) \to P \left(Z < \frac{1}{2} \right) < 1.$$

Thus, the HT procedure does not achieve variable selection consistency.

**Proof of Proposition 3.** Similar to the proof of Theorem 1,

$$\sqrt{n} \left( \beta_{i}^{\text{HT}} - \beta_0^0 \right) = \sqrt{n} \left( \hat{\beta}_i - \beta_0^0 \right) - \sqrt{n} E(s_i \mid \tau_n, Y) \hat{\beta}_i - \sqrt{n} \hat{\beta}_i^\text{PM} I(E(1 - s_i \mid \tau_n, Y) \leq 1/2).$$

For $i \in A$, the third term in the right-hand side converges to zero in probability. The first term has a normal distribution with mean 0 and variance $\sigma^2$. The posterior density function of $s_i$ is

$$p(s_i \mid \tau_n, Y) \propto s_i^{-\frac{3}{2}} \exp \left\{ -\frac{n \hat{\beta}_i^2}{2\sigma^2} s_i - \frac{b}{n\tau_n s_i} \right\} L \left( \frac{1}{n\tau_n} \left( \frac{1}{s_i} - 1 \right) \right), 0 \leq s_i \leq 1.$$

It is obvious that $\frac{m}{M} \tilde{S}_{n}^{(i)} \leq E(s_i \mid \tau_n, Y) \leq \frac{M}{m} \tilde{S}_{n}^{(i)}$ almost surely, where

$$\tilde{S}_{n}^{(i)} = \frac{\int_0^1 s_i^{-\frac{1}{2}} \exp \left\{ -\frac{n \hat{\beta}_i^2}{2\sigma^2} s_i - \frac{b}{n\tau_n s_i} \right\} ds_i}{\int_0^1 s_i^{-\frac{3}{2}} \exp \left\{ -\frac{n \hat{\beta}_i^2}{2\sigma^2} s_i - \frac{b}{n\tau_n s_i} \right\} ds_i}.$$

Notice that

$$s_i^{-\frac{3}{2}} \exp \left\{ -\frac{n \hat{\beta}_i^2}{2\sigma^2} s_i - \frac{b}{n\tau_n s_i} \right\} I(0 < s_i < \infty) = \left( \frac{\lambda_D}{2\pi} \right)^{-\frac{1}{2}} \exp \left( -\frac{\lambda_D}{\mu_D} \right) f(s_i; \lambda_D, \mu_D),$$

where $\lambda_D = 2b/n\tau_n$, $\mu = \frac{\sqrt{2b}\sigma}{|\beta_i|} \left( n^2 \tau_n \right)^{-\frac{1}{2}}$ and

$$f(x; \lambda, \mu) = \left( \frac{\lambda}{2\pi x^3} \right)^{\frac{1}{2}} \exp \left\{ -\frac{\lambda(x - \mu)^2}{2\mu^2 x} \right\} I(0 < x < \infty)$$
is the probability density function of an inverse Gaussian (IG) distribution with mean \( \mu \) and shape parameter \( \lambda \). According to Shuster (1968), the cdf of \( \text{IG}(\mu, \lambda) \) can be expressed as

\[
F(x; \lambda, \mu) = \Phi \left( \sqrt{\frac{\lambda}{\mu}} \left( \frac{x}{\mu} - 1 \right) \right) + \exp \left( \frac{2\lambda}{\mu} \right) \Phi \left( -\sqrt{\frac{\lambda}{\mu}} \left( \frac{x}{\mu} + 1 \right) \right).
\]

Thus, the denominator of \( \tilde{S}_n^{(i)} \) can be written as \((\lambda D_{2\pi})^{-\frac{1}{2}} \exp(-\frac{\lambda D}{\mu_D}) F(1; \lambda_D, \mu_D)\). With the transformation \( s_i = 1/t_i \) and similar arguments for the denominator, the numerator of \( \tilde{S}_n^{(i)} \) can be expressed as \((\lambda N_{2\pi})^{-\frac{1}{2}} \exp(-\frac{\lambda N}{\mu_N}) F(1; \lambda_N, \mu_N)\), where \( \lambda_N = n\beta_i^2/\sigma^2 \) and \( \mu_N = \frac{n\sqrt{\tau_n}^2|\beta_i|}{\sqrt{2b_0\sigma}} \). By some simple calculations,

\[
\tilde{S}_n^{(i)} = \frac{\sqrt{2b_0\sigma}}{n\sqrt{\tau_n}^2|\beta_i|} \left\{ \Phi(b_n) - \exp(c_n)\Phi(-d_n) \right\},
\]

where \( b_n = \sqrt{\frac{2b}{n\tau_n}} \left( \frac{n\sqrt{\tau_n}^2|\beta_i|}{\sqrt{2b_0\sigma}} - 1 \right) \), \( c_n = \frac{2\sqrt{2b}\sqrt{\tau_n}^2|\beta_i|}{\sqrt{2b_0\sigma}} \), and \( d_n = \sqrt{\frac{2b}{n\tau_n}} \left( \frac{n\sqrt{\tau_n}^2|\beta_i|}{\sqrt{2b_0\sigma}} + 1 \right) \).

If \( n^2\tau_n \to \infty \) as \( n \to \infty \), \( b_n \xrightarrow{p} +\infty \) and thus, \( \Phi(b_n) \xrightarrow{p} 1 \). Combining this with the fact that \( \Phi(b_n) + \exp(c_n)\Phi(-d_n) \) is in \([0, 1]\) since it is equal to \( F(1; \lambda_D, \mu_D) \), we have \( \exp(c_n)\Phi(-d_n) \xrightarrow{p} 0 \). As a result, \( \sqrt{n}\tilde{S}_n^{(i)} \xrightarrow{p} \infty \).

If \( n^2\tau_n \to c \) in \((0, \infty)\) as \( n \to \infty \), the limit of \( b_n \) can be \(+\infty\), \(-\infty\), or some constant \( r \). We will discuss the three cases separately. If \( b_n \xrightarrow{p} +\infty \), by similar arguments as in the case \( n^2\tau_n \to \infty \), we have \( \sqrt{n}\tilde{S}_n^{(i)} \xrightarrow{p} \infty \). If \( b_n \xrightarrow{p} -\infty \),

\[
\frac{\exp(c_n)\Phi(-d_n)}{\Phi(b_n)} = \frac{\exp(c_n)\phi(d_n)/d_n}{\phi(-b_n)/(-b_n)} = \frac{\sqrt{2b_0\sigma} - n\sqrt{\tau_n}\sqrt{\tau_n}^2|\beta_i|}{\sqrt{2b_0\sigma} + n\sqrt{\tau_n}\sqrt{\tau_n}^2|\beta_i|} \xrightarrow{p} \frac{\sqrt{2b_0\sigma} - \sqrt{c}\beta_0^2}{\sqrt{2b_0\sigma} + \sqrt{c}\beta_0^2}.
\]

Therefore, \( \tilde{S}_n^{(i)} \xrightarrow{p} 1 \) and \( \sqrt{n}\tilde{S}_n^{(i)} \xrightarrow{p} \infty \). If \( b_n \xrightarrow{p} r \), since \( c_n - \frac{1}{2}d_n^2 + \frac{1}{2}\beta_n^2 = 0 \), we have \( c_n - \frac{1}{2}d_n^2 \xrightarrow{p} -\frac{1}{2}r^2 \). Therefore, \( \exp(c_n)\Phi(-d_n) \sim \exp(c_n)\phi(d_n)/d_n \xrightarrow{p} 0 \) and \( \sqrt{n}\tilde{S}_n^{(i)} \xrightarrow{p} \infty \).

If \( n^2\tau_n \to 0 \), \( b_n \xrightarrow{p} -\infty \). By the famous inequality \( \frac{x^2}{1+x^2} \leq \frac{x(1-\Phi(x))}{\phi(x)} \leq 1 \),

\[
\tilde{S}_n^{(i)} \geq \frac{\sqrt{2b_0\sigma}}{n\sqrt{\tau_n}|\beta_i|} \left\{ \phi(b_n) - \exp(c_n)\phi(d_n)/d_n \right\} = 1 \frac{1 + \frac{\sqrt{2b_0\sigma}}{n\sqrt{\tau_n}|\beta_i|} - 2\frac{\sqrt{2b}}{n\tau_n}|\beta_i|}{2\left( \frac{\sqrt{2b}}{n\tau_n}|\beta_i| - 1 \right)^2} \xrightarrow{p} 1.
\]
In all the cases, $\sqrt{n}\tilde{S}_n^{(i)} \xrightarrow{p} \infty$, as $n \to \infty$. Thus, $\sqrt{n}E(s_i | \tau_n, Y) \xrightarrow{p} \infty$ and $\sqrt{n}(\beta_{i}^{PM} - \beta_{i}^{0}) \not\xrightarrow{} N(0, \sigma^2)$.

**Proof of Theorem 3.** Following the proof of Theorem 3, if $n\tau \to 0$ and $n^2\tau_n \to \infty$ as $n \to \infty$, then $n\sqrt{\tau_n}\tilde{S}_n^{(i)} \xrightarrow{p} \sqrt{2b\sigma}/|\beta_{i}^{0}|$. This completes the proof.

**Proof of Theorem 4.** Theorem 4 is a direct result of Propositions 2 and 3.

**Xueying Tang**
**Malay Ghosh**
**Department of Statistics**
**University of Florida**
**Gainesville, FL 32611, USA**
E-mail: xytang@ufl.edu

**Xiaofan Xu**
**Stubhub Inc.**
**Boston, MA 02109, USA**

**Prasenjit Ghosh**
**Department of Statistics**
**Presidency University**
**86/1 College Street, Kolkata 700073, India**

Paper received: 27 February 2017.