Statistical Analysis of Complex Data: Bayesian Model Selection and Functional Data Depth

by

Naveen Naidu Narisetty

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2016

Doctoral Committee:

Professor Xuming He, Co-Chair
Professor Vijayan N. Nair, Co-Chair
Professor Bhramar Mukherjee
Associate Professor XuanLong Nguyen


THE UNIVERSITY OF MICHIGAN
ARTES · SCIENTIA · VERITAS
1817

To

Amma, Naanna, Siva

ACKNOWLEDGEMENTS

I would first like to thank Xuming and Vijay for their guidance, support, and affection during this wonderful journey. I am extremely fortunate to have them as advisers. Xuming is not only an exemplary scholar, teacher, and a caring adviser, but also an inspirational person who always has abundance of positive energy around. Meetings, lunch time discussions, and walks to the arboretum with him will always be very memorable.

Vijay has been a father figure greatly influencing both my academic and personal life. He is a role model for me for his professional and personal achievements. I will greatly miss our chai time chats on anything and everything which were not only fun but also taught me a lot of perspective about research and academic career. Many thanks to Vijay, Corinne, and all the family for being so loving and welcoming.

I am immensely grateful to Long for many insightful discussions, and helpful advice throughout the time, and particularly during the job search process. His passion for research and teaching motivated me. I would like to thank Liza for trusting me with the QR sessions and for everything she does to make students an integral part of the department, Mouli da for his role in bringing me to Michigan and his affection thereafter, Kerby for his teaching that had a transformational effect on me. I cherish wonderful interactions with all the faculty members. Thanks a lot to Shyamala for all her kind support, and for the help with STAT 470 teaching. All the staff members Bebe, Gina, Judy, and Lorie for always fondly welcoming me into the office with big smiles. I am fortunate to have so many friends in the department; special thanks to everyone in our cohort, in our research group, and the participants of QR sessions for all the good times.

I am indebted to Bhramar di for her contribution in both my professional and personal life. I have had a big learning experience from working with her. My Bengali connections have been kept alive with the 'adda's at her place together with great food. I would specially thank Pramita and Sebanti for the wonderful decade long friendship and all the ISI friends for the nice company. Aditya, Ashwini, Atul da, Gautham, Hwa, Juan, and Sandipan have been very dear friends and each of them occupies big parts of my life frame at graduate school.

I have had many great memories as part of AID and am thankful to my friends Aniket, Anurag, Archit, Benil, Bikash, Karthik, Priyanka, Rashmi, Sayantan, Sharan, Suchandan, Supreet, Surbhi, Vikas, and all the volunteers for the wonderful work together and for the nice memories. Many thanks to Nandini and Uday for the lifelong friendship, and to Ria, Ma, and Baba for their ever fresh affection, and for the delicious Bengali food.

Finally, I would like to thank my family for their constant love, support, friendship, and inspiration. I am proud of my sister Siva, for all her achievements. Kudos to Amma and Naanna for not compromising on our education even in the toughest of times, and for overcoming those times. It is their dedication and the importance they give to education that shaped our lives as they are today. I lovingly dedicate this to my parents and my sister!

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Statistical Analysis of Complex Data:
Bayesian Model Selection and Functional Data Depth

by

Naveen Naidu Narisetty

Chairs: Xuming He and Vijayan N. Nair

Big data of the modern era exhibit different types of complex structures. This dissertation addresses two important problems that arise in this context. Consider high-dimensional data where the number of variables is much larger than the sample size. For model selection in a Bayesian framework, a novel approach using sample size dependent spike and slab priors is proposed. It is shown that the corresponding posterior has strong variable selection consistency even when the number of covariates grows nearly exponentially with the sample size, and that the posterior induces shrinkage similar to the shrinkage due to the L0 penalty. A new computational algorithm for posterior computation is proposed, which is much more scalable in memory and in computational efficiency than existing Markov chain Monte Carlo algorithms. For the analysis of functional data, a new notion of data depth is devised which possesses desirable properties, and is especially well suited for obtaining central regions. In particular, the central regions achieve desired simultaneous coverage probability and are useful in a wide range of applications including boxplots and outlier detection for functional data, and simultaneous confidence bands in regression problems.

CHAPTER I

Introduction

The rapid developments in collecting, storing, transmitting, and managing massive amounts of data have led to unique opportunities and challenges in Statistics and the emerging field of Data Science. This thesis deals with statistical models, methods, theory, and algorithms for analyzing complex data structures including high dimensional data and functional data.

Variable selection is a fundamentally important problem in high-dimensional settings as the number of variables being considered could be even much larger than the sample size, which, for example, is a common feature of modern gene expression data sets. In Chapter II, a novel variable selection approach called "Bayesian Variable Selection with Shrinking and Diffusing priors" (BASAD) is investigated, which uses spike and slab priors with a distinct feature: the prior parameters depend on the sample size and the number of variables. It is showed that the shrinkage due to BASAD is similar to the shrinkage due to L_0 penalty, and that BASAD possesses strong variable selection consistency even when the number of covariates grows nearly exponentially with sample size. This filled a theoretical gap by providing the first formal justification for using spike and slab priors for high dimensional Bayesian variable selection.

Efficient computation is a crucial component of any statistical procedure in the

Big Data era. In Chapter III, a fast and scalable algorithm called Skinny Gibbs is developed, which only requires linear order computations in the number of variables. In contrast with the standard Gibbs sampling algorithm, Skinny Gibbs does not require large matrix operations and is much more scalable to high-dimensional problems both in memory and in computational efficiency while retaining all the strong theoretical properties previously shown for BASAD.

Chapter IV concerns with analyzing functional data using the data depth concept. A given notion of data depth provides an ordering of observations in terms of their closeness to the center of the data cloud. Different notions of data depth have been creatively used to obtain robust nonparametric statistical methods for analyzing multivariate data. The situation for functional data is more complex due to their infinite dimensionality. A new depth measure for functional data called Extremal Depth (ED) is proposed which extends the notions such as ranks and order statistics to functional data, and can be used for summarizing and analyzing functional data. ED is based on an extreme tail-ordering. ED possesses many desired properties, is particularly well-suited for obtaining central regions of functional data. The performance and usefulness of ED is demonstrated on two applications: (i) to construct functional boxplots and to detect outliers, and (ii) to obtain simultaneous confidence bands in regression problems.

CHAPTER II

Bayesian Variable Selection with Shrinking and Diffusing Priors

2.1 Introduction

We consider the linear regression setup with high dimensional covariates where the number of covariates p can be large relative to the sample size n. When $p > n$, the estimation problem is ill-posed without performing variable selection. A natural assumption to limit the number of parameters in high dimensional settings is that the regression function (i.e., the conditional mean) is sparse in the sense that only a small number of covariates (called active covariates) have non-zero coefficients. We aim to develop a new Bayesian methodology for selecting the active covariates that is asymptotically consistent and computationally convenient. A large number of methods have been proposed for variable selection in the literature from both frequentist and Bayesian viewpoints. Many frequentist methods based on penalization have been proposed following the well-known least absolute shrinkage and selection operator (LASSO, Tibshirani (1996a)). We mention the smoothly clipped absolute deviation (SCAD, Fan and Li (2001)), adaptive LASSO (Zou (2006)), octagonal shrinkage and clustering algorithm for regression (OSCAR, Bondell and Reich (2008)) and the Dantzig selector (Candes and Tao (2007); James et al. (2009)) just to name a few.

Fan and Lv (2010) provided a selective overview of high dimensional variable selection methods. Various authors reported inconsistency of LASSO and its poor performance for variable selection under high dimensional settings; see Zou (2006) and Johnson and Rossell (2012). On the other hand, several penalization based methods were shown to have the oracle property (Fan and Li (2001)) under some restrictions on p. For example, Fan and Peng (2004) and Huang and Xie (2007) showed the oracle property for some nonconcave penalized likelihood methods when $p = O(n^{1/3})$ and $p = o(n)$, respectively. Shen et al. (2012) showed that L_0 penalized likelihood method has the oracle property under exponentially large $p = e^{o(n)}$.

Many Bayesian methods have also been proposed for variable selection including the stochastic search variable selection (George and McCulloch (1993)), empirical Bayes variable selection (George and Foster (2000)), spike and slab selection method (Ishwaran and Rao (2005)), penalized credible regions (Bondell and Reich (2012)), non-local prior method (Johnson and Rossell (2012)), among others. We shall describe the typical framework used for Bayesian variable selection methods before discussing their theoretical properties.

We use the standard notation $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$ to represent the linear regression model. Bayesian variable selection methods usually introduce latent binary variables for each of the covariates to be denoted by $Z = (Z_1, \cdots, Z_p)$. The idea is that each Z_i would indicate whether the i^{th} covariate is active in the model or not. For this reason, the prior distribution on the regression coefficient β_i under $Z_i = 0$ is usually a point mass at zero, but a diffused (non-informative) prior under $Z_i = 1$. The concentrated prior of β_i under $Z_i = 0$ is referred to as the spike prior, and the diffused prior under $Z_i = 1$ is called the slab prior. Further, a prior distribution on the binary random vector Z is assumed, which can be interpreted as a prior distribution on the space of models. A Bayesian variable selection method then selects the model with the highest posterior probability. Various selection procedures with this structure

have been proposed; they essentially differ in the form of the spike and slab priors, or in the form of the prior on the model space.

Mitchell and Beauchamp (1988) considered a uniform distribution for the slab prior. George and McCulloch (1993) used the Gaussian distribution with a zero mean and a small but fixed variance as the spike prior, and another Gaussian distribution with a large variance as the slab prior. This allowed the use of a Gibbs sampler to explore the posterior distribution of Z. However, as we argue in Section 2.3, this prior specification does not guarantee model selection consistency at any fixed prior. Ishwaran and Rao (2005) also used Gaussian spike and slab priors, but with continuous bimodal priors for the variance of β to alleviate the difficulty of choosing specific prior parameters. More recently, Ishwaran and Rao (2011) established the oracle property for the posterior mean as n converges to infinity (but p is fixed) under certain conditions on the prior variances. They noted that in the orthogonal design case, a uniform complexity prior leads to correct complexity recovery (i.e., the expected size of the posterior model size converges to the true model size) under weaker conditions on the prior variances. In another development, Yang and He (2012) used shrinking priors to explore commonality across quantiles in the context of Bayesian quantile regression, but the use of such priors for achieving model selection consistency has not been explored. In this paper, we continue to work with the framework where both the spike and slab priors are Gaussian, but our prior parameters depend explicitly on the sample size through which appropriate shrinkage is achieved. We shall establish model selection consistency properties for general design matrices while allowing p to grow with n at a nearly exponential rate. In particular, the strong selection consistency property we establish is a stronger result for model selection than complexity recovery.

One of the most commonly used priors on the model space is the independent prior given by $P[Z = z] = \prod_{i=1}^{p} w_i^{z_i}(1 - w_i)^{z_i}$, where the marginal probabilities w_i

are usually taken to be the same constant. However, when p is diverging, this implies that the prior probability on models with sizes of order less than p goes to zero, which is against model sparsity. We consider marginal probabilities w_i in the order of p^{-1}, which will impose vanishing prior probability on models of diverging size. Yuan and Lin (2005) used a prior that depends on the Gram matrix to penalize models with unnecessary covariates at the prior level. The vanishing prior probability in our case achieves similar prior penalization.

A common notion of consistency for Bayesian variable selection is defined in terms of pairwise Bayes factors, i.e., the Bayes factor of any under- or over-fitted model with respect to the true model goes to zero. Moreno et al. (2010) proved that intrinsic priors give pairwise consistency when $p = O(n)$, and similar consistency of the Bayesian information criterion (BIC, Schwarz (1978)) when $p = O(n^\alpha), \alpha < 1$. Another notion of consistency for both frequentist and Bayesian methods is that the selected model equals the true model with probability converging to one. We refer to this as selection consistency. Bondell and Reich (2012) proposed a method based on penalized credible regions that is shown to be selection consistent when $\log p = O(n^c), c < 1$. Johnson and Rossell (2012) proposed a stronger consistency for Bayesian methods under which the posterior probability of the true model converges to one, which we shall refer to as strong selection consistency. The authors used non local distributions (distributions with small probability mass close to zero) as slab priors, and proved strong selection consistency when $p < n$. However, apart from the limitation $p < n$, their method involves approximations of the posterior distributions and an application of MCMC methods, which are computationally intensive if at all feasible for modest size problems.

We make the following contributions to variable selection in this article. We introduce shrinking and diffusing priors as spike and slab priors, and establish strong selection consistency of the approach for $p = e^{o(n)}$. This approach is computationally

advantageous because a standard Gibbs sampler can be used to sample from the posterior. In addition, we find that the resultant selection on the model space is closely related to the L_0 penalized likelihood function. The merits of the L_0 penalty for variable selection have been discussed by many authors including Schwarz (1978), Liu and Wu (2007), Dicker et al. (2013), Kim et al. (2012) and Shen et al. (2012).

We now outline the remaining sections of the paper as follows. The first part of Section 2.2 describes the model, conditions on the prior parameters and motivation for these conditions. The later part describes our proposed methodology for variable selection based on the proposed model. Section 2.3 motivates the use of sample size dependent prior parameters by considering orthogonal design matrices, and provides insight into the variable selection mechanism using those priors. Section 2.4 presents our main results on the convergence of the posterior distribution of the latent vector Z, and the strong selection consistency of our model selection methodology. Section 2.5 provides an asymptotic connection between the proposed method and the L_0 penalization. Section 2.6 provides a discussion on the conditions assumed for proving the results of Section 2.4. Some computational aspects of the proposed method are noted in Section 3.2.2. We present simulation studies in Section 2.8 to illustrate how the proposed method compares with some existing methods. Application to a gene expression data set is given in Section 2.9, followed by a conclusion in Section 2.10. Section 2.11 provides proofs of some results not given in the earlier sections.

2.2 The model

From now on, we use p_n to denote the number of covariates to indicate that it grows with n. Consider the $n \times 1$ response vector Y, and the $n \times p_n$ design matrix X corresponding to the p_n covariates of interest. Let β be the regression vector, i.e., the conditional mean of Y given X is given by $X\beta$. We assume that β is sparse in the sense that only a few components of β are non-zero; this sparsity assumption can

be relaxed as in Condition 2.4.3. Our goal is to identify the non-zero coefficients to learn about the active covariates. We describe our working model as follows

$$
\begin{aligned}
Y \mid (X, \beta, \sigma^2) &\sim N(X\beta, \sigma^2 I), \\
\beta_i \mid \left(\sigma^2, Z_i = 0\right) &\sim N(0, \sigma^2 \tau_{0,n}^2), \\
\beta_i \mid \left(\sigma^2, Z_i = 1\right) &\sim N(0, \sigma^2 \tau_{1,n}^2), \\
P(Z_i = 1) = 1 - P(Z_i = 0) &= q_n, \\
\sigma^2 &\sim IG(\alpha_1, \alpha_2),
\end{aligned}
\tag{2.1}
$$

where i runs from 1 to p_n, $q_n, \tau_{0,n}, \tau_{1,n}$ are constants that depend on n, and $IG(\alpha_1, \alpha_2)$ is the Inverse Gamma distribution with shape parameter α_1 and scale parameter α_2.

The intuition behind this set-up is that the covariates with zero or very small coefficients will be identified with zero Z values, and the active covariates will be classified as Z = 1. We use the posterior probabilities of the latent variables Z to identify the active covariates.

Notation: We now introduce the following notation to be used throughout the paper.
Rates: For sequences a_n and b_n, $a_n \sim b_n$ means $\frac{a_n}{b_n} \to c$ for some constant $c > 0$, $a_n \succeq b_n$ (or $b_n \preceq a_n$) means $b_n = O(a_n)$, and $a_n \succ b_n$ (or $b_n \prec a_n$) means $b_n = o(a_n)$.
Convergence: Convergence in probability is denoted by $\xrightarrow{\text{P}}$, and equivalence in distribution is denoted by $\stackrel{d}{=}$.
Models: We use k to index an arbitrary model which is viewed as a $p_n \times 1$ binary vector. The i^{th} entry k_i of k indicates whether the i^{th} covariate is active (1) or not (0). We use X_k as the design matrix corresponding to the model k, and β_k to denote the corresponding regression coefficients. In addition, t is used to represent the true model.

Model operations: We use $|k|$ to represent the size of the model k. For two models k and j, the operations $k \vee j$ and $k \wedge j$ denote entry-wise maximum and minimum, respectively. Similarly, $k^c = \mathbf{1} - k$ is entrywise operation, where $\mathbf{1}$ is the vector of 1's. We also use the notation $k \supset j$ (or $k \geq j$) to denote that the model k includes all the covariates in model j, and $k \not\supset j$ otherwise.

Eigenvalues: We use $\phi_{min}(A)$ and $\phi_{max}(A)$ to denote the minimum and maximum eigenvalues, respectively, and $\phi^{\#}_{min}(A)$ to denote the minimum nonzero eigenvalue (MNEV) of the matrix A. Moreover, we use λ^n_M to be the maximum eigenvalue of the Gram matrix $X'X/n$, and for $\nu > 0$, we define

$$m_n(\nu) = p_n \wedge \tfrac{n}{(2+\nu)\log p_n}, and\ \lambda^n_m(\nu) := \inf_{|k| \leq m_n(\nu)} \phi^{\#}_{min}\left(\tfrac{X'_k X_k}{n}\right).$$

Matrix inequalities: For square matrices A and B of the same order, $A \geq B$ or $(A - B) \geq 0$ means that $(A - B)$ is positive semidefinite.

Residual sum of squares: We define $\tilde{R}_k = Y'(I - X(D_k + X'X)^{-1}X')Y$, where $D_k = Diag\,(k\tau_{1n}^{-2} + (\mathbf{1} - k)\tau_{0n}^{-2})$. $\tilde{R}_k$ approximates the usual residual sum of squares $R^*_k = Y'\,(I - P_k)\,Y$, where P_k is the projection matrix corresponding to the model k.

Generic constants: We use c' and w' to denote generic positive constants that can take different values each time they appear.

2.2.1 Prior parameters

We consider $\tau^2_{0,n} \to 0$ and $\tau^2_{1,n} \to \infty$ as n goes to ∞, where the rates of convergence depend on n and p_n. To be specific, we assume that for some $\nu > 0$, and $\delta > 0$,

$$n\tau^2_{0n}\lambda^n_M = o(1), \text{ and } n\tau^2_{1n}\lambda^n_m(\nu) \sim \left(n \vee p_n^{2+2\delta}\right).$$

As will be seen later, these rates ensure desired model selection consistency for any $\delta > 0$, where larger values of δ will correspond to higher penalization and vice versa.

Note that the variance τ_{0n}^2 depends on the sample size n and the scale of the Gram matrix. Since the prior distribution of a coefficient under $Z = 0$ is mostly concentrated in

$$\left(-\frac{3\sigma}{\sqrt{n\lambda_M^n}}, \frac{3\sigma}{\sqrt{n\lambda_M^n}}\right),$$

one can view this as the shrinking neighborhood around 0 that is being treated as the region of inactive coefficients. The variance τ_{1n}^2 increases to ∞, where the rate depends on p_n. However, when $p_n \prec \sqrt{n}$, τ_{1n}^2 can be of constant order (if $\lambda_m^n(\nu)$ is bounded away from zero).

Now consider the prior probability that a coefficient is nonzero (denoted by q_n). The following calculation gives insight into the choice of q_n. Let K_n be a sequence going to ∞, then

$$P\left(\sum_{i=1}^{p_n} Z_i > K_n\right) \approx 1 - \Phi\left(\frac{K_n - p_n q_n}{\sqrt{p_n q_n (1-q_n)}}\right) \longrightarrow 0,$$

if $p_n q_n$ is bounded. Therefore, we typically choose q_n such that $q_n \sim p_n^{-1}$. This can be viewed as apriori penalization of the models with large size in the sense that the prior probability on models with diverging number of covariates goes to zero. To this respect, if K is an initial upper bound for the size of the model t, by choosing $q_n = c/p_n$ such that $\Phi\left((K-c)/\sqrt{c}\right) \approx 1 - \alpha$, our prior probability on the models with sizes greater than K will be α.

We would like to note that the hierarchical model considered by George and McCulloch (1993) is similar to our model (2.1), but their prior parameters are fixed and therefore do not satisfy our conditions. In Section 2.3, we give an example illustrating model selection inconsistency under fixed prior parameters.

2.2.2 Methodology for variable selection

We use the posterior distribution of the latent variables Z_i to select the active covariates. Note that the sample space of Z, denoted by M, has 2^{p_n} points, each of which corresponds to a model. For this reason, we call M the model space. To find the model with the highest posterior probability is computationally challenging for large p_n. In this paper, we use a simpler alternative, that is, we use the p_n marginal posterior probabilities $P(Z_i = 1|Y, X)$, and select the covariates with the corresponding probability more than a fixed threshold $\underline{p} \in (0, 1)$. A threshold probability of 0.5 is a natural choice for $\underline{p}$. This corresponds to what Barbieri and Berger (2004) call the median probability model. In the orthogonal design case, Barbieri and Berger (2004) showed that the median probability model is an optimal predictive model. The median probability model may not be the same as the maximum a posteriori (MAP) model in general, but the two models are the same with probability converging to one under strong selection consistency.

On the other hand, Dey et al. (2008) argued that the median probability model tends to underfit in finite samples. We also consider an alternative by first ranking the variables based on the marginal posterior probabilities and then using BIC to choose among different model sizes. This option avoids the need to specify a threshold. In either case, it is computationally advantageous to use the marginal posterior probabilities, because we need fewer Gibbs iterations to estimate only p_n of them. The proposed methods based on marginal posteriors achieve model selection consistency because the results in Section 2.4 assure that (i) the posterior probability of the true model converges to 1, and (ii) the marginal posterior based variable selection selects the true model with probability going to 1. We now motivate these results and the necessity of sample size dependent priors in a simple but illustrative case with orthogonal designs.

2.3 Orthogonal design

In this section, we consider the case where the number of covariates $p_n < n$, and assume that the design matrix X is orthogonal, i.e., $X'X = nI$. We also assume σ^2 to be known. Though this may not be a realistic set-up, this simple case provides motivation for the necessity of sample size dependent prior parameters as well as an insight into the mechanism of model selection using these priors. At this moment, we do not impose any assumptions on the prior parameters. **All the probabilities used in the rest of the paper are conditional on** X. Under this simple set-up, the joint posterior of β and Z can be written as:

$$\begin{aligned}
&P(\beta, Z \mid \sigma^2, Y) \\
&\quad \propto \exp\left\{-\tfrac{1}{2\sigma^2}\|Y - X\beta\|_2^2\right\} \textstyle\prod_{i=1}^{p_n} \left((1-q_n)\pi_0(\beta_i)\right)^{1-Z_i} \left(q_n\pi_1(\beta_i)\right)^{Z_i} \\
&\quad \propto \exp\left\{-\tfrac{1}{2\sigma^2}(\beta'X'X\beta - \beta'X'Y)\right\} \textstyle\prod_{i=1}^{p_n} \left((1-q_n)\pi_0(\beta_i)\right)^{1-Z_i} \left(q_n\pi_1(\beta_i)\right)^{Z_i} \\
&\quad \propto \exp\left\{-\tfrac{n}{2\sigma^2}\sum_{i=1}^{p}(\beta_i - \hat{\beta}_i)^2\right\} \textstyle\prod_{i=1}^{p_n} \left((1-q_n)\pi_0(\beta_i)\right)^{1-Z_i} \left(q_n\pi_1(\beta_i)\right)^{Z_i},
\end{aligned}$$

where for $k = 0, 1$, $\pi_k(x) = \boldsymbol{\phi}(x, 0, \sigma^2\tau_{k,n}^2)$ is the probability density function (pdf) of the normal distribution with mean zero and variance $\sigma^2\tau_{k,n}^2$ evaluated at x, and $\hat{\beta}_i$ is the OLS estimator of β_i, i.e., $\hat{\beta}_i = X_i'Y/n$.

The product form of the joint posterior of (Z_i, β_i) implies that (Z_i, β_i) and $\{(Z_j, \beta_j), j \neq i\}$ are independent given data. Hence the marginal posterior of Z_i is given by

$$P(Z_i \mid \sigma^2, Y) \propto \int \exp\left\{-\frac{n}{2\sigma^2}(b - \hat{\beta}_i)^2\right\} \left((1-q_n)\pi_0(b)\right)^{1-Z_i} \left(q_n\pi_1(b)\right)^{Z_i} db.$$

Therefore,

$$P(Z_i = 0 \mid \sigma^2, Y) = \frac{(1-q_n)E_{\hat{\beta}_i}(\pi_0(B))}{(1-q_n)E_{\hat{\beta}_i}(\pi_0(B)) + q_nE_{\hat{\beta}_i}(\pi_1(B))}, \tag{2.2}$$

where $E_{\hat{\beta}_i}$ is the expectation under B following the normal distribution with mean $\hat{\beta}_i$

and variance σ^2/n. These expectations can be calculated explicitly, that is, for $k = 0$ and 1,

$$\begin{aligned} E_{\hat{\beta}_i}(\pi_k(B)) &= \frac{\sqrt{n}}{2\pi\sigma\tau_{k,n}} \int \exp\left\{ -\frac{n}{2\sigma^2}(b-\hat{\beta}_i)^2 - \frac{b^2}{2\tau_{k,n}^2} \right\} db \\ &= \frac{1}{\sqrt{2\pi}a_{k,n}} \exp\left\{ -\frac{\hat{\beta}_i^2}{2a_{k,n}^2} \right\}, \end{aligned}$$

where $a_{k,n} = \sqrt{\sigma^2/n + \tau_{k,n}^2}$.

This simple calculation gives much insight into the role of our priors and the influence of the prior parameters on variable selection, which we explain in some detail below. In the following subsections, we assume that the i^{th} covariate is identified as active if and only if $P(Z_i = 1 \mid \sigma^2, Y) > 0.5$ for simplicity, and similar arguments can be produced for threshold values other than 0.5.

2.3.1 Fixed parameters

Let us first consider the case of fixed parameters $\tau_{0n}^2 = \tau_0^2 < \tau_{1n}^2 = \tau_1^2$ and $q_n = q = 0.5$. We then have for $k = 0, 1$,

$$E_{\hat{\beta}_i}(\pi_k(B)) \xrightarrow{\mathrm{P}} \tfrac{1}{\tau_k} \exp\left\{ -\tfrac{\beta_i^2}{2\tau_k^2} \right\} \text{ as } n \to \infty \text{ for } \beta_i \neq 0. \tag{2.3}$$

Now for $\beta_i = \tau_0 \neq 0$, we have $\exp\{-\beta_i^2/2\tau_0^2\}/\tau_0 > \exp\{-\beta_i^2/2\tau_1^2\}/\tau_1$ for any $\tau_1 \neq \tau_0$. Therefore, the limiting value of $P(Z_i = 1 \mid \sigma^2, Y)$ will be less than 0.5 (with high probability) as $n \to \infty$. This implies that even as $n \to \infty$, we would not be able to identify the active coefficient in this case.

2.3.2 Shrinking $\tau_{0,n}^2$, fixed $\tau_{1,n}^2$ & q_n

Now consider the prior parameters such that $\tau_{1,n}^2$ & q_n are fixed, but $\tau_{0,n}^2$ goes to 0 with n. If $\beta_i = 0$, $\sqrt{n}\hat{\beta}_i$ converges in distribution to the standard normal distribution,

and we have, for $k = 0, 1$,

$$\exp\left\{-\frac{\hat{\beta}_i^2}{2(\sigma^2/n)+2\tau_{k,n}^2}\right\} = O_P(1).$$

In this case, (2.3) will imply that $E_{\hat{\beta}_i}(\pi_1(B)) = O_p(1)$, while $E_{\hat{\beta}_i}(\pi_0(B)) \xrightarrow{\text{P}} \infty$. Therefore, from (2.2), we have $P(Z_i = 0 \mid \sigma^2, Y) \xrightarrow{\text{P}} 1$. For $\beta_i \neq 0$, using $\hat{\beta}_i^2 \xrightarrow{\text{P}} \beta_i^2$ and the fact that $xe^{-rx^2} \to 0$ as $x \to \infty$ (for fixed $r > 0$), we obtain $E_{\hat{\beta}_i}(\pi_0(B)) \to 0$. As $E_{\hat{\beta}_i}(\pi_1(B)) \sim c'$, for some $c' > 0$, we have $P(Z_i = 1 \mid \sigma^2, Y) \xrightarrow{\text{P}} 1$.

To summarize, we have argued that $P(Z_i = 0 \mid \sigma^2, Y) \xrightarrow{\text{P}} I(\beta_i = 0)$, where $I(.)$ is the indicator function. That is, for orthogonal design matrices, the marginal posterior probability of including an active covariate or excluding an inactive covariate converges to one under shrinking prior parameter $\tau_{0,n}^2$, with fixed parameters $\tau_{1,n}^2$ and q_n. However, it should be noted that this statement is restricted to the convergence of marginals of Z, and does not assure consistency of overall model selection. To achieve this, we will need to allow $\tau_{1,n}^2$, q_n to depend on the sample size too.

2.3.3 Shrinking and diffusing priors

Note that the i^{th} covariate is identified as active if and only if

$$\begin{aligned}
&P(Z_i = 1 \mid \sigma^2, Y) > 0.5 \\
&\quad \Leftrightarrow q_n E_{\hat{\beta}_i}(\pi_1(B)) > (1 - q_n) E_{\hat{\beta}_i}(\pi_0(B)) \\
&\quad \Leftrightarrow \hat{\beta}_i^2 \left(a_{0,n}^{-2} - a_{1,n}^{-2}\right) > 2\left(\log(1 - q_n)a_{1,n} - \log q_n a_{0,n}\right) \\
&\quad \Leftrightarrow \hat{\beta}_i^2 > 2\left(\log(1 - q_n)a_{1,n} - \log q_n a_{0,n}\right) / (a_{0,n}^{-2} - a_{1,n}^{-2}) := \varphi_n.
\end{aligned}$$

In particular, when $\tau_{0,n}^2 = o(1/n)$, but the other parameters $\tau_{1,n}^2$ and q_n are fixed, we have $\varphi_n \sim \sigma^2 \log n/n$. Without loss of generality, assume that the first $|t|$ coeffi-

cients of β are non-zero. For $i > |t|$, $\beta_i = 0$ which implies that $n\hat{\beta}_i^2 \stackrel{d}{=} \chi_1^2$. Therefore,

$$\begin{aligned} P[\hat{\beta}_i^2 > \tfrac{\sigma^2 \log n}{n}] &= P[\chi_1^2 > \log n] \\ &\geq (\tfrac{1}{\sqrt{\log n}} - \tfrac{1}{\sqrt{\log n^3}}) e^{-\frac{\log n}{2}} \\ &\geq n^{-1/2-\epsilon}, \end{aligned}$$

for $\epsilon > 0$ and sufficiently large n. Therefore, we have

$$\begin{aligned} P[Z = t|\sigma^2, Y] &\leq P\left[\hat{\beta}_i^2 \leq \tfrac{\sigma^2 \log n}{n}, \forall i > |t|\right] \\ &\leq (1 - n^{-1/2-\epsilon})^{p_n - |t|} \\ &\to 0, \text{ if } p_n > n^{1/2+2\epsilon}. \end{aligned}$$

The above argument shows that having $\tau_{1,n}^2$ and q_n fixed leads to inconsistency of selection if the number of covariates is much greater than $\sqrt{n}$. In this case, the threshold φ_n should be larger to bound the magnitude of all the inactive covariates simultaneously. By using the diffusing prior parameters Section 2.2.1, the threshold will be $(2+\delta)\sigma^2 \log p_n/n$ in place of $\sigma^2 \log n/n$. Model selection consistency with this threshold can be proved using similar arguments in the orthogonal design case. We will defer the rigorous arguments to the next section.

2.4 Main results

In this section we consider our model given by (2.1) and general design matrices. Because the model selection consistency holds easily with $p_n = O(1)$, we assume throughout the paper that $p_n \to \infty$ as $n \to \infty$.

2.4.1 Conditions

We first state the main conditions we use.

Condition 2.4.1 (On dimension p_n)**.** $p_n = e^{nd_n}$ *for some* $d_n \to 0$ *as* $n \to \infty$, *i.e.,* $\log p_n = o(n)$.

Condition 2.4.2 (Prior parameters)**.** $n\tau_{0n}^2 = o(1)$, $n\tau_{1n}^2 \sim \left(n \vee p_n^{2+3\delta}\right)$, *for some* $\delta > 0$, *and* $q_n \sim p_n^{-1}$.

Condition 2.4.3 (On true model)**.** $Y|X \sim N(X_t\beta_t + X_{t^c}\beta_{t^c}, \sigma^2 I)$ *where the size of the true model* $|t|$ *is fixed. The coefficients corresponding to the inactive covariates can be nonzero but satisfy* $b_0 := \|X_{t^c}\beta_{t^c}\|_2 = O(1)$.

For any fixed K, define

$$\Delta_n(K) := \inf_{\{k:|k|<K|t|,k\not\supset t\}} \|(I - P_k)X_t\beta_t\|_2^2,$$

where P_k is the projection matrix onto the column space of X_k.

Condition 2.4.4 (Identifiability)**.** *There is* $K > 1 + 8/\delta$ *such that* $\Delta_n(K) > \gamma_n := 5\sigma^2|t|(1+\delta)\log\left(\sqrt{n} \vee p_n\right)$.

Condition 2.4.5 (Regularity of the Design)**.** *For some* $\nu < \delta$, $\kappa < (K-1)\delta/2$,

$$\lambda_M^n \prec \left((n\tau_{0n}^2)^{-1} \wedge n\tau_{1n}^2\right); \text{ and } \lambda_m^n(\nu) \succeq \left(\tfrac{n \vee p_n^{2+2\delta}}{n\tau_{1n}^2} \vee p_n^{-\kappa}\right).$$

The moderateness of these conditions will be examined in some detail in Section 2.6.

2.4.2 Results for fixed σ^2

We suppress ν and K from the notation of $\lambda_m^n(\nu)$, $m_n(\nu)$ and $\Delta_n(K)$ for stating the results for convenience. In addition, we introduce the following notation. The Bayes factor of model k with respect to the true model t is defined as

$$BF(k,t) := P(Z = k \mid Y, \sigma^2)/P(Z = t \mid Y, \sigma^2).$$

The following lemma gives an upper bound on the Bayes factors.

Lemma 2.4.1. *Under Conditions 3.3.4 & 3.3.2, for any model $k \neq t$ we have*

$$\begin{aligned} BF(k,t) &= \tfrac{Q_k}{Q_t}\, s_n^{|k|-|t|} \exp\left\{-\tfrac{1}{2\sigma^2}(\tilde{R}_k - \tilde{R}_t)\right\} \\ &\leq w' \left(n\tau_{1n}^2 \lambda_m^n (1-\phi_n)\right)^{-\frac{1}{2}(r_k^* - r_t)} (\lambda_m^n)^{-\frac{1}{2}|t\wedge k^c|}\, s_n^{|k|-|t|} \\ &\qquad \times \exp\left\{-\tfrac{1}{2\sigma^2}(\tilde{R}_k - \tilde{R}_t)\right\}, \end{aligned}$$

where $Q_k = |I + XD_k^{-1}X'|^{-1/2}$, $s_n = q_n/(1-q_n) \sim p_n^{-1}$, $w' > 0$ is a constant, $r_k = rank(X_k)$, $r_k^* = r_k \wedge m_n$, $\phi_n = o(1)$, $\tilde{R}_k = Y'(I - X(D_k + X'X)^{-1}X')Y$, and $D_k = Diag\,(k\tau_{1n}^{-2} + (\mathbf{1}-k)\tau_{0n}^{-2})$.

The following arguments give some heuristics for the convergence of pair-wise Bayes factors. Note that $\tilde{R}_k$ is the residual sum of squares from a shrinkage estimator of β, and the term $LR_n := \exp\{-(\tilde{R}_k - \tilde{R}_t)/2\sigma^2\}$ corresponds to the usual likelihood ratio of the two models k and t. Consider a model k that does not include one or more active covariates, then $(\tilde{R}_k - \tilde{R}_t)$ goes to ∞ at the same rate as n, because it is (approximately) the difference in the residual sums of squares of model k and model t. We then have the Bayes factor converging to zero since $LR_n \sim e^{-cn}$ for some $c > 0$, and due to Conditions 3.3.1–3.3.2, $P_n := (n\tau_{1n}^2 \lambda_m^n (1-\phi_n))^{(r_t - r_k^*)/2} (\lambda_m^n)^{-|t\wedge k^c|/2}\, s_n^{|k|-|t|}(1-\phi_n)^{-|t|/2} = o(e^{cn})$. On the other hand, if the model k includes all the active covariates and one or more inactive covariates, we have $|k| > |t|$, but $(\tilde{R}_k - \tilde{R}_t)$ is probabilistically bounded. The Bayes factor in this case also converges to zero because P_n goes to zero. Note that when $r_k > r_t$, larger values of τ_{1n}^2 will imply smaller P_n. That is, the Bayes factors for large sized models go to zero faster for larger values of τ_{1n}^2. A similar observation is made by Ishwaran and Rao (2011). To state our main result, we first consider the posterior distributions of the models Z, assuming the variance parameter σ^2 to be known. We consider the case with the prior on σ^2 in Theorem 2.4.2.

Theorem 2.4.1. *Assume Conditions 3.3.1- 3.3.2. Under Model (2.1), we have*

$P(Z = t \mid Y, \sigma^2) \xrightarrow{\text{P}} 1$ *as* $n \to \infty$*, i.e., the posterior probability of the true model goes to 1 as the sample size increases to* ∞*.*

Remark 1. The statement of Theorem 2.4.1 is equivalent to

$$\frac{1-P(Z=t|Y,\sigma^2)}{P(Z=t|Y,\sigma^2)} = \sum_{k \neq t} BF(k,t) \xrightarrow{\text{P}} 0. \tag{2.4}$$

Remark 2. It is worth noting that for Theorem 2.4.1 to hold, we do not actually need the true σ^2 to be known. Even for a misspecified $\tilde{\sigma}^2 \neq \sigma^2$, $P(Z = t \mid Y, \tilde{\sigma}^2) \xrightarrow{\text{P}} 1$ under the conditions $\Delta_n > \tilde{\sigma}^2 \gamma_n / \sigma^2$ and $2(1+\delta)\tilde{\sigma}^2 > (2+\delta)\sigma^2$. The same proof for Theorem 2.4.1 works.

To see why (2.4) holds, we provide specific rates of convergence of individual Bayes factors summed over subsets of the model space. We divide the set of models (excluding the model t) into the following subsets

1. Unrealistically large models: $M_1 = \{k : r_k > m_n\}$, all the models with dimension (i.e., the rank) greater than m_n.

2. Over-fitted models: $M_2 = \{k : k \supset t, r_k \leq m_n\}$, i.e., the models of dimension smaller than m_n which include all the active covariates plus one or more inactive covariates.

3. Large models: $M_3 = \{k : k \not\supset t, K|t| < r_k \leq m_n\}$, the models which do not include one or more active covariates, and dimension greater than $K|t|$ but smaller than m_n.

4. Under-fitted models: $M_4 = \{k : k \not\supset t, r_k \leq K|t|\}$, the models of moderate dimension which miss an active covariate.

The proof of Theorem 4.1 shows the following results.

Lemma 2.4.2 (Rates of convergence)**.** *For some constants* $c', w' > 0$ *(which may depend on* δ*), we have*

1. *The sum of Bayes factors* $\sum_{k \in M_1} BF(k,t) \preceq \exp\{-w'n\}$, *with probability at least* $1 - 2\exp\{-c'n\}$.

2. *The sum* $\sum_{k \in M_2} BF(k,t) \preceq v_n := \left(p_n^{-\delta/2} \wedge \frac{p_n^{1+\delta/2}}{\sqrt{n}}\right)$, *with probability greater than* $1 - \exp\{-c' \log p_n\}$.

3. *The sum* $\sum_{k \in M_3} BF(k,t) \preceq \nu_n^{(K-1)|t|/2+1}$, *with probability greater than* $1 - \exp\{-c' K|t| \log p_n\}$.

4. *For some* $w'' < 1$, *we have* $\sum_{k \in M_4} BF(k,t) \preceq \exp\{-w'(\Delta_n - w''\gamma_n)\}$, *with probability greater than* $1 - \exp\{-c'\Delta_n\}$.

2.4.3 Results with prior on σ^2

We now consider the case with the Inverse Gamma prior on the variance parameter σ^2. Define the constant w as $w := \delta/8(1+\delta)^2$ in the rest of the section.

Theorem 2.4.2. *Under the same conditions as in Theorem 2.4.1, if we only consider models of dimension at most* $|t| + w\, n/\log p_n$, *we have* $P(Z = t \mid Y) \xrightarrow{\text{P}} 1$ *as* $n \to \infty$.

Remark 3. Note that the dimension of the models that need to be excluded for Theorem 2.4.2 to hold is in the order of $n/\log p_n$. These are unrealistically large models that are uninteresting to us. From now on, we implicitly assume this restriction when a prior distribution is used for σ^2.

The following corollary ensures that the variable selection procedure based on the marginal posterior probabilities finds the right model with probability tending to 1. It is a direct consequence of Theorems 2.4.1 and 2.4.2, but is particularly useful for computations because it ensures that the marginal posterior probabilities can be used for selecting the active covariates.

Corollary 2.4.1. *Under the conditions of Theorem 2.4.2, we have for any* $0 < \underline{p} < 1$, $P\Big[P(Z_i = t_i \mid Y) > \underline{p} \text{ for all } i = 1, \cdots, p_n\Big] \to 1$ *as* $n \to \infty$.

Proof. Let E_i be the event that the marginal posterior probability of i^{th} covariate $P(Z_i = t_i \mid Y) > \underline{p}$. We shall show that $P[\cup_{i=1}^{p_n} E_i^c] \to 0$ as $n \to \infty$. For each $i = 1, \cdots, p_n$, we have

$$\begin{aligned} P(Z_i \neq t_i \mid Y) &= \sum_{k:k_i \neq t_i} P(Z = k \mid Y) \\ &\leq \sum_{k \neq t} P(Z = k \mid Y) \\ &= 1 - P(Z = t|Y). \end{aligned}$$

Then, $P[\cup_{i=1}^{p_n} E_i^c] = P\Big[P(Z_i = t_i \mid Y) \leq \underline{p} \text{ for some } i = 1, \cdots, p_n\Big] \leq P\Big[P(Z = t|Y) \leq \underline{p}\Big] \to 0$, due to Theorem 2.4.2. □

2.5 Connection with penalization methods

Due to Lemma 2.4.1, the maximum aposteriori (MAP) estimate of the model using our Bayesian set-up is equivalent to minimizing the objective function

$$\begin{aligned} B(k) &:= \tilde{R}_k + 2\sigma^2 \left(-(|k| - |t|) \log s_n - \log(Q_k/Q_t)\right) \\ &= \tilde{R}_k + (|k| - |t|)\, \psi_{n,k}, \end{aligned} \tag{2.5}$$

where

$$\psi_{n,k} = 2\sigma^2 \left(-\log s_n - \tfrac{\log(Q_k/Q_t)}{(|k|-|t|)}\right).$$

Lemma 2.4.2 implies that with exponentially small probability, the sum of Bayes factors of the models with dimension greater than m_n goes to zero (exponentially) for the fixed σ case. We therefore focus on all the models with dimension less than m_n in this section. In addition, assume that the maximum and minimum non-zero eigenvalues of models of size $2|t|$ are bounded away from ∞ and 0, respectively. Then,

due to Condition 3.3.2 and the proof of Lemma 2.11.1 (iii), we have

$$c \log(n \vee p_n) \leq -\frac{\log(Q_k/Q_t)}{(r_k - r_t)} \leq C \log(n \vee p_n), \tag{2.6}$$

for some $0 < c \leq C < \infty$.

In particular, if the models with dimension less than m_n are of full rank, i.e., $|k| = r_k$, then due to (2.6), we have

$$2\sigma^2 c' \log(n \vee p_n) \leq \psi_{n,k} \leq 2\sigma^2 C' \log(n \vee p_n), \tag{2.7}$$

where $0 < c' \leq C' < \infty$. As $n\tau_{0n}^2 \lambda_M^n \to 0$, and $n\tau_{1n}^2 \lambda_m^n \to \infty$,

$$\tilde{R}_k \sim Y'(I - X(1/\tau_{1n}^2 + X'X)^{-1}X')Y = \|Y - \hat{Y}_k\|^2 + O(1).$$

Therefore, the MAP estimate can be (asymptotically) described as the model corresponding to minimizing the following objective function.

$$m(\beta) := \|Y - X\beta\|_2^2 + \psi_{n,k} \left(\|\beta\|_0 - |t|\right). \tag{2.8}$$

Due to the bounds (2.7) on $\psi_{n,k}$, any inactive covariate will be penalized in the order of $\log(n \vee p_n)$ irrespective of the size of the coeffecient. This is however not the case with the L_1 penalty or SCAD penalty, which are directly proportional to the magnitude of the coeffecient in some interval around zero.

The commonly used model selection criteria AIC and BIC are special cases of L_0 penalization. The objective functions of AIC and BIC are similar to $m(\beta)$, which have the quotient of penalty equal to 2 and $\log n$ in place of $\psi_{n,k}$. Due to the results in Section 2.4 and the above arguments, selection properties of our proposed method are similar to those of the L_0 penalty. In particular, it attempts to find the model

with the least possible size that could explain the conditional mean of the response variable. A salient feature of our approach is that the L_0-type penalization is implied by the hierarchical model. The tuning parameters are more transparent than those in penalization methods. Another feature to note is that our model allows high (or even perfect) correlations among inactive covariates. This is practically very useful in high dimensional problems because the number of inactive covariates is often large and the singularity of the design matrix is a common occurrence. Also, high correlations between active and inactive covariates is not as harmful to the proposed method as they are to the L_1-type penalties. This point is illustrated in Table 2.4 of our simulation studies in Section 2.8.

2.6 Discussion of the conditions

The purpose of this section is to demonstrate that Conditions 3.3.1–3.3.2 that we use in Section 2.4 are quite mild. Condition 3.3.1 restricts the number of covariates to be no greater than exponential in n, and Condition 3.3.4 provides the shrinking and diffusing rates for the spike and slab priors, respectively. We note that Conditions 2.4.3–3.3.2 allow β to depend on n. For instance, consider $p_n < n$ and the design matrix X with $X'X/n \to D$, where D is a positive definite matrix. Ishwaran and Rao (2005), Zou (2006), Bondell and Reich (2012) and Johnson and Rossell (2012) assumed this condition on the design under which Conditions 2.4.3 & 3.3.3 only require β to be such that

$$\|\beta_{t^c}\|_2^2 = O\left(\tfrac{1}{n}\right) \text{ and } \|\beta_t\|_2^2 > c' \tfrac{\log n}{n},$$

for some $c' > 0$. Condition 3.3.2 is also satisfied in this case, so Conditions 2.4.3–3.3.2 allow a wider class of design matrices.

In general, Condition 3.3.3 is a mild regularity condition that allows us to identify

the true model. It serves to restrict the magnitude of the correlation between active and inactive covariates, and also to bound the signal to noise ratio from below. The following two remarks provide some insight into the role of Condition 3.3.3 in these aspects.

Remark 4. Consider the case where the active coefficients β_t are fixed. We then have some $w' > 0$, such that

$$\begin{aligned}\Delta_n(K) &\geq \|\beta_t\|_2^2 \inf_{\{k:|k|<K|t|,k\not\supset t\}} \phi_{min}\left(X_t'(I-P_k)X_t\right) \\ &\geq w'n \inf_{\{k:|k|<K|t|,k\not\supset t\}} \phi_{min}\left(\frac{X_{k\vee t}'X_{k\vee t}}{n}\right),\end{aligned}$$

where we have used the fact that $\phi_{min}\left(X_{k\vee t}'X_{k\vee t}\right) \leq \phi_{min}\left(X_t'(I-P_k)X_t\right)$. To see this, we just need to consider the cases where $X_{k\vee t}$ is of full rank. Then, it follows from the observation that $\left(X_t'(I-P_k)X_t\right)^{-1}$ is a submatrix of $\left(X_{k\vee t}'X_{k\vee t}\right)^{-1}$. Therefore, Condition 3.3.3 is satisfied if the minimum eigenvalues of the submatrices of $X'X/n$ with size smaller than $(K+1)|t|$ are uniformly larger than $c'\log(n\vee p_n)/n$. In the other end of the spectrum, where the inactive covariates can be perfectly correlated, Condition 3.3.3 could still hold.

Remark 5. If the infimum of $\phi_{min}\left(X_t'(I-P_k)X_t/n\right)$ is uniformly bounded away from zero, then $\Delta_n(K) \geq w'n\|\beta_t\|_2^2$. Then Condition 3.3.3 is satisfied if

$$\left\|\frac{\beta_t}{\sigma}\right\|_2^2 \geq \frac{c'\log(n\vee p_n)}{n}.$$

Condition 3.3.2 provides conditions on the eigenvalues of the Gram matrix in terms of the prior parameters. The condition is weaker than the assumption that the maximum and minimum non-zero eigenvalues of the Gram matrix are bounded away from infinity and zero, respectively. In Condition 3.3.2, $\lambda_M^n \prec (n\tau_{0n}^2)^{-1}$ will be satisfied if τ_{0n}^2 is small enough. However, the assumption on $\lambda_m^n(\nu)$ is non-trivial as it needs to be greater than $p_n^{-\kappa}$. We now show that this requirement is satisfied with

high probability if the design matrix consists of independent sub-Gaussian rows.

Lemma 2.6.1 (MNEV for sub-Gaussian random matrices)**.** *Suppose that the rows of $X_{n\times p_n}$ are independent isotropic sub-Gaussian random vectors in R^{p_n}. Then, there exists a $\nu > 0$ such that, with probability greater than $1 - \exp(-w'n)$,*

$$\inf_{|k|\leq m_n(\nu)} \phi_{min}\left(\frac{X_k' X_k}{n}\right) > 0.$$

A proof of Lemma 2.6.1 is provided in Section 2.11. Lemma 2.6.1 implies that the Gram matrix of a sub-Gaussian design matrix has the minimum eigenvalues of all the $m_n(\nu)$ dimensional submatrices to be uniformly bounded away from zero. This clearly is stronger than Condition 3.3.2, which only requires the minimum non-zero eigenvalues to be uniformly greater than $p_n^{-\kappa}$. In particular, unlike the restricted isometry conditions which control the minimum eigenvalue, Condition 3.3.2 allows the minimum eigenvalue to be exactly zero to allow even perfect correlation among inactive (or active) covariates.

2.7 Computation

The implementation of our proposed method involves using the Gibbs sampler to draw samples from the posterior of Z. The full conditionals are standard distributions due to the use of conjugate priors. The conditional distribution of β is given by,

$$f(\beta \mid Z, \sigma^2, Y) \propto \exp\left\{-\tfrac{1}{2\sigma^2}\|Y - X\beta\|_2^2\right\} \prod_{i=1}^{p_n} \phi\left(\beta_i, 0, \sigma^2\tau_{Z_i,n}^2\right),$$

where $\phi(x, 0, \tau^2)$ is the pdf of the normal distribution with mean zero, and variance τ^2 evaluated at x. This can be rewritten as

$$f(\beta \mid Z = k, \sigma^2, Y) \propto \exp\left\{-\tfrac{1}{2\sigma^2}(\beta' X' X\beta - 2\beta' X' Y)\right\} \exp\left\{-\tfrac{1}{2\sigma^2}\beta' D_k \beta\right\},$$

where $D_k = Diag\,(\tau_{k_i,n}^{-2})$. Hence, the conditional distribution of β is given by $\beta \sim N(m, \sigma^2 V)$, where $V = (X'X + D_k)^{-1}$, and $m = VX'Y$. Furthermore, the conditionl distribution of Z_i is

$$P(Z_i = 1 \mid \beta, \sigma^2) = \frac{q_n \phi(\beta_i, 0, \sigma^2 \tau_{1,n}^2)}{q_n \phi(\beta_i, 0, \sigma^2 \tau_{1,n}^2) + (1 - q_n)\phi(\beta_i, 0, \sigma^2 \tau_{0,n}^2)}.$$

The conditional of σ^2 is the Inverse Gamma distribution $IG(a, b)$ with $a = \alpha_1 + n/2 + p_n/2$, and $b = \alpha_2 + \beta' D_k \beta/2 + (Y - X\beta)'(Y - X\beta)/2$.

The only possible computational difficulty in the Gibbs sampling algorithm is the step of drawing from the conditional distribution of β, which is a high dimensional normal distribution for large values of p_n. However, due to the structure of the covariance matrix $(X'X + D_k)^{-1}$, it can be efficiently sampled using block updating that only requires drawing from smaller dimensional normal distributions. Details of the block updating can be found in Ishwaran and Rao (2005).

2.8 Simulation study

In this section, we study performance of the proposed method in several experimental settings, and compare it with some existing variable selection methods. We will refer to the proposed method as BASAD for BAyesian Shrinking And Diffusing priors.

The proposed BASAD method has three tuning parameters. In all our empirical work, we use

$$\tau_{0n}^2 = \tfrac{\hat{\sigma}^2}{10n},\ \tau_{1n}^2 = \hat{\sigma}^2 \max\left(\tfrac{p_n^{2.1}}{100n}, \log n\right),$$

where $\hat{\sigma}^2$ is the sample variance of Y, and we choose $q_n = P[Z_i = 1]$ such that $P\left[\sum_{i=1}^{p_n} Z_i = 1 > K\right] = 0.1$, for a pre-specified value of K. Our default value is $K = \max(10, \log(n))$, unless otherwise specified in anticipation of a less sparse model. The purpose of using $\hat{\sigma}^2$ is to provide appropriate scaling. If a preliminary model is

available, it is better to use as $\hat{\sigma}^2$ the residual variance from such a model. It is clear that those choices are not optimized for any given problem, but they provide a reasonable assessment on how well BASAD can do. In the simulations, we use 1000 burn-in iterations for the Gibbs sampler followed by 5000 updates for estimating the posterior probabilities. As mentioned in Section 2.2, we consider both the median probability model (denoted by BASAD) and the BIC-based model (denoted by BASAD.BIC) where the threshold for marginal posterior probability is chosen by the BIC.

In this paper, we report our simulation results for six cases under several (n, p) combinations, varied correlations, signal strengths and sparsity levels.

- Case 1: In the first case, we use the set-up of Johnson and Rossell (2012) with $p = n$. Two sample sizes, $n = 100$ and $n = 200$, are considered, and the covariates are generated from the multivariate normal distributions with zero mean and unit variance. The compound symmetric covariance with pairwise covariance of $\rho = 0.25$ is used to represent correlation between covariates. Five covariates are taken active with coefficients $\beta_t = (0.6, 1.2, 1.8, 2.4, 3.0)$. This is a simple setting with moderate correlation between covariates and strong signal strength.

- Case 2: We consider the $p > n$ scenario with $(n, p) = (100, 500)$ and $(n, p) = (200, 1000)$, but the other parameters are same as in Case 1.

  For the next three cases (Cases 3-5), we keep $(n, p) = (100, 500)$ but vary model sparsity, signal strength, and correlation among covariates.

- Case 3: We keep $\rho = 0.25$ and $|t| = 5$ but have low signals $\beta_t = (0.6, 0.6, 0.6, 0.6, 0.6)$.

- Case 4: We consider a block covariance setting where the active covariates have common correlation (ρ_1) equal to 0.25, the inactive covariates have common

correlation (ρ_3) equal to 0.75 and each pair of active and inactive covariate has correlation (ρ_2) 0.50. The other aspects of the model are the same as in Case 1.

- Case 5: We consider a less sparse true model with $|t| = 25$ and β_t is the vector containing 25 equally spaced values between 1 and 3 (inclusive of 1 and 3).

- Case 6: We consider the more classical case of $n > p$ with $(n, p) = (100, 50)$ and $(n, p) = (200, 50)$. Following Bondell and Reich (2012), the covariates are drawn from a normal distribution with the covariance matrix distributed as the Wishart distribution centered at the identity matrix with p degrees of freedom. Three of the 50 covariates are taken to be active with their coefficients drawn from the uniform distribution $U(0, 3)$ to imply a mix of weak and strong signals.

Table 2.1: *Performance of BASAD for Case 1: $n = p$*

$(n, p) = (100, 100); \rho = 0.25; \lvert t \rvert = 5$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.016	0.985	0.866	0.954	0.015	1.092
BASAD.BIC	0.016	0.985	0.066	0.996	0.256	1.203
piMOM	0.012	0.991	0.836	0.982	0.030	1.083
BCR.Joint			0.442	0.940	0.157	1.165
SpikeSlab			0.005	0.216	0.502	1.660
Lasso.BIC			0.010	0.992	0.430	1.195
EN.BIC			0.398	0.982	0.154	1.134
SCAD.BIC			0.356	0.990	0.160	1.157
$(n, p) = (200, 200); \rho = 0.25; \lvert t \rvert = 5$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.002	1.000	0.944	1.000	0.009	1.037
BASAD.BIC	0.002	1.000	0.090	1.000	0.187	1.087
piMOM	0.003	1.000	0.900	1.000	0.018	1.038
BCR.Joint			0.594	0.994	0.102	1.064
SpikeSlab			0.008	0.236	0.501	1.530
Lasso.BIC			0.014	1.000	0.422	1.101
EN.BIC			0.492	1.000	0.113	1.056
SCAD.BIC			0.844	1.000	0.029	1.040

Table 2.2: *Performance of BASAD for Case 2: $p > n$*

$(n, p) = (100, 500); \rho = 0.25; \lvert t \rvert = 5$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.001	0.948	0.730	0.775	0.011	1.130
BASAD.BIC	0.001	0.948	0.190	0.915	0.146	1.168
BCR.Joint			0.070	0.305	0.268	1.592
SpikeSlab			0.000	0.040	0.626	3.351
Lasso.BIC			0.005	0.845	0.466	1.280
EN.BIC			0.135	0.835	0.283	1.223
SCAD.BIC			0.045	0.980	0.328	1.260
$(n, p) = (200, 1000); \rho = 0.25; \lvert t \rvert = 5$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.000	0.986	0.930	0.950	0.000	1.054
BASAD.BIC	0.000	0.986	0.720	0.990	0.046	1.060
BCR.Joint			0.090	0.250	0.176	1.324
SpikeSlab			0.000	0.050	0.574	1.933
Lasso.BIC			0.020	1.000	0.430	1.127
EN.BIC			0.325	1.000	0.177	1.077
SCAD.BIC			0.650	1.000	0.091	1.063

The summary of our results are presented in Tables 2.1- 2.6. In those tables, BASAD denotes the median probability model, BASAD.BIC denotes the model obtained by using the threshold probability chosen by the BIC. Three competing Bayesian model selection methods are: (1) piMOM, the non-local prior method proposed by Johnson and Rossell (2012) but only when $p \leq n$; (2) BCR.Joint, the Bayesian joint credible region method of Bondell and Reich (2012) (using the default priors followed by an application of BIC); (3) SpikeSlab, the generalized elastic net model obtained using the R package spikeslab (Ishwaran et al., 2010) for the spike and slab method of Ishwaran and Rao (2005). Three penalization methods under consideration are: (1) LASSO; (2) Elastic Net (EN); and (3) SCAD, all tuned by the BIC. Our simulation experiment used 500 data sets from each model when $n \geq p$, but used 200 data sets when $p > n$ to aggregate the results.

The columns of the tables show the average marginal posterior probability assigned to inactive covariates and active covariates (pp_0 and pp_1, respectively), proportion of

Table 2.3: *Performance of BASAD for Case 3:* $(n, p) = (100, 500)$

$\rho = 0.25; \lvert t\rvert = 5; \beta_t = (0.6, 0.6, 0.6, 0.6, 0.6)$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.002	0.622	0.185	0.195	0.066	2.319
BASAD.BIC	0.002	0.622	0.160	0.375	0.193	1.521
BCR.Joint			0.030	0.315	0.447	1.501
SpikeSlab			0.000	0.000	0.857	2.466
Lasso.BIC			0.000	0.520	0.561	1.555
EN.BIC			0.040	0.345	0.478	1.552
SCAD.BIC			0.045	0.340	0.464	1.561

Table 2.4: *Performance of BASAD for Case 4:* $(n, p) = (100, 500)$

$\rho_1 = 0.25, \rho_2 = 0.50, \rho_3 = 0.75$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.002	0.908	0.505	0.530	0.012	1.199
BASAD.BIC	0.002	0.908	0.165	0.815	0.179	1.210
BCR.Joint			0.000	0.000	0.515	2.212
SpikeSlab			0.000	0.000	0.995	10.297
Lasso.BIC			0.000	0.015	0.869	8.579
EN.BIC			0.000	0.000	0.898	8.360
SCAD.BIC			0.000	0.000	0.899	8.739

choosing the true model ($Z = t$), proportion of including the true model ($Z \supset t$) and false discovery rate (FDR). The last column (MSPE) gives the average test mean squared prediction error based on n new observations as testing data. From our simulation experiment, we have the following findings.

(i) The Bayesian model selection methods BASAD and piMOM (whenever available) tend to perform better than the other methods in terms of selecting the true model and controlling the false discovery rate in variable selection, and our proposed BASAD stands out in this regard. The penalization methods often have higher probabilities of selecting all the active covariates at the cost of overfitting and false discoveries. In terms of the prediction error however, BASAD does not always outperform its competitors, but remains competitive.

(ii) When the signals are low (Case 3), all the methods under consideration have

Table 2.5: *Performance of BASAD for Case 5:* $(n, p) = (100, 500)$. *In this case, two versions of BASAD are included, where BASAD.K10 uses our default value of* $K = 10$, *and BASAD.K50 uses a less sparse specification of* $K = 50$.

$\rho = 0.25; \lvert t \rvert = 25$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD.K50	0.020	0.988	0.650	0.950	0.036	3.397
BASAD.BIC.K50	0.020	0.988	0.005	0.960	0.283	4.019
BASAD.K10	0.003	0.548	0.405	0.420	0.011	170.862
BASAD.BIC.K10	0.003	0.548	0.035	0.430	0.076	88.881
BCR.Joint			0.000	0.000	0.622	49.299
SpikeSlab			0.000	0.000	0.816	111.911
Lasso.BIC			0.000	0.005	0.685	58.664
EN.BIC			0.000	0.000	0.693	59.058
SCAD.BIC			0.000	0.000	0.666	72.122
$\rho = 0.75; \lvert t \rvert = 25$						
	pp0	pp1	Exact	Include	FDR	MSPE
BASAD.K50	0.048	0.914	0.005	0.355	0.289	6.103
BASAD.BIC.K50	0.048	0.914	0.000	0.445	0.498	6.611
BASAD.K10	0.003	0.298	0.025	0.030	0.018	349.992
BASAD.BIC.K10	0.003	0.298	0.000	0.060	0.087	61.709
BCR.Joint			0.000	0.000	0.772	34.113
SpikeSlab			0.000	0.000	0.899	48.880
Lasso.BIC			0.000	0.000	0.734	24.310
EN.BIC			0.000	0.000	0.754	29.171
SCAD.BIC			0.000	0.000	0.736	27.236

trouble finding the right model, and BASAD.BIC results in lower prediction error than BASAD with 0.5 as the threshold for posterior probabilities.

(iii) In Case 4, there is a moderate level of correlation among inactive covariates and some level of correlation between active and inactive covariates. This is where BASAD outperforms the other methods under consideration because BASAD is similar to the L_0 penalty and is able to accommodate such correlations well. Please refer to our discussion in Sections 2.5 and 2.6.

(iv) When the true model is not so sparse and has $|t| = 25$ active covariates (Case 5), our default choice of $K = 10$ in BASAD did not perform well, which is not surprising. In fact, no other methods under consideration did well in this case,

Table 2.6: *Performance of BASAD for Case 6: $n > p$*

$(n, p) = (100, 50)$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.037	0.899	0.654	0.714	0.026	1.086
BASAD.BIC	0.037	0.899	0.208	0.778	0.267	1.151
piMOM	0.011	0.892	0.656	0.708	0.021	1.066
SpikeSlab			0.064	0.846	0.567	1.226
BCR.Joint			0.336	0.650	0.216	1.124
Lasso.BIC			0.076	0.744	0.397	1.152
EN.BIC			0.378	0.742	0.194	1.110
SCAD.BIC			0.186	0.772	0.284	1.147
$(n, p) = (200, 50)$						
	pp_0	pp_1	$Z = t$	$Z \supset t$	FDR	MSPE
BASAD	0.026	0.926	0.738	0.784	0.017	1.029
BASAD.BIC	0.026	0.926	0.338	0.842	0.193	1.055
piMOM	0.005	0.908	0.694	0.740	0.020	1.036
BCR.Joint			0.484	0.770	0.133	1.045
SpikeSlab			0.038	0.900	0.629	1.121
Lasso.BIC			0.082	0.752	0.378	1.059
EN.BIC			0.428	0.748	0.165	1.039
SCAD.BIC			0.358	0.812	0.193	1.046

highlighting the difficulty of finding a non-sparse model with a limited sample size. On the other hand, there is some promising news. If we anticipate a less sparse model with $K = 50$, the proposed method BASAD improved the performance considerably. Our empirical experience suggests that if we are uncertain about the level of sparsity of our model, we may use a generous choice of K or use BIC to choose between different values of K.

2.9 Real data example

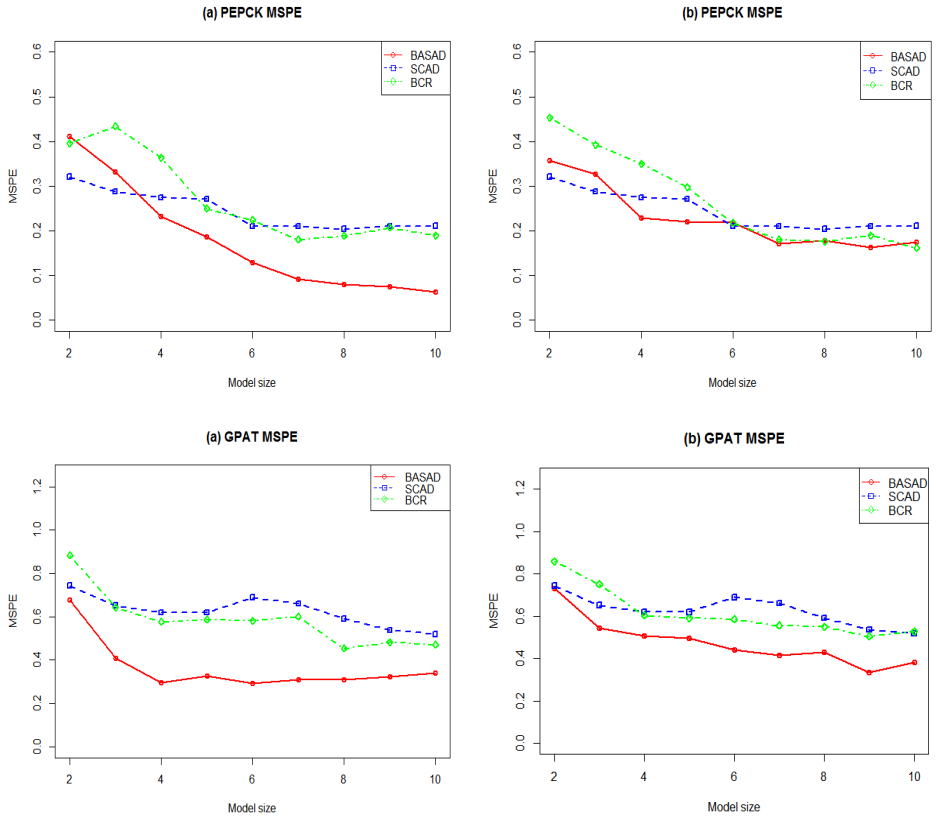
In this section, we apply our variable selection method to a real data set to examine how it works in practice. We consider the data from an experiment conducted by Lan et al. (2006) to study the genetics of two inbred mouse populations (B6 and BTBR). The data include expression levels of 22,575 genes of 31 female and

29 male mice resulting in a total of 60 arrays. Some physiological phenotypes, including the numbers of phosphoenopyruvate carboxykinase (PEPCK) and glycerol-3-phosphate acyltransferase (GPAT) were also measured by quantitative real-time PCR. The gene expression data and the phenotypic data are available at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330). Zhang et al. (2009) used orthogonal components regression to predict each phenotype based on the gene expression data. Bondell and Reich (2012) used the Bayesian credible region method for variable selection on the same data.

Because this is an ultra-high dimensional problem with $p_n = 22,575$, we prefer to perform simple screenings of the genes first based on the magnitude of marginal correlations with the response. The power of marginal screening has been recognized by Fan and Lv (2008). After the screening, the dataset for each of the responses consisted of $p = 200$ and 400 predictors (including the intercept and gender) by taking 198 and 398 genes based on marginal screening. We performed variable selection with BASAD along with LASSO, SCAD and the BCR method. Following Bondell and Reich (2012), we randomly split the sample into a training set of 55 observations and a test set with the remaining five observations. The fitted models using the training set were used to predict the response in the test set. This process was repeated 100 times to estimate the prediction error.

In Figure 2.1, we plot the average mean square prediction error (MSPE) for models of various sizes chosen by BASAD, BCR and SCAD methods for the two responses PEPCK and GPAT. We find that the MSPE of BASAD is mostly smaller than that for other methods across different model sizes. In particular, BASAD chooses less correlated variables and achieves low MSPE with fewer predictive genes than the other methods. We also note that the 10-covariate models chosen by BASAD is very different (with the overlap of just one covariate for PEPCK and three covariates for GPAT) from those of SCAD which chose mostly the same covariates as LASSO. There

Figure 2.1: *Mean squared prediction error (MSPE) versus model size for analyzing PEPCK and GPAT in the upper and lower panel, respectively, (a)* $p = 200$ *and (b)* $p = 400$



are four common covariates identified by both BASAD and BCR methods. When we perform a linear regression by including the covariates chosen by BASAD and SCAD, we noticed that majority of the covariates chosen by BASAD are significant, which indicates that those genes chosen by BASAD are significant in explaining the response even in the presence of those chosen using SCAD. Most of the genes selected by SCAD however are not significant in the presence of those chosen by BASAD. Despite the evidence in favor of the genes selected by BASAD in this example, we must add that the ultimate assessment of a chosen model would need to be made by additional information from the subject matter science and/or additional experiment.

2.10 Conclusion

In this paper, We consider a Bayesian variable selection method for high dimensional data based on the spike and slab priors with shrinking and diffusing priors. We show under mild conditions that this approach achieves strong selection consistency in the sense that the posterior probability of the true model converges to one. The tuning parameters needed for the prior specifications are transparent, and a standard Gibbs sampler can be used for posterior sampling. We also provide the asymptotic relationship between the proposed approach and the L_0 penalty for model selection. Simulation studies in Section 2.8 and real data example in Section 2.9 show evidence that the method performs well in a variety of settings even though we do not attempt to optimize the tuning parameters in the proposed method.

The strong selection consistency of Bayesian methods has not been established in the cases of $p > n$ until very recently. For higher dimensional cases, we just became aware of Liang et al. (2013), which provided the strong selection consistency for Bayesian subset selection based on the theory developed by Jiang (2007) for posterior density consistency. However, to translate density consistency into selection consistency, Liang et al. (2013) imposed a condition on the posterior distribution itself, which is not verifiable directly. The techniques we use in this paper might also be used to complete the development of their theory on strong selection consistency.

Throughout the paper, we assume Gaussian errors in the regression model, but this assumption is not necessary to obtain selection consistency. For proving Lemma 2.4.1, we did not need assumptions on the error distribution, and to prove Theorem 2.4.2, we just need deviation inequalities of the quadratic forms $\epsilon' P_k \epsilon$, which follow the chi-squared distribution for normal errors. Similar proofs with an application of deviation inequalities for other error distributions would work. For instance, Hsu et al. (2012) provide deviation inequalities for quadratic forms of sub-Gaussian random variables.

The primary focus of our paper is model selection consistency. The model is

selected by averaging over the latent indicator variables drawn from the posterior distributions. The strengths of different model selection methods need to be evaluated differently if prediction accuracy is the goal. In our empirical work, we have included comparisons of the mean squared prediction errors, and found that our proposed method based on default tuning parameters is highly competitive in terms of prediction. However, improvements are possible, mainly in the cases of low signals, if the parameters are tuned by BIC or cross-validation, or if model-averaging is used instead of the predictions from a single model.

2.11 Proofs

In this section, we prove all the theoretical results.

Proof of Lemma 2.4.1 . The joint posterior of β, σ^2, Z under model (2.1) is given by

$$P(\beta, Z = k, \sigma^2 \mid Y) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\|Y - X\beta\|_2^2 - \beta' D_k \beta - 2\alpha_2\right)\right\} \sigma^{-2\left(\frac{n}{2}+\frac{p_n}{2}+\alpha_1+1\right)} |D_k|^{\frac{1}{2}} s_n^{|k|}, \quad (2.9)$$

where $D_k = Diag\left(k\tau_{1n}^{-2} + (\mathbf{1} - k)\tau_{0n}^{-2}\right)$, $s_n = q_n/(1 - q_n)$, α_1, α_2 are the parameters of IG prior, and $|k|$ is the size of the model k. By a simple rearrangement of terms in the above expression, we obtain

$$P(\beta, Z = k \mid Y, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\left((\beta - \tilde{\beta})'(D_k + X'X)(\beta - \tilde{\beta}) - \tilde{\beta}'(D_k + X'X)\tilde{\beta}\right)\right\} |D_k|^{\frac{1}{2}} s_n^{|k|},$$

where $\tilde{\beta} = (D_k + X'X)^{-1}X'Y$. Note that $\tilde{\beta}$ is a shrinkage estimator of the regression vector β. Shrinkage of $\tilde{\beta}$ depends on D_k, which is the precision matrix of β given

$Z = k$. The components of $\tilde{\beta}_i$ corresponding to $k_i = 0$ are shrunk towards zero while the shrinkage of coefficients corresponding to $k_i = 1$ is negligible (as τ_{1n}^{-2} is small).

$$\begin{aligned} P(Z = k \mid Y, \sigma^2) &\propto Q_k \, s_n^{|k|} \, \exp\left\{-\tfrac{1}{2\sigma^2}\left(Y'Y - \tilde{\beta}'(D_k + X'X)\tilde{\beta}\right)\right\} \\ &= Q_k \, s_n^{|k|} \, \exp\left\{-\tfrac{1}{2\sigma^2}\left(Y'Y - Y'X(D_k + X'X)^{-1}X'Y\right)\right\} \\ &= Q_k \, s_n^{|k|} \, \exp\left\{-\tfrac{1}{2\sigma^2}\tilde{R}_k\right\}, \end{aligned} \tag{2.10}$$

where $Q_k = |D_k + X'X|^{-\frac{1}{2}}|D_k|^{\frac{1}{2}}$. Next, we obtain bounds on Q_k.

Lemma 2.11.1. *Let A be an invertible matrix, and B be any matrix with appropriate dimension. Further, let k and j be any pair of models. Then,*

(i) $|(A + B'B)^{-1}A| = |I + BA^{-1}B'|^{-1}$,

(ii) $\left(I + \tau_{1n}^2 X_k X_k' + \tau_{0n}^2 X_j X_j'\right)^{-1} \geq \left(I + \tau_{1n}^2 X_k X_k'\right)^{-1}(1 - \xi_n)$, *where* $\xi_n = n\tau_{0n}^2 \lambda_M^n = o(1)$, *and*

(iii) $Q_k \leq w' \left(n\tau_{1n}^2 \lambda_m^n (1 - \phi_n)\right)^{-\frac{1}{2}(r_k^* - r_t)} (\lambda_m^n)^{-\frac{1}{2}|t \wedge k^c|} Q_t$, *where* $w' > 0$, $r_k = rank(X_k)$, $r_k^* = r_k \wedge m_n$, *and* $\phi_n = o(1)$.

Proof. (i) We use the Sylvester's determinant theorem, and the multiplicative property of the determinant to obtain

$$\begin{aligned} |(A + B'B)^{-1}A| &= |I + A^{-\frac{1}{2}}B'BA^{-\frac{1}{2}}|^{-1} \\ &= |I + BA^{-1}B'|^{-1}. \end{aligned}$$

(ii) By the Sherman-Morrison-Woodbury (SMW) identity, assuming A, C and $(C^{-1} + DA^{-1}B)$ to be non-singular,

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}, \tag{2.11}$$

we have, for for any vector a,

$$a'(I + \tau_{1n}^2 X_k X_k' + \tau_{0n}^2 X_j X_j')^{-1} a \;\; = \;\; a' G^{-1} a - \tau_{0n}^2 H,$$

where $G = I + \tau_{1n}^2 X_k X_k'$, and $H = a' G^{-1} X_j (I + \tau_{0n}^2 X_j' G^{-1} X_j)^{-1} X_j' G^{-1} a$. Note that

$$\begin{aligned} 0 \leq \tau_{0n}^2 H &\leq \tau_{0n}^2 a' G^{-1} X_j X_j' G^{-1} a \\ &\leq n \tau_{0n}^2 \lambda_M^n a' G^{-1} a, \end{aligned} \tag{2.12}$$

where λ_M^n is the maximum eigenvalue of the Gram matrix $X'X/n$. Therefore,

$$a'(I + \tau_{1n}^2 X_k X_k')^{-1} a (1 - n\tau_{0n}^2 \lambda_M^n) \leq a'(I + \tau_{1n}^2 X_k X_k' + \tau_{0n}^2 X_j X_j')^{-1} a,$$

and hence (ii) is proved.

(iii) From part (i) of the lemma, we have

$$\begin{aligned} Q_k &= |I + X D_k^{-1} X'|^{-\frac{1}{2}} \\ &= |I + \tau_{1n}^2 X_k X_k' + \tau_{0n}^2 X_{k^c} X_{k^c}'|^{-\frac{1}{2}}. \end{aligned} \tag{2.13}$$

Define $A = I + \tau_{1n}^2 X_{k \wedge t} X_{k \wedge t}' + \tau_{0n}^2 X_{k^c \vee t^c} X_{k^c \vee t^c}'$. Then, by (ii) we have

$$(1 - \xi_n)(I + \tau_{1n}^2 X_{k \wedge t} X_{k \wedge t}'^{-1}) \leq A^{-1} \leq (I + \tau_{1n}^2 X_{k \wedge t} X_{k \wedge t}')^{-1}.$$

This, along with Condition 3.3.2 implies

$$\begin{aligned}
\tfrac{Q_k}{Q_{k\wedge t}} &= |I+\tau_{1n}^2 X_k X_k' + \tau_{0n}^2 X_{k^c} X_{k^c}'|^{-\frac{1}{2}} |A|^{\frac{1}{2}} \\
&= |A + (\tau_{1n}^2 - \tau_{0n}^2) X_{k\wedge t^c} X_{k\wedge t^c}'|^{-\frac{1}{2}} |A|^{\frac{1}{2}} \\
&= |I + (\tau_{1n}^2 - \tau_{0n}^2) X_{k\wedge t^c}' A^{-1} X_{k\wedge t^c}|^{-\frac{1}{2}} \\
&\leq |I + (\tau_{1n}^2 - \tau_{0n}^2)(1-\xi_n) X_{k\wedge t^c}' (I+\tau_{1n}^2 X_{k\wedge t} X_{k\wedge t}')^{-1} X_{k\wedge t^c}|^{-\frac{1}{2}} \\
&= |I+\tau_{1n}^2 X_t X_t' + \tau_{1n}^2 (1-\phi_n) X_{k\wedge t^c} X_{k\wedge t^c}'|^{-\frac{1}{2}} |I+\tau_{1n}^2 X_{k\wedge t} X_{k\wedge t}'|^{\frac{1}{2}} \\
&\leq |I+\tau_{1n}^2 (1-\phi_n) X_k X_k'|^{-\frac{1}{2}} |I+\tau_{1n}^2 X_{k\wedge t} X_{k\wedge t}'|^{\frac{1}{2}} \\
&\leq (n\tau_{1n}^2 \lambda_m^n (1-\phi_n))^{-(r_k^* - r_{t\wedge k})/2} (1-\phi_n)^{-|t\wedge k|/2},
\end{aligned}$$

where $(1-\phi_n) = (\tau_{1n}^2 - \tau_{0n}^2)(1-\xi_n)/\tau_{1n}^2 \to 1$. Similarly, let $A = I + \tau_{1n}^2 X_t X_t' + \tau_{0n}^2 X_{t^c} X_{t^c}'$ to obtain

$$\begin{aligned}
\tfrac{Q_{k\wedge t}}{Q_t} &= |A - (\tau_{1n}^2 - \tau_{0n}^2) X_{k\wedge t^c} X_{k\wedge t^c}'|^{-\frac{1}{2}} |A|^{\frac{1}{2}} \\
&\leq |I+\tau_{1n}^2 X_{k\wedge t} X_{k\wedge t}'|^{-\frac{1}{2}} |I+\tau_{1n}^2 X_t X_t'|^{\frac{1}{2}} \\
&\leq |I+\tau_{1n}^2 X_{t\wedge k^c} X_{t\wedge k^c}'|^{\frac{1}{2}} \\
&\leq (n\tau_{1n}^2 c')^{|t\wedge k^c|/2}.
\end{aligned}$$

The above two inequalities give

$$\tfrac{Q_k}{Q_t} \leq w' (n\tau_{1n}^2 \lambda_m^n (1-\phi_n))^{-(r_k^* - r_t)/2} (\lambda_m^n)^{-|t\wedge k^c|/2}.$$

□

Due to (2.10), we have

$$BF(k,t) = \tfrac{Q_k}{Q_t}\, s_n^{|k|-|t|} \exp\left\{-\tfrac{1}{2\sigma^2}(\tilde{R}_k - \tilde{R}_t)\right\} \quad .$$

Therefore, Lemma 2.11.1(iii) implies Lemma 2.4.1. □

2.11.1 Preliminary Results:

To prove Theorem 2.4.1 & Lemma 2.4.2, we first prove some preliminary results in this section. We define

$$R_k = Y' \left(I - X_k (I/\tau_{1n}^2 + X_k' X_k)^{-1}) X_k' \right) Y. \tag{2.14}$$

As the Bayes factors provided by Lemma 2.4.1 involve the quantities $\tilde{R}_k - \tilde{R}_t$, we shall show that $(\tilde{R}_k - \tilde{R}_t)$ approximates $(R_k - R_t)$, which is easier to work with.

Lemma 2.11.2. *Let A be any matrix such that $A \geq I$, and let j and k be models such that $j \subset k$, we then have*

(i) $\tilde{R}_k = Y'(I + \tau_{1n}^2 X_k X_k' + \tau_{0n}^2 X_{k^c} X_{k^c}')^{-1} Y$,

(ii) $(A + \tau_{0n}^2 X_k X_k')^{-1} - (A + \tau_{1n}^2 X_k X_k')^{-1} \geq (A^{-1} - (A + \tau_{1n}^2 X_k X_k')^{-1}) (1 - w_n)$, *and*

(iii) $(R_j - R_k)(1 - w_n)(1 - \xi_n)^2 \leq \tilde{R}_j - \tilde{R}_k \leq (R_j - R_k)(1 - \xi_n)^{-1}$, *where $w_n = o(1)$ uniformly in k, and ξ_n is as defined in Lemma 2.11.1.*

Proof. (i) Due to (2.11), we have

$$\begin{aligned} X(D_k + X'X)^{-1} X' &= X {D_k}^{-1} X' - X {D_k}^{-1} X' (I + X {D_k}^{-1} X')^{-1} X {D_k}^{-1} X' \\ &= X {D_k}^{-1} X' - X {D_k}^{-1} X' (I - (I + X {D_k}^{-1} X')^{-1}) \\ &= X {D_k}^{-1} X' (I + X {D_k}^{-1} X')^{-1} \\ &= I - (I + X {D_k}^{-1} X')^{-1}, \end{aligned}$$

which implies

$$\begin{aligned} \tilde{R}_k &= Y'Y - Y'X(D_k + X'X)^{-1} X'Y \\ &= Y'(I + X {D_k}^{-1} X')^{-1} Y \\ &= Y'(I + \tau_{1n}^2 X_k X_k' + \tau_{0n}^2 X_{k^c} X_{k^c}')^{-1} Y. \end{aligned}$$

(ii)

$$
\begin{aligned}
LHS &:= (A+\tau_{1n}^2 X_k X_k')^{-1} - (A+\tau_{0n}^2 X_k X_k')^{-1} \\
&= A^{-1}X_k(\tau_{0n}^{-2}I + X_k'A^{-1}X_k)^{-1}X_k'A^{-1} - A^{-1}X_k(\tau_{1n}^{-2}I + X_k'A^{-1}X_k)^{-1}X_k'A^{-1} \\
&= A^{-1}X_k U\left((\tau_{0n}^{-2}I + D)^{-1} - (\tau_{1n}^{-2}I + D)^{-1}\right)U'X_k'A^{-1},
\end{aligned}
$$

where UDU' is the eigen decomposition of $X_k'A^{-1}X_k$, and the diagonal entries of D are denoted by d_i, which are bounded by $n\lambda_M^n$. Hence, the i^{th} diagonal entry of $(\tau_{1n}^{-2}I + D)^{-1} - (\tau_{0n}^{-2}I + D)^{-1}$ is given by

$$
\begin{aligned}
\tfrac{1}{\tau_{1n}^{-2}+d_i} - \tfrac{1}{\tau_{0n}^{-2}+d_i} &= \tfrac{\tau_{1n}^2-\tau_{0n}^2}{(1+\tau_{1n}^2 d_i)(1+\tau_{0n}^2 d_i)} \\
&\geq \tfrac{1-\tau_{0n}^2/\tau_{1n}^2}{(\tau_{1n}^{-2}+d_i)(1+n\tau_{0n}^2\lambda_M^n)} \\
&= \tfrac{1-w_n}{\tau_{1n}^{-2}+d_i},
\end{aligned}
$$

where $w_n = 1 - \frac{(1-\tau_{0n}^2/\tau_{1n}^2)}{(1+n\tau_{0n}^2\lambda_M^n)} \to 0$. Therefore,

$$
\begin{aligned}
LHS &\geq A^{-1}X_k U(I/\tau_{1n}^2 + D)^{-1}U'X_k'A^{-1}(1-w_n) \\
&= \left(A^{-1} - (A+\tau_{1n}^2 X_k X_k')^{-1}\right)(1-w_n).
\end{aligned}
$$

(iii) Define $A = (I+\tau_{1n}^2 X_j X_j' + \tau_{0n}^2 X_{k^c} X_{k^c}')$, and $B = (I+\tau_{1n}^2 X_j X_j')$. By Lemma 2.11.1 (ii),

we have $(1-\xi_n)B^{-1} \leq A^{-1}$. On one hand, due to part (i) of the lemma, we have

$$\begin{aligned}
\tilde{R}_j - \tilde{R}_k &= Y'(A+\tau_{0n}^2 X_{k\wedge j^c} X'_{k\wedge j^c})^{-1}Y - Y'(A+\tau_{1n}^2 X_{k\wedge j^c} X'_{k\wedge j^c})^{-1}Y \\
&\leq Y'A^{-1}Y - Y'(A+\tau_{1n}^2 X_{k\wedge j^c} X'_{k\wedge j^c})^{-1}Y \\
&= Y'A^{-1}X_{k\wedge j^c}(\tau_{1n}^{-2}I + X'_{k\wedge j^c}A^{-1}X_{k\wedge j^c})^{-1}X'_{k\wedge j^c}A^{-1}Y \\
&\leq Y'B^{-1}X_{k\wedge j^c}(\tau_{1n}^{-2}I + X'_{k\wedge j^c}B^{-1}(1-\xi_n)X_{k\wedge j^c})^{-1}X'_{k\wedge j^c}B^{-1}Y \\
&= Y'B^{-1}X_{k\wedge j^c}(\tau_{1n}^{-2}(1-\xi_n)^{-1}I + X'_{k\wedge j^c}B^{-1}X_{k\wedge j^c})^{-1}X'_{k\wedge j^c}B^{-1}Y(1-\xi_n)^{-1} \\
&= \left(Y'B^{-1}Y - Y'(B+\tau_{1n}^2(1-\xi_n)X_{k\wedge j^c}X'_{k\wedge j^c})^{-1}Y\right)(1-\xi_n)^{-1} \\
&\leq (R_j - R_k)(1-\xi_n)^{-1}.
\end{aligned}$$

On the other hand, part (ii) of the lemma implies

$$\begin{aligned}
\tilde{R}_j - \tilde{R}_k &\geq \left(Y'A^{-1}Y - Y'(A+\tau_{1n}^2 X_{k\wedge j^c} X'_{k\wedge j^c})^{-1}Y\right)(1-w_n) \\
&= Y'A^{-1}X_{k\wedge j^c}(\tau_{1n}^{-2}I + X'_{k\wedge j^c}A^{-1}X_{k\wedge j^c})^{-1}X'_{k\wedge j^c}A^{-1}Y(1-w_n) \\
&\geq Y'B^{-1}X_{k\wedge j^c}(\tau_{1n}^{-2}I + X'_{k\wedge j^c}B^{-1}X_{k\wedge j^c})^{-1}X'_{k\wedge j^c}B^{-1}Y(1-w_n)(1-\xi_n)^2 \\
&= \left(Y'B^{-1}Y - Y'(B+\tau_{1n}^2(1-\xi_n)X_{k\wedge j^c}X'_{k\wedge j^c})^{-1}Y\right)(1-w_n)(1-\xi_n)^2 \\
&= (R_j - R_k)(1-w_n)(1-\xi_n)^2.
\end{aligned}$$

□

We use R_k^* to denote the residual sum of squares obtained by OLS for model k, i.e., $R_k^* = Y'(I-P_k)Y$, where P_k is the projection matrix corresponding to the model k. The following remark relates $\tilde{R}_k, R_k$ and R_k^*.

Remark 6. For any model k, $\tilde{R}_k \leq R_k$ by definitions, and $R_k^* \leq R_k$ because if $X_k = U_k\Lambda_k V_k'$ is the SVD of X_k, $P_k = U_k U_k'$ whereas $X_k(I/\tau_{1n}^2 + X_k'X_k)^{-1})X_k' = U_k\Lambda_k(\tau_{1n}^{-2}I + \Lambda_k^2)^{-1}\Lambda_k U_k'$.

The following lemma bounds the difference between R_t and R_t^*.

Lemma 2.11.3. *For any sequence $g_n \to \infty$, and $\epsilon > 0$, we have*

(i) $P[R_t - R_t^* > g_n] \leq \exp\{-c' n\tau_{1n}^2 g_n\}$, *and*

(ii) $P\left[\left|\frac{R_t^*}{n\sigma^2} - 1\right| > \epsilon\right] \leq \exp\{-c'n\}$, *for some* $c' > 0$.

Proof. (i) Due to (2.11), we have

$$\begin{aligned} 0 \leq R_t - R_t^* &= Y'X_t\left((X_t'X_t)^{-1} - (I/\tau_{1n}^2 + X_t'X_t)^{-1}\right)X_t'Y \\ &= Y'X_t(X_t'X_t)^{-1}(\tau_{1n}^2 I + (X_t'X_t)^{-1})^{-1}(X_t'X_t)^{-1}X_t'Y \\ &= (n\tau_{1n}^2)^{-1}Y'MY, \end{aligned}$$

where $M = nX_t(X_t'X_t)^{-2}X_t'$ has rank $|t|$ and bounded eigenvalues. Therefore,

$$\begin{aligned} P[R_t - R_t^* > g_n] &\leq P[Y'MY > n\tau_{1n}^2 g_n] \\ &\leq \exp\{-c' n\tau_{1n}^2 g_n\}. \end{aligned}$$

(ii) Note that R_t^*/σ^2 follows the $\chi^2_{n-|t|}$ distribution. By Lemma 1 of Laurent and Massart (2000), we have

$$P\left[\left|\frac{R_t^*}{\sigma^2} - (n-|t|)\right| \geq 2(n-|t|)(\sqrt{x} + 2x)\right] \leq 2\exp(-(n-|t|)x).$$

Therefore, we obtain $P\left[|R_t^*/n\sigma^2 - 1| > \epsilon\right] \leq \exp\{-c'n\}$. □

2.11.2 Proof of Theorem 2.4.1

To prove Theorem 2.4.1, we divide the set of possible incorrect models into four subsets $M_1, \cdots, M_4$ as defined in Section 2.4. We shall prove $\sum_{k \in M_u} BF(k,t) \xrightarrow{\text{P}} 0$ for each $u = 1, 2, 3, 4$.

2.11.2.1 Unrealistically Large models

We first consider the models in M_1, which correspond to all the models containing at least m_n linearly independent covariates. Note that M_1 is empty if $p_n \preceq n/\log n$. Owing to the penalization of such large models, we shall show that the sum of the Bayes factors of such models converges to zero exponentially fast. First note that for any $s > 0$,

$$\begin{aligned} P\left[\cup_{k\in M_1}\left\{\tilde{R}_t - \tilde{R}_k > n(1+2s)\sigma^2\right\}\right] &\leq P\left[\tilde{R}_t > n(1+2s)\sigma^2\right] \\ &\leq P\left[R_t > n(1+2s)\sigma^2\right] \\ &\leq P\left[R_t^* > (1+s)n\sigma^2\right] + P\left[R_t - R_t^* > s\, n\sigma^2\right] \\ &\leq 2\, e^{-c'n}, \end{aligned}$$

uniformly for all k, due to Lemma 2.11.3.

Consider the term $n\tau_{1n}^2\lambda_m^n(1-\phi_n)$. Conditions 3.3.4 & 3.3.2 imply that

$$\left(p_n^{2+2\delta} \vee n\right) \preceq n\tau_{1n}^2\lambda_m^n(1-\phi_n) \preceq \left(p_n^{2+3\delta} \vee n\right). \tag{2.15}$$

Restricting to the high probability event $\{\tilde{R}_t - \tilde{R}_k \leq n(1+2s)\sigma^2\}$, Lemma 2.4.1 and (2.15) give

$$\begin{aligned} \sum_{k\in M_1} BF(k,t) &\preceq \sum_{k\in M_1} p_n^{-(1+\delta)(m_n-|t|)}\, s_n^{(|k|-|t|)}\, (\lambda_m^n)^{-|t|/2}\, e^{n(1+2s)/2} \\ &\preceq \sum_{k\in M_1} e^{-n(1+\delta)/(2+\delta)}\, s_n^{(|k|-|t|)}\, (\lambda_m^n)^{-|t|/2}\, e^{n(1+2s)/2}, \end{aligned}$$

because for $k \in M_1$, $r_k^* = m_n > n/\log(p_n^{2+\nu}) \geq n/\log(p_n^{2+\delta})$. Therefore, due to

Condition 3.3.2, and $s_n \sim p_n^{-1}$, we have

$$\begin{aligned}
\sum_{k \in M_1} BF(k,t) &\preceq e^{-n(1+\delta)/(2+\delta)} e^{n(1+2s)/2}\, p_n^{\kappa|t|} \sum_{k \in M_1} s_n^{(|k|-|t|)} \\
&\preceq e^{-n(1+\delta)/(2+\delta)} e^{n(1+2s)/2}\, p_n^{c'|t|} \sum_{|k|=m_n}^{p_n} \binom{p_n}{|k|} s_n^{|k|} \\
&\preceq e^{-n(1+\delta)/(2+\delta)} e^{n(1+2s)/2}\, p_n^{c'|t|} \, (1+s_n)^{p_n} \\
&\preceq e^{-w'n} \to 0 \; as \; n \to \infty,
\end{aligned}$$

for some $w' > 0$, if s satisfies $1+2s < 2(1+\delta)/(2+\delta)$, i.e., $s < \delta/2(2+\delta)$. Therefore, we have

$$\sum_{k \in M_1} BF(k,t) \xrightarrow{\mathrm{P}} 0. \tag{2.16}$$

2.11.2.2 Over-fitted models

We use deviation inequalities of the chi-squared distribution to simultaneously bound $(\tilde{R}_t - \tilde{R}_k)$ over all the models in M_2. For $k \in M_2$, we have

$$\begin{aligned}
R_t^* - R_k^* &= \| \left(P_k - P_t\right) Y \|_2^2 \\
&\leq \left(\| \left(P_k - P_t\right) X_{-t} \beta_{-t} \|_2 + \| \left(P_k - P_t\right) \epsilon \|_2 \right)^2 \\
&\leq \left(\|X_{-t}\beta_{-t}\|_2 + \sqrt{\epsilon' P_{k \wedge t^c} \epsilon} \right)^2 \\
&= \left(b_0 + \sqrt{\epsilon' P_{k \wedge t^c} \epsilon} \right)^2,
\end{aligned}$$

where $b_0 = \|X_{-t}\beta_{-t}\|_2 = O(1)$ due to Condition 2.4.3. Since $\epsilon' P_{k \wedge t^c} \epsilon$ follows the chi-squared distribution with $r_k - r_t$ degrees of freedom, for any $x > 0$, and for some

$\sqrt{2/3} < w < 1$, we have

$$\begin{aligned}
&P\left[R_t^* - R_k^* > \sigma^2(2+3x)(r_k - r_t)\log p_n\right] \\
&\quad \le P\left[\epsilon' P_{k\wedge t^c}\epsilon > \sigma^2\left(2+3wx\right)\left(r_k - r_t\right)\log p_n\right] \\
&\quad \le P\left[\chi^2_{r_k - r_t} - (r_k - r_t) > (2+3w^2x)(r_k - r_t)\log p_n\right] \\
&\quad \le c' \exp\{-(1+x)(r_k - r_t)\log p_n\} \\
&\quad = c' p_n^{-(1+x)(r_k - r_t)}.
\end{aligned} \tag{2.17}$$

Consider $0 < s \le \delta/8$, a sequence ζ_n such that $\zeta_n = o(1)$, and define the event

$$\begin{aligned}
A(k) \;&:=\; \left\{\tilde{R}_t - \tilde{R}_k > 2\sigma^2(1+4s)(r_k - r_t)(1-\zeta_n)\log p_n\right\} \\
&\subset \; \{R_t - R_k > 2\sigma^2(1+4s)(r_k - r_t)(1-\xi_n)(1-\zeta_n)\log p_n\} \\
&\subset \; \{R_t - R_k > 2\sigma^2(1+2s)(r_k - r_t)\log p_n\},
\end{aligned}$$

by Lemma 2.11.2 (iii). For a fixed dimension $d > r_t$, consider the event $U(d) := \cup_{\{k:r_k=d\}} A(k)$. Since $R_k \ge R_k^*$, we have

$$\begin{aligned}
P\left[\, U(d)\,\right] \;&\le P\left[\cup_{\{k:r_k=d\}}\left\{R_t - R_k^* > 2\sigma^2(1+2s)(r_k - r_t)\log p_n\right\}\right] \\
&\le P\left[\cup_{\{k:r_k=d\}}\left\{R_t^* - R_k^* > \sigma^2(2+3s)(d - r_t)\log p_n\right\}\right] \\
&\quad + P\left[R_t - R_t^* > s\sigma^2(d - r_t)\log p_n\right].
\end{aligned} \tag{2.18}$$

The event $\{R_t^* - R_k^* > \sigma^2(2+3s)(d-r_t)\log p_n\}$ depends only on the projection matrix $P_{k\wedge t^c}$, so the union $\cup_{\{k:r_k=d\}}$ can be written as a smaller set of events indexed by $P_{k\wedge t^c}$. Note that the cardinality of such projections is at most $p_n^{d-r_t}$ because there are at most p_n^m subspaces of rank m, and any projection matrix $P_{k\wedge t^c}$ corresponds to a subspace of rank $(d - r_t)$. Then, (2.17), (2.18) & Lemma 2.11.3 (i) lead to

$$\begin{aligned}
P\left[\, U(d)\,\right] \;&\le\; c' p_n^{-(1+s)(d-r_t)}\, p_n^{(d-r_t)} + \exp\{-c' n \log p_n\} \\
&\le\; 2c'\, p_n^{-s(d-r_t)}.
\end{aligned} \tag{2.19}$$

Next, we consider the union of all such events $U(d)$, that is,

$$\begin{aligned} P\left[\cup_{\{d>r_t\}}U(d)\right] &\leq \textstyle\sum_{\{d>r_t\}} P\left[U(d)\right] \\ &\leq 2c' \textstyle\sum_{d>r_t} p_n^{-s\,(d-r_t)} \\ &\leq 2c' p_n^{-s} \tfrac{1}{1-p_n^{-s}} \\ &= \tfrac{2c'}{p_n^s-1} \to 0 \ \ as \ \ n \to \infty. \end{aligned} \tag{2.20}$$

Restricting our attention to the high probability event $\cap_{\{d>r_t\}}U(d)^c$, due to Lemma 2.4.1 and (2.15), we have

$$\begin{aligned} \sum_{k\in M_2} BF(k,t) &\preceq \sum_{k\in M_2} \left(n\tau_{1n}^2\lambda_m^n(1-\phi_n)\right)^{-(r_k-r_t)}\ s_n^{(|k|-|t|)} \exp\left\{-\tfrac{1}{2\sigma^2}(\tilde{R}_k-\tilde{R}_t)\right\} \\ &\preceq \sum_{k\in M_2} \left(p_n^{1+\delta}\vee\sqrt{n}\right)^{-(r_k-r_t)}\ s_n^{(|k|-|t|)} p_n^{(1+4s)(r_k-r_t)} \\ &\preceq \sum_{k\in M_2} \left(p_n^{\delta-4s}\vee\sqrt{n}p_n^{-1-4s}\right)^{-(r_k-r_t)}\ s_n^{(|k|-|t|)} \\ &\preceq \left(p_n^{-\delta/2}\wedge\tfrac{p_n^{1+\delta/2}}{\sqrt{n}}\right) \sum_{|k|=|t|+1}^{p_n} \binom{p_n}{|k|-|t|}\ s_n^{(|k|-|t|)} \\ &\sim \nu_n \to 0, \ \ as \ n\to\infty, \end{aligned}$$

where $\nu_n = p_n^{-\delta/2}\wedge(p_n^{1+\delta/2}/\sqrt{n}) \to 0$. Note that, $r_k^* = r_k$ as $r_k < m_n$ for $k \in M_2$, and $(1+s_n)^{p_n} \sim 1$, because $s_n \sim p_n^{-1}$. Therefore, we have

$$\sum_{k\in M_2} BF(k,t) \xrightarrow{\text{P}} 0. \tag{2.21}$$

2.11.2.3 Large models

Models in M_3 do not contain one or more active covariates with dimension at least $K|t|$. Similar to the proof in Section 2.11.2.2, we define the event

$$\begin{aligned} B(k) &:= \left\{ \tilde{R}_t - \tilde{R}_k > 2\sigma^2(1+4s)(r_k - r_t)(1-\zeta_n)\log p_n \right\} \\ &\subset \left\{ \tilde{R}_t - \tilde{R}_{k\vee t} > 2\sigma^2(1+4s)(r_k - r_t)(1-\zeta_n)\log p_n \right\} \\ &\subset \left\{ R_t - R_{k\vee t} > 2\sigma^2(1+2s)(r_k - r_t)\log p_n \right\}, \end{aligned}$$

and consider the union of such events $V(d) := \cup_{\{k:r_k=d, k\in M_3\}} B(k)$. Similar to (2.19), for $d > K|t|$, and $s = \delta/8$, we have

$$\begin{aligned} P\left[\, V(d)\,\right] &\leq P\left[\cup_{\{k:r_k=d\}} \left\{ R_t - R_{k\vee t} > 2\sigma^2(1+2s)(r_k - r_t)\log p_n \right\}\right] \\ &\leq c' p_n^{-(1+s)(d-r_t)} p_n^d \\ &\leq c' p_n^{-(1+w')d} p_n^d = c'\, p_n^{-w'd}, \end{aligned}$$

where the inequality $(1+s)(d-r_t) > (1+w')r_k$ holds for some $w' > 0$, because $(d-r_t)/d > (K-1)/K > 1/(1+\delta/8)$, which implies that $(1+s)(d-r_t) > r_k$. Then,

$$P\left[\cup_{\{d>K|t|\}} V(d)\right] \leq p_n^{-w'K|t|} \to 0, \ \ as\ n \to \infty.$$

Restricting our attention to the high probability event $\cap_{\{d>r_t\}} V(d)^c$, with probability at least $1-\exp\{-w'K|t|\log p_n\} \to 1$, we have

$$\begin{aligned}
\sum_{k\in M_3} BF(k,t) &\preceq \sum_{k\in M_3} \left(p_n^{1+\delta} \vee \sqrt{n}\right)^{-(r_k-r_t)} (\lambda_m^n)^{-|t\wedge k^c|/2} s_n^{(|k|-|t|)} p_n^{(1+4s)(r_k-r_t)} \\
&\preceq \sum_{k\in M_3} \left(p_n^{\delta-4s} \vee \sqrt{n} p_n^{-1-4s}\right)^{-(r_k-r_t)} (\lambda_m^n)^{-|t|/2}\, s_n^{(|k|-|t|)} \\
&\preceq \left(p_n^{-\delta/2} \wedge \tfrac{p_n^{1+2s}}{\sqrt{n}}\right)^{(K-1)r_t+1} p_n^{\kappa|t|/2} \sum_{k\in M_3} s_n^{(|k|-|t|)} \\
&\preceq \nu_n^{(K-1)r_t+1}\, p_n^{\delta(K-1)|t|/4} (1+s_n)^{p_n} \\
&\sim \nu_n^{(K-1)|t|/2} \\
&\to 0, \ as\ n\to\infty.
\end{aligned}$$

Hence,

$$\sum_{k\in M_3} BF(k,t) \xrightarrow{\mathrm{P}} 0. \tag{2.22}$$

2.11.2.4 Under-fitted models

We shall first prove that, if $c \in (0,1)$,

$$P\left[\cup_{k\in M_4}\{\tilde{R}_k - \tilde{R}_t < \ \Delta_n(1-c)\}\right] \to 0, \tag{2.23}$$

where Δ_n is as defined in Condition 3.3.3. Because

$$\begin{aligned}
R_k^* - R_{k\vee t}^* &= \|(P_{X_{k\vee t}} - P_{X_k})Y\|_2^2 \\
&= \|(P_{X_{k\vee t}} - P_{X_k})X_t\beta_t + (P_{X_{k\vee t}} - P_{X_k})\epsilon\|_2^2 \\
&\geq \left(\|(P_{X_{k\vee t}} - P_{X_k})X_t\beta_t\|_2 - \|(P_{X_{k\vee t}} - P_{X_k})\epsilon\|_2\right)^2,
\end{aligned}$$

and by Condition 3.3.3, $\|(P_{X_{k\vee t}} - P_{X_k})X_t\beta_t\|_2 = \|(I - P_{X_k})X_t\beta_t\|_2 \geq \sqrt{\Delta_n}$, we have for any $w' \in (0,1)$,

$$\begin{aligned} P\Big[\cup_{k\in M_4} \{R_k^* - R_{k\vee t}^* < (1-w')^2 \Delta_n\}\Big] \\ \leq\ & P\Big[\cup_{k\in M_4} \{\|(P_{X_{k\vee t}} - P_{X_k})\epsilon\|_2 > w'\sqrt{\Delta_n}\}\Big] \\ \leq\ & P\Big[\|P_t\epsilon\|_2 > w'\sqrt{\Delta_n}\Big] \\ \leq\ & \exp\{-c'\Delta_n\}. \end{aligned}$$

Since $R_k \geq R_k^*$, this implies that for $w \in (0,1)$, we have

$$\begin{aligned} & P\left[\cup_{k\in M_4}\{R_k - R_{k\vee t} < \Delta_n(1-w)\}\right] \\ & \leq\ P\left[\cup_{k\in M_4}\{R_k^* - R_{k\vee t}^* < \Delta_n(1-w/2)\right] \\ & \quad + P\left[\cup_{k\in M_4}\{R_{k\vee t}^* - R_{k\vee t} < \Delta_n w/2\}\right] \\ & \leq\ 2\exp\{-c'\Delta_n\} \to 0. \end{aligned} \tag{2.24}$$

For the last inequality, we use the fact that $n\tau_{1n}^2\lambda_m^n(R_{k\vee t}^* - R_{k\vee t})$ has exponential tails, similar to (2.11.1). To see this, let $X_{k\vee t} = U_{n\times r}\Lambda_{r\times r}V'_{r\times|k\vee t|}$ be the SVD of $X_{k\vee t}$, where $r = rank(X_{k\vee t})$. Then, $P_{k\vee t} = UU'$ is the projection matrix onto the column space of $X_{k\vee t}$, and hence

$$\begin{aligned} 0 \leq R_{k\vee t}^* - R_{k\vee t} &= Y'U\left(\Lambda^2(\tau_{1n}^{-2}I + \Lambda^2)^{-1} - I\right)U'Y \\ &= \tau_{1n}^{-2}Y'U(\tau_{1n}^{-2}I + \Lambda^2)^{-1}U'Y \\ &\leq (n\tau_{1n}^2\lambda_m^n)^{-1}Y'UU'Y. \end{aligned}$$

Since U is a unitary matrix with rank at most $(K+1)|t|$, we have

$$\begin{aligned} P\left[\cup_{k\in M_4}\{R_{k\vee t}^* - R_{k\vee t} < -\Delta_n w/2\}\right] &\preceq \exp\{-w'n\tau_{1n}^2\lambda_m^n\Delta_n\}\, p_n^{(K+1)|t|} \\ &\preceq \exp\{-p_n^{2+\delta}\Delta_n + (K+1)|t|\log p_n\} \\ &\preceq \exp\{-c'\Delta_n\}. \end{aligned}$$

Due to (2.24), Lemma 2.11.2 & 2.11.3, for $0 < c = 3w < 1$, we have

$$
\begin{aligned}
& P\left[\cup_{k\in M_4}\{\tilde{R}_k - \tilde{R}_t < \ \Delta_n(1-c)\}\right] \\
& \le P\left[\cup_{k\in M_4}\{\tilde{R}_k - \tilde{R}_{k\vee t} < \Delta_n(1-2w)\}\right] + P\left[\cup_{k\in M_4}\{\tilde{R}_{k\vee t} - \tilde{R}_t < -\Delta_n w\}\right] \\
& \le P\left[\cup_{k\in M_4}\{R_k - R_{k\vee t} < \Delta_n(1-w)\}\right] + P\left[\cup_{k\in M_4}\{R_t - R_{k\vee t} > \ w^2\Delta_n\}\right] \\
& \le \exp\{-c'\Delta_n\} + P\left[\cup_{k\in M_4} R_t^* - R_{k\vee t}^* > w^2\Delta_n\right] + P\left[R_t - R_t^* > w^2\Delta_n\right] \\
& \le P\left[\chi^2_{K|t|} > w^2\Delta_n\right] + 2\exp\{-c'\Delta_n\} \\
& \le 3\exp\{-c'\Delta_n\} \ \to \ 0, \ \text{uniformly in } k \in M_4.
\end{aligned}
$$

By restricting to the event $C_n := \left\{\tilde{R}_k - \tilde{R}_t \ge \Delta_n(1-c), \forall k \in M_4\right\}$, which has $P(C_n) \ge 1 - 3\exp\{-c'\Delta_n\}$, we have due to Condition 3.3.3,

$$
\begin{aligned}
\sum_{k\in M_4} BF(k,t) &\preceq \sum_{k\in M_4} (n\tau_{1n}^2\lambda_m^n)^{|t|/2}\,(\lambda_m^n)^{-|t|/2}\,s_n^{|k|-|t|}\exp\left\{-\frac{1}{2\sigma^2}(\tilde{R}_k - \tilde{R}_t)\right\} \\
&\preceq \sum_{k\in M_4}\left(p_n^{2+3\delta}\vee n\right)^{|t|/2} p_n^{\delta|t|/2} s_n^{|k|-|t|}\exp\left\{-\Delta_n(1-c)/2\sigma^2\right\} \\
&\preceq \exp\left\{-\tfrac{1}{2\sigma^2}\left(\Delta_n(1-c) - \sigma^2|t|\ \log\left(p_n^{2+3\delta}\vee n\right) - \sigma^2|t|(2+\delta)\log p_n\right)\right\} \\
&\preceq \exp\left\{-\tfrac{1}{2\sigma^2}\left(\Delta_n(1-c) - w'\gamma_n\right)\right\} \to 0,
\end{aligned}
$$

where $w' \in (0,1)$, and $c < 1 - w'$. Therefore, we have

$$
\sum_{k\in M_4} BF(k,t) \xrightarrow{\text{P}} 0. \tag{2.25}
$$

Now, by combining (2.16), (2.21), (2.22) and (2.25), we get $\sum_{k\neq t} BF(k,t) \xrightarrow{\text{P}} 0$, which implies Theorem 2.4.1. □

2.11.3 Proof of Theorem 2.4.2

The posterior of Z in this case is obtained by integrating out σ^2 along with β from the joint posterior in (2.9), that is,

$$P(Z \mid Y) \propto Q_k s_n^{|k|} \left(Y'Y - \tilde{\beta}'(D_k + X'X)\tilde{\beta} + \alpha_2\right)^{-\left(\frac{n}{2}+\alpha_1\right)}. \tag{2.26}$$

Due to Lemmas 2.11.1 & 2.11.2, we obtain

$$\frac{P(Z=k|Y)}{P(Z=t|Y)} \preceq (n\tau_{1n}^2 \lambda_m^n (1-\phi_n))^{-(r_k^*-r_t)/2} \, (\lambda_m^n)^{-|t|/2} \, s_n^{|k|-|t|} \left(\frac{\tilde{R}_k+\alpha_2}{\tilde{R}_t+\alpha_2}\right)^{-\left(\frac{n}{2}+\alpha_1\right)}.$$

Next define

$$\rho_n := \frac{\tilde{R}_t+\alpha_2}{n\sigma^2} - 1.$$

We shall now show that $\rho_n = o_P(1)$. Due to Lemmas 2.11.2 & 2.11.3, we have

$$\frac{R_t^*(1-\xi_n)+\alpha_2}{n\sigma^2} - 1 \leq \rho_n \leq \frac{R_t^*+\alpha_2}{n\sigma^2} + \frac{R_t - R_t^*}{n\sigma^2} - 1.$$

Therefore, for $\epsilon > 2\xi_n \to 0$,

$$\begin{aligned} P\left(|\rho_n| > 2\epsilon\right) &\leq P\left[\left|\frac{R_t^*}{n\sigma^2} - 1\right| > \epsilon\right] + P[R_t - R_t^* \geq \epsilon n\sigma^2] \\ &\leq 2\exp(-c'n), \end{aligned} \tag{2.27}$$

due to Lemma 2.11.3. This implies

$$\begin{aligned} \frac{P(Z=k|Y)}{P(Z=t|Y)} \preceq\ & (n\tau_{1n}^2 \lambda_m^n (1-\phi_n))^{-(r_k^*-r_t)/2} \, (\lambda_m^n)^{-|t\wedge k^c|/2} \, s_n^{|k|-|t|} \\ & \left(1 + (\tilde{R}_k - \tilde{R}_t)/n(1+\rho_n)\sigma^2\right)^{-\left(\frac{n}{2}+\alpha_1\right)}, \end{aligned} \tag{2.28}$$

where ρ_n satisfies (2.27).

We will now consider only the models of dimension at most $m_n^* = |t| + wn/\log p_n$, with $w = \delta/8(1+\delta)^2$, and consider the subsets of models $M_u^* := M_u \cap \{k : r_k \leq m_n^*\}$,

for $u = 2, 3, 4$. We first define $x_n := (r_k - r_t) \log p_n / n < \delta/8(1+\delta)^2$, and note that for $s < \delta/4$,

$$\begin{aligned} t_n := -\log(1 - 2(1+s)x_n) &< \tfrac{2(1+s)x_n}{1-2(1+s)x_n} \\ &< 2(1+\delta/2)x_n. \end{aligned} \tag{2.29}$$

Consider $\tilde{\epsilon}$ small such that $(1+2s)(1-\tilde{\epsilon}) > (1+s)$. Restricting to the high probability event $\{\cup_{\{d>r_t\}} U(d)\} \cap \{|\rho_n| < \tilde{\epsilon}\}$, and proceeding in the similar way as in Section 2.11.2.2, we obtain due to (2.28) & (2.29),

$$\begin{aligned} \sum_{k \in M_2^*} \tfrac{P(Z=k|Y)}{P(Z=t|Y)} &\leq \sum_{k \in M_2^*} (p_n^{1+\delta} \vee \sqrt{n})^{-(r_k - r_t)} \, s_n^{(|k|-|t|)} \exp\left\{\left(\tfrac{n}{2} + \alpha_1\right) t_n\right\} \\ &\preceq \sum_{k \in M_2^*} (p_n^{1+\delta} \vee \sqrt{n})^{-(r_k - r_t)} \, s_n^{(|k|-|t|)} p_n^{-(1+\delta/2)(r_k - r_t)} \\ &\sim \nu_n \to 0, \ as \ n \to \infty. \end{aligned}$$

This, together with the proof for large models in Section 2.11.2.3, imply

$$\sum_{k \in M_2^* \cup M_3^*} \tfrac{P(Z=k|Y)}{P(Z=t|Y)} \xrightarrow{\mathrm{P}} 0.$$

Now, we consider the models in M_4. If $\Delta_n = o(n)$, similar to the proof in Section 2.11.2.4, we have

$$\begin{aligned} \sum_{k \in M_4^*} \tfrac{P(Z=k|Y)}{P(Z=t|Y)} &\preceq \sum_{k \in M_4^*} \left(p_n^{2+3\delta} \vee n\right)^{|t|/2} p_n^{\delta|t|/2} \, s_n^{(|k|-|t|)} \left(1 + \tfrac{\tilde{R}_k - \tilde{R}_t}{n(1+\rho_n)\sigma^2}\right)^{-\left(\frac{n}{2}+\alpha_1\right)} \\ &\preceq \left(p_n^{2+3\delta} \vee n\right)^{|t|/2} p_n^{|t|(1+\delta/2)} \exp\left\{-\tfrac{(1-c)\Delta_n}{2\sigma^2(1+\tilde{\epsilon})}\right\} \\ &\preceq \exp\left\{-\tfrac{1}{2\sigma^2}\left(\Delta_n(1-c') - \gamma_n\right)\right\} \\ &\to 0, \ as \ n \to \infty. \end{aligned}$$

Also, if $\Delta_n \sim n$, by taking $\tilde{\epsilon} < 1/2$, we have

$$\begin{aligned} \sum_{k \in M_4^*} \tfrac{P(Z=k|Y)}{P(Z=t|Y)} &\preceq \left(p_n^{2+3\delta} \vee n\right)^{|t|/2} p_n^{(1+\delta)|t|} \left(1 + \tfrac{\Delta_n(1-c')}{4n\sigma^2}\right)^{-\left(\frac{n}{2}+\alpha_1\right)} \\ &\preceq (p_n \vee n)^{(2+3\delta)|t|} \, e^{-w'n} \to 0. \end{aligned}$$

Hence Theorem 2.4.2 is proved. □

2.11.4 Proof of Lemma 2.6.1

Proof. The rows of X_k are n independent sub-Gaussian random isotropic random vectors in $R^{|k|}$. Note that $|k| \leq m_n$ implies $|k| = o(n)$. Due to Theorem 5.39 of Vershynin (2012), with probability at least $1 - 2\exp(-c\,s)$, we have

$$\phi_{min}\left(\frac{X_k' X_k}{n}\right) > \left(1 - C\sqrt{\frac{|k|}{n}} - \sqrt{\frac{s}{n}}\right)^2, \tag{2.30}$$

where c and C are absolute constants that depend only on the sub-Gaussian norms of the rows of the matrix X_k.

Let us fix $s = n(1-\phi)$ for some $\phi > 0$, and define the event given by Equation (2.30) as A_k. We then have $P[A_k^c] < 2\exp(-c(1-\phi)n)$ for all k. By taking an union bound over $\{k : |k| \leq m_n\}$, we obtain

$$\begin{aligned} P[\cup_{|k|\leq m_n} A_k^c] &\leq p_n^{m_n} \exp(-c(1-\phi)n) \\ &= \exp\left\{\tfrac{n}{2+\nu} - c(1-\phi)n\right\} \to 0, \end{aligned}$$

if $\nu > (\frac{1}{c(1-\phi)} - 2)$. Therefore, in the event $\cap_{|k|\leq m_n} A_k$, whose probability goes to 1, we have $\phi_{min}\left(X_k' X_k/n\right) \geq \phi^2/4 - O(\sqrt{m_n/n}) > 0$, for all k. □

CHAPTER III

Scalable and Consistent Variable Selection for High Dimensional Logistic Regression

3.1 Introduction

With the increased ability to collect and store large amounts of data, we have the opportunities and challenges to analyze data with a large number of covariates or features per subject. When the number of covariates in a regression model is greater than the sample size, the parameter estimation problem becomes ill posed, and variable selection is usually a natural first-step. There have been extensive studies on variable selection in high dimensional settings, especially since the advent of Lasso Tibshirani (1996b), an L_1 regularized regression method for variable selection. Other penalization methods for sparse model selection include smoothly clipped absolute deviation (SCAD) Fan and Li (2001), adaptive Lasso Zou (2006), minimum concave penalty (MCP) Zhang (2010), and many variations of such methods. Though many of these methods are first introduced in the context of linear regression, their theoretical properties and optimization methods for logistic regression and other generalized linear models (GLM) have also been studied. van de Geer S. A. (2008) proved oracle inequalities for L_1 penalized high dimensional GLM, whereas the oracle properties of Fan and Peng (2004) also hold for GLM. Friedman et al. (2008) and Breheny

and Huang (2011) proposed coordinate descent algorithms for convex and nonconvex penalized regression methods, respectively. Park and Hastie (2007) and Huang and Zhang (2012) studied other optimization approaches for L_1 penalized GLM. The computational complexity of these algorithms typically grows linearly in p.

The literature on high dimensional Bayesian variable selection has been focusing mostly on linear models, but most techniques generalize, with some efforts, to logistic regression. It has been understood that most penalization methods have Bayesian interpretations, because all the methods share the basic desire of shrinkage towards sparse models. We refer to Bhattacharya et al. (2014); Johnson and Rossell (2012); Park and Casella (2008); Ročková and George (2014) for some recent work on Bayesian shrinkage. An advantage of Bayesian methods for variable selection is that Markov Chain Monte Carlo (MCMC) techniques can be used to explore the posterior distributions, which often offer a more informative approach to model selection than the corresponding penalization method with a highly non-convex optimization problem. For instance, the methods proposed recently by Liang et al. (2013), Narisetty and He (2014), and Shen et al. (2012) are similar to the L_0 penalty, which is generally considered to be desirable for model selection consistency.

It is of great importance to address the following two issues related to Bayesian model selection methods. The first one is theoretical ability in handling high dimensional covariates, especially when p is greater than n. The Bayesian model selection consistency has been examined only very recently in the cases of $p > n$. The other issue is its computational complexity for large p problems. While optimization algorithms with the complexity that is linear in p are usually available to solve the penalization problems, many of the existing Bayesian methods that use non-degenrate priors require repeated sampling of a p-dimensional variate. Drawing from a p-variate normal distribution with a non-sparse covariance matrix requires operations in the order of at least p^2. In this paper we attempt to cross both hurdles by developing

a new pseudo-Bayesian model selection method that has strong selection consistency when p grows sub-exponentially with n but avoids the needs to use operations of order p^2 in each iteration of an MCMC algorithm. Sampling a p-dimensional variate may not be required if one uses a point mass spike prior similar to what is used by Hans et al. (2007) and Liang et al. (2013). The sampling methods for these methods are analogous to stepwise selection, but our proposed method in this paper allows for a more general move in its MCMC iterations. We defer further discussion to Section 3.3.2.

We adopt the well-known spike and slab priors on the regression coefficients. Gaussian spike and slab priors have a special place in linear models because of the conjugacy of these priors George and McCulloch (1993). Even though the conjugacy is not preserved for logistic regression, we find that the t approximation to the logistic function proposed in Albert and Chib (1993) and the normal scale-mixture representation of the t distribution make the standard Gibbs sampler computationally convenient. The Gibbs sampler however requires sampling from a p-variate normal distribution with a non-sparse covariance matrix, which is not so scalable for large p. A major contribution of this paper is our proposal to replace the covariance matrix in the Gibbs sampler by a sparse one so that no sampling of high dimensional variates will be required. The resulting algorithm is called Skinny Gibbs, because we use a skinny covariance matrix in the Gibbs algorithm. We might view Skinny Gibbs as an approximation to the usual Gibbs sampler, but more importantly, we show that Skinny Gibbs is indeed a Gibbs sampler on its own with a different stationary distribution, but there is no sacrifice on the strong model selection consistency that we would expect from the usual Gibbs sampler. Before we move on, we would like to mention that the proposed model selection method is strictly speaking not a Bayesian method, because we are using priors that depend on the sample size and the number of variables. For the lack of better terminology, we continue to use the misnomer in

this article. For a discussion of the Bayesian viewpoint on model selection, we refer to Kass and Raftery (1995); O'hara and Sillanpaa (2009).

The rest of the paper is organized as follows. In Section 3.2, we describe our model setup, including the prior distributions and the standard Gibbs sampler, and then propose Skinny Gibbs as a new model selection algorithm. In Section 3.3, we present the strong selection consistency results for the proposed method. In Section 3.4, we compare the proposed Skinny Gibbs approach to model selection with a number of leading penalization methods in simulated settings. In Section 3.5, we present empirical studies on two examples to demonstrate how the proposed methodology works with real data. We provide a conclusion in Section 3.9. In Section 3.8, proofs for all the theoretical results are given. In the supplementary materials, we demonstrate stability and convergence of the Skinny Gibbs chain, and provide a small study to show time improvement of Skinny Gibbs from the standard Gibbs sampler.

3.2 Variable selection for logistic regression

Our data contains an $n \times 1$ binary response vector denoted by $\mathbf{E} = (E_1, \cdots, E_n)^T$ and an $n \times p_n$ design matrix X. We use p_n for the model dimension to emphasize its dependence on the sample size n. We assume that the columns of X are standardized to have zero mean and unit variance. We use x_i to denote the i^{th} row of X, which contains covariates for the i^{th} response E_i. Moreover, X_A will be used to denote the $n \times |A|$ dimensional submatrix of X containing the columns indexed by A, and $|A|$ is the cardinality of A. Logistic regression models the conditional distribution of E given X with the logit link, that is,

$$P[E_i = 1|x_i] = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}, \ i = 1, \cdots, n, \tag{3.1}$$

for some unknown parameter $\beta \in R^{p_n}$. The logistic model is one of the most widely used statistical models for binary outcomes. This paper attempts to address the problem of variable selection when the number of predictors p_n is large. If p_n is large relative to n, even the estimation problem is ill posed without any further assumptions on the model parameters. We work under the assumption that there is a true parameter vector β that is sparse in the sense that it has only a small number of non-zero components. Even under this assumption, it is a challenge to find the active predictors in the model.

In the Bayesian variable selection literature, spike and slab priors on β are commonly used. The idea is to introduce binary latent variables Z_j for the j-th component of β, which indicates whether the j^{th} covariate is active (i.e., having a nonzero coefficient). Then, priors on β_j given Z_j are specified as

$$\beta_j \mid Z_j = 0 \sim \pi_0(\beta_j); \quad \beta_j \mid Z_j = 1 \sim \pi_1(\beta_j), \tag{3.2}$$

where π_0 and π_1 are called the spike and slab priors, respectively. We refer to Mitchell and Beauchamp (1988), George and McCulloch (1993), Ishwaran and Rao (2005) and Narisetty and He (2014) for further details. For linear regression with Gaussian errors, both the spike and slab priors are often taken to be Gaussian with a small and a large variance, respectively. An advantage of this approach for linear regression is that the conditionals of the Gibbs sampler are standard distributions due to conjugacy of those priors. Though they are not conjugate for the logistic model, the well-known normal scale mixture representation of the logistic distribution due to Stefanski (1991) enables us to derive the conditional distributions used in the Gibbs sampler. More specifically, let Y_i follow the logistic distribution with location parameter $x_i\beta$, and $E_i = \mathbb{1}_{\{Y_i>0\}}$ in distribution. Then, Y_i can be equivalently represented as

$$Y_i|s_i \sim N(x_i\beta, s_i^2), \;\; s_i/2 \sim F_{KS},$$

where F_{KS} is the Kolomogorov-Smirnov distribution whose CDF is given by

$$G(\sigma) = 1 - 2\sum_{n=1}^{\infty}(-1)^{n+1}\exp(-2n^2\sigma^2). \tag{3.3}$$

By introducing the latent variables Y_i, we can implement the usual Gibbs sampler for logistic regression with simple conditionals. We will however need to draw s_i's from their conditional distributions, which we shall discuss in Section 3.2.2.

To achieve appropriate shrinkage and ensure model selection consistency, we consider the priors in Equation (3.2) to be sample-size dependent and a prior on Z_j to induce sparsity on the model space.

3.2.1 Shrinking and diffusing priors

The priors on the binary latent variables Z_j and the corresponding regression coefficients β_j are given by

$$\begin{aligned} &\beta_j \mid Z_j = 0 \sim N(0, \tau_{0,n}^2), \;\; \beta_j \mid Z_j = 1 \sim N(0, \tau_{1,n}^2) \\ &P(Z_j = 1) = 1 - P(Z_j = 0) = q_n, \end{aligned} \tag{3.4}$$

for $j = 1, \cdots, p_n$ (with independence across different j), where the constants $\tau_{0,n}^2, \tau_{1,n}^2$ and q_n are further specified below. Broadly speaking, we consider the settings where $\tau_{0,n}^2 \to 0$, and $\tau_{1,n}^2 \to \infty$ as $n \to \infty$. The specific rates for $\tau_{0,n}^2$, $\tau_{1,n}^2$ are given by Condition 3.3.4. The intuition behind such choices is that the inactive covariates will be identified with zero Z_j values, where small values of β_j relative to τ_{0n}^2 are truncated to zero. The diverging parameter $\tau_{1,n}^2$ forces the inactive covariates to be classified under $Z_j = 0$ because the prior probability around zero becomes negligible as $n \to \infty$. Finally, we shall use $q_n \sim p_n^{-1}$ to encourage the models to be sparse, i.e., it bounds the apriori size of $|Z| := \sum_{j=1}^{p_n} Z_j$ to be small, where Z denotes the vector of Z_j. The posterior probabilities of the binary variables Z_j will be used to select the

active covariates.

In the linear regression case, Narisetty and He (2014) argued that the prior specification similar to (3.4) implies a posterior that is asymptotically similar to the L_0 penalized likelihood. More specifically, when $n\tau_{0,n}^2 = o(1)$, the prior parameters imply a penalty in the order of $\log(\sqrt{n}\tau_{1,n}q_n^{-1})$ for each additional covariate added in the model. This is the reason for allowing these prior parameters to change with the sample size n so as to obtain the appropriate amount of penalization. In this paper, we propose a fast and scalable Gibbs sampler that preserves the similarity to the L_0 penalty and achieves the strong selection consistency (see Section 3.3).

3.2.2 Gibbs sampler

As a prelude to our proposed Skinny Gibbs sampler, we first present the usual Gibbs sampler corresponding to (3.4), which will provide motivation for our proposal of Skinny Gibbs. **In the rest of the paper, all the distributions are conditional on X but we suppress it in the notations for convenience.** By considering

$$E_i = \begin{cases} 1 & \text{if } Y_i \geq 0 \\ 0 & \text{if } Y_i < 0 \end{cases} \tag{3.5}$$

$$Y_i \overset{\text{ind}}{\sim} N(x_i\beta, s_i^2), \qquad s_i/2 \overset{\text{ind}}{\sim} F_{KS},$$

together with the priors in (3.4), the joint posterior of β, Z, Y and

$$W = Diag\left(s_1^{-2}, ..., s_{p_n}^{-2}\right) \tag{3.6}$$

is given by

$$\begin{aligned} f(\beta, W, Y, Z \mid \mathbf{E}) \quad \propto \quad & \prod_{i=1}^{n} \phi\left(Y_i, x_i\beta, s_i^2\right) \mathbb{1}\{E_i = \mathbb{1}\{Y_i \geq 0\}\}\ g(s_i) \\ & \times \prod_{j=1}^{p_n} \left((1-q_n)\pi_0(\beta_j)\right)^{1-Z_j} \left(q_n\pi_1(\beta_j)\right)^{Z_j}, \end{aligned} \tag{3.7}$$

where $\pi_0(x) = \phi(x, 0, \tau_{0,n}^2)$, $\pi_1(x) = \phi(x, 0, \tau_{1,n}^2)$, $\phi(x, \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2 evaluated at x, and $g(\sigma) = (d/d\sigma)F_{KS}(\sigma/2)$ is the density function of two times the KS variable.

The conditional distributions of the Gibbs sampler can be derived from (3.7) as follows. The conditional distribution of β is

$$f(\beta \mid W, Y, Z, \mathbf{E}) \propto \exp\left\{-\frac{1}{2}(\beta' X' W X \beta - 2\beta' X' W Y)\right\} \exp\left\{-\frac{1}{2}\beta' D_z \beta\right\},$$

where $D_z = Diag(Z\tau_{1,n}^{-2} + (1 - Z)\tau_{0,n}^{-2})$. That is,

$$\beta \mid (W, Y, Z, \mathbf{E}) \sim N\left((X'WX + D_z)^{-1} X'WY, (X'WX + D_z)^{-1}\right). \tag{3.8}$$

The conditional distributions of Y_i are independent with the marginals given by

$$f(Y_i \mid \beta, W, Z, \mathbf{E}) \propto \begin{cases} \phi\left(Y_i, x_i\beta, s_i^2\right) \mathbb{1}\{Y_i > 0\} & \text{if } E_i = 1, \\ \phi\left(Y_i, x_i\beta, s_i^2\right) \mathbb{1}\{Y_i < 0\} & \text{if } E_i = 0, \end{cases} \tag{3.9}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. The conditional distributions of Z_j are independent (across j) and given by

$$P(Z_j = 1 \mid \beta, W, Y, \mathbf{E}) = \frac{q_n \phi(\beta_j, 0, \tau_{1,n}^2)}{(1 - q_n)\phi(\beta_j, 0, \tau_{0,n}^2) + q_n \phi(\beta_j, 0, \tau_{1,n}^2)}. \tag{3.10}$$

The conditional distribution of W is described in terms of the independent distributions of s_i as

$$f(s_i \mid \beta, Y, Z, \mathbf{E}) \propto \phi(Y_i, x_i\beta, s_i^2)\ g(s_i). \tag{3.11}$$

In this Gibbs sampler, sampling from the distribution (3.11) is not as straightforward as the others. Holmes and Held (2006) proposed a rejection sampling algorithm. Albert and Chib (1993) noted that the univariate logistic density can be approximated

well by a t-density. O'Brien and Dunson (2004) later used the t-approximation of the logistic density for multivariate logistic regression. We simply adopt those ideas to proceed as follows.

Let us denote the t-distribution that approximates the KS distribution by $T = w\ t_\nu$, i.e., t with ν degrees of freedom and scale parameter w. Due to the Gaussian scale mixture representation of the t-distribution, it can be equivalently represented as

$$T|\phi \sim N(0, \phi^2),\ \ \phi^2 \sim w^2 IG\left(\nu/2, \nu/2\right), \tag{3.12}$$

where IG is the inverse gamma distribution. Following O'Brien and Dunson (2004), we take $w^2 = \pi^2(\nu-2)/3\nu$ and $\nu = 7.3$ so that the resulting distribution of T is nearly indistinguishable from the KS distribution. Using this approximation, the sampling of (3.11) can be done using an inverse Gamma distribution.

However, when p_n is large, the real bottleneck with the usual Gibbs sampler lies in its need to sample from the p_n-variate normal distribution for β given by (3.8). For linear regression, Guan and Stephens (2011) avoided such sampling by integrating β out and devise an MCMC method that samples Z directly. However, this technique does not seem to generalize easily to logistic regression. A direct sampling scheme would require handling a $p_n \times p_n$ covariance matrix of general forms, which is expensive in both CPU and memory. Even if this task is decomposed into componentwise sampling by a further Gibbs iteration, it requires operations in the order of p_n^2, making Bayesian model selection algorithm less competitive with the penalty based optimization methods. Moreover, the complexity of this computation is not reduced even if the model size ($|Z|$) in each iteration is small.

3.2.3 Skinny Gibbs algorithm

We propose the Skinny Gibbs algorithm as a simple yet effective modification of the Gibbs sampler to avoid the computational complexity in the case of large p_n. The

idea is to split β into two parts in each Gibbs iteration, corresponding to the "active" (with the current $Z_j = 1$) and "inactive" (with the current $Z_j = 0$) sub-vectors. The active part has a low dimension, and is sampled from the multivariate normal distribution. The inactive part has a high dimension, but we simply sample it from a normal distribution with independent marginals. More specifically, the Skinny Gibbs sampler proceeds as follows, after an initialization.

(a) Decompose $\beta = (\beta_A, \beta_I)$, where β_A and β_I contain the components of β corresponding to $Z_j = 1$ and $Z_j = 0$, respectively. Similarly let $X = [X_A, X_I]$. Then, generate

$$\beta_A \mid (W, Y, Z, \mathbf{E}) \sim N(m_A, V_A^{-1}), \qquad \beta_I \mid (W, Y, Z, \mathbf{E}) \sim N(0, V_I^{-1}),$$

where $V_A = (X_A' W X_A + \tau_{1n}^{-2} I)$, $m_A = V_A^{-1} X_A' W Y$, and $V_I = Diag\,(X_I' X_I + \tau_{0n}^{-2} I) = (n + \tau_{0n}^{-2}) I$. Note that the dimension of V_A is only $|Z|$.

(b) Generate Z_j $(j = 1, \cdots, p_n)$ sequentially based on

$$\frac{P[Z_j = 1 \mid Z_{-j}, \beta, W, Y, \mathbf{E}]}{P[Z_j = 0 \mid Z_{-j}, \beta, W, Y, \mathbf{E}]} = \frac{q_n \phi(\beta_j, 0, \tau_{1,n}^2)}{(1 - q_n)\phi(\beta_j, 0, \tau_{0,n}^2)} \times \exp\left\{ \beta_j X_j' W (Y - X_{C_j} \beta_{C_j}) + \frac{1}{2} X_j' (I - W) X_j \beta_j^2) \right\},$$

where Z_{-j} is the Z vector without the jth component, and C_j is the index set corresponding to the active components of Z_{-j}, i.e., $C_j = \{k : k \neq j, Z_k = 1\}$.

(c) The conditional distribution of Y is changed to

$$f(Y_i \mid \beta, W, Z, \mathbf{E}) \propto \begin{cases} \phi\left(Y_i, x_{Ai}\beta_A, s_i^2\right) \mathbb{1}\{Y_i > 0\} & \text{if } E_i = 1, \\ \phi\left(Y_i, x_{Ai}\beta_A, s_i^2\right) \mathbb{1}\{Y_i < 0\} & \text{if } E_i = 0, \end{cases}$$

(d) The conditional distribution of s_i is

$$f(s_i \mid \beta, Y, Z, \mathbf{E}) \propto \phi(Y_i, x_{Ai}\beta_A, s_i^2)\ g(s_i).$$

In (a), the update of β is changed such that the coefficients corresponding to $Z_j = 1$ (denoted by β_A) and those corresponding to $Z_j = 0$ (denoted by β_I) are sampled independently. Furthermore, the components of β_I are updated independently. This is in contrast with the usual Gibbs, where the entire β is updated jointly. It is worth noting that the precision matrix of β_A is just the corresponding sub-matrix of the precision matrix of β, which is $V_z = (X'WX + D_z)$. Essentially, Skinny Gibbs sparsifies the precision matrix V_z as

$$V_z = \begin{pmatrix} X_A' W X_A + \tau_{1n}^{-2} I & X_A' W X_I \\ X_I' W X_A & X_I' W X_I + \tau_{0n}^{-2} I \end{pmatrix}$$

$$\Downarrow$$

$$\begin{pmatrix} X_A' W X_A + \tau_{1n}^{-2} I & 0 \\ 0 & (n + \tau_{0n}^{-2}) I \end{pmatrix}.$$

This modification in step (a) alters the Gibbs sampler in such a non-trivial way that the correlation structure among the coefficients β_j is lost. Without any compensation, the modified sampler would not converge to the right stationary distribution. The step (b) of the proposed Skinny Gibbs is designed to compensate for the loss in step (a), but the computational complexity in step (b) is minimal. In the next section, we provide theoretical justification for the Skinny Gibbs sampler.

3.3 Theoretical results

In this section, we provide theoretical results about the asymptotic properties of Skinny Gibbs. We show that Skinny Gibbs has a stationary posterior distribution that preserves the strong model selection consistency. We first introduce the following notations.

Notations: We use k (and s) to denote a generic model and t to denote the true model. A model is treated both as a $p_n \times 1$ binary vector similar to Z and as the set containing the active covariates, but this will be clear depending on the context. The size of the model k is denoted by $|k|$. For any $p_n \times 1$ vector v, $v(k)$ is used to denote the $|k| \times 1$ vector containing the components of v corresponding to model k. We denote the true regression vector as $\beta_0(t)$, and for any $k \supset t$, $\beta_0(k)$ denotes the $|k| \times 1$ vector having $\beta_0(t)$ for t and zeroes for $k \cap t^c$. For sequences a_n and b_n, $a_n \sim b_n$

means $\frac{a_n}{b_n} \to c$ for some $c > 0$, $b_n \succeq a_n$ (or $a_n \preceq b_n$) means $b_n = O(a_n)$, and $b_n \succ a_n$ (or $a_n \prec b_n$) means $b_n = o(a_n)$.

The log-likelihood for a model k is

$$L_n(\beta(k)) := \sum_{i=1}^{n} E_i \log F\left(x_i\beta(k)\right) + (1 - E_i) \log(1 - F\left(x_i\beta(k)\right)), \tag{3.13}$$

where $F(.)$ is the cdf of the logistic distribution. Let

$$s_n(\beta(k)) = \frac{\partial L_n(\beta(k))}{\partial \beta(k)} = \sum_{i=1}^{n} \left(E_i - \mu_i(\beta(k))\right) x_i, \tag{3.14}$$

with $\mu_i(\beta(k)) = \frac{\exp\{x_i\beta(k)\}}{1+\exp\{x_i\beta(k)\}}$. The negative Hessian of $L_n(\beta(k))$ is

$$H_n(\beta(k)) = -\frac{\partial^2 L_n(\beta(k))}{\partial \beta(k) \partial \beta(k)'} = \sum_{i=1}^{n} \sigma_i^2(\beta(k)) x_i x_i', \tag{3.15}$$

where $\sigma_i^2(\cdot) = \mu_i(\cdot)(1 - \mu_i(\cdot)$. Note that in our notations, x_i and X are restricted to the model under consideration, even though it is not explicitly displayed. That is, x_i in Equations (3.14) and (3.15) is a $|k| \times 1$ vector containing the components corresponding to model k. Therefore, the dimension of $s_n(\beta(k))$ is $|k| \times 1$ and that of $H_n(\beta(k))$ is $|k| \times |k|$. We shall also use μ_i and σ_i^2 in place of $\mu_i(\beta_0(t))$ and $\sigma_i^2(\beta_0(t))$, respectively, for the sake of simplicity.

We first prove the following to provide the posterior that corresponds to the Skinny Gibbs sampler.

Theorem 3.3.1. *The joint posterior of β, Z, Y and W corresponding to the Skinny*

Gibbs algorithm is given by

$$
\begin{aligned}
& f(\beta, W, Y, Z = k \mid \mathbf{E}) \\
& \quad \propto |W|^{1/2} \exp\left\{-\tfrac{1}{2}(Y - X\beta(k))' W (Y - X\beta(k))\right\} \; v_n^{-|k|} \\
& \qquad \times \prod_i g(s_i) \exp\left\{-\tfrac{1}{2}\left(\beta' D_k \beta + n\beta(k^c)'\beta(k^c)\right)\right\} \mathbb{1}\{E_i = \mathbb{1}\{Y_i \geq 0\}\},
\end{aligned} \tag{3.16}
$$

where $W = Diag\,(s_1^{-2}, ..., s_{p_n}^{-2})$, $D_k = Diag(k\tau_{1n}^{-2} + (1-k)\tau_{0n}^{-2})$ *and* $v_n = \tau_{1n}(1 - q_n)/(q_n\tau_{0n})$.

Remark 7. The posterior (3.16) suggests that with everything else the same, a unit increase in the model size ($|k|$) reduces the posterior by a multiple of $v_n^{-1} = (q_n\tau_{0n})/\tau_{1n}(1-q_n)$. This hints at the following: (a) the similarity of the posterior to L_0 penalty as discussed in Subsection 3.3.1, and (b) the reason for allowing the prior parameters to depend on n (see Condition 3.3.4) so that the shrinkage implied by v_n^{-1} is at an appropriate level.

We now provide the conditions assumed for proving strong selection consistency property of Skinny Gibbs. By strong selection consistency, we mean that the posterior probability of the true model converges to one as sample size increases to infinity, as used in Johnson and Rossell (2012) and Narisetty and He (2014).

Condition 3.3.1 (On Dimension p_n)**.** $p_n \to \infty$ *and* $\log p_n = o(n)$ *as* $n \to \infty$.

Condition 3.3.2 (On Regularity of the Design)**.**

(a) The predictors are bounded, that is, $\max\{|x_{ij}|,\ 1 \leq i \leq n, 1 \leq j \leq p_n\} \leq C$, *for some* $0 < C < \infty$;

(b) for some fixed $0 \leq d < d' \leq 1$,

$$
\begin{aligned}
0 < \lambda \; &\leq \min_{k:|k| \leq m_n + |t|} \lambda_{min}\left(n^{-1} H_n(\beta_0(k))\right) \\
&\leq \max_{k:|k| \leq m_n + |t|} \lambda_{max}\left(n^{-1} X_k' X_k\right) \leq C^2 \left(\frac{n}{\log p_n}\right)^d
\end{aligned} \tag{3.17}
$$

where, $\lambda_{min}(\cdot), \lambda_{max}(\cdot)$ *are the minimum and maximum eigenvalues of their arguments respectively,*

$$m_n = \left(\left(\frac{n}{\log p_n} \right)^{\frac{1-d'}{2}} \wedge p_n \right),$$

and

(c) *for any possible model* k *with* $|k| \leq m_n + |t|$ *and any* $u \in R^n$ *in the space spanned by the columns of* $\Sigma^{1/2} X_k$*, there exists* $\delta^* > 0$ *and* $N(\delta^*)$ *such that*

$$\mathbb{E}\left[\exp\{u'\Sigma^{-\frac{1}{2}}(\mathbf{E} - \mu)\}\right] \leq \exp\left\{\frac{(1+\delta^*)u'u}{2}\right\},$$

for any $n \geq N(\delta^*)$*, where* $\mathbb{E}$ *denotes expectation over* $\mathbf{E}$ *(conditional on the design).*

Condition 3.3.3 (On True Model and Signal Strength)**.** *We assume that there exists constant* $c > 1$ *such that*

$$c|t| \leq m_n, \quad \text{and} \quad \min_{1 \leq i \leq |t|} |\beta_{0i}(t)| \geq \sqrt{\frac{c|t|\Lambda_{c|t|} \log p_n}{n}},$$

where $\beta_0(t) = (\beta_{0i}(t))_{i=1}^{|t|}$ *is the nonzero coefficients of* β *under the true model, and* $\Lambda_{c|t|} := \max_{k:|k| \leq c|t|} \lambda_{max}\left(n^{-1} X_k' X_k\right)$.

Condition 3.3.4 (Prior Parameters)**.** *The prior parameters* τ_{0n}^2*,* τ_{1n}^2 *and* q_n *are such that for some* $\delta > \delta^*$*,*

$$n\tau_{0n}^2 = o(1), \quad n\tau_{1n}^2 \sim (n \vee p_n^{2+2\delta}), \quad q_n \sim p_n^{-1}.$$

Remark 8. The upper bound on the maximum eigenvalue in Condition 3.3.2 (b) is always satisfied if $1/3 < d < d'$. This is because, $\lambda_{max}\left(n^{-1} X_k' X_k\right) \leq Trace\left(n^{-1} X_k' X_k\right) \leq C^2|k| \leq C^2(n/\log p_n)^d$ holds for any $|k| \leq m_n + |t|$ when $1/3 < d < d'$. This condition is of course weaker than the bounded maximum eigenvalue condition as assumed in

Bondell and Reich (2012). If the maximum eigenvalue here is bounded, we have the case $d = 0$, and m_n can be almost as large as $(n/\log p_n)^{1/2}$.

The lower bound in Condition 3.3.2 (b) is essentially a restricted eigenvalue condition for L_0-sparse vectors. Restricted eigenvalue (RE) conditions are routinely assumed in high-dimensional theory to guarantee some level of curvature of the objective function in lower dimensions. The RE condition with L_1-sparse vectors is assumed for L_1 penalized problems for estimation consistency (see Bickel et al. (2009), Section 6.2.3 of Bühlmann and van de Geer (2011)). The intuition behind the L_0-sparse eigenvalue condition for Skinny Gibbs is attributable to the similarity between Skinny Gibbs and the L_0 type penalization as discussed in Section 3.3.1. The lower bound in Condition 3.3.2 (b) is satisfied by sub-Gaussian random design matrices with high probability. A formal statement about this is stated below and the proof is given in Section 3.8.

Lemma 3.3.1. *Let $X_{n\times p}$ be a random design matrix with rows i.i.d. from a sub-Gaussian distribution with covariance matrix Σ. Let the principal submatrices of Σ of order $m_n + |t|$ have minimum eigenvalues bounded (away from zero). Also, assume that $\beta_0(t)$ is a $|t| \times 1$ vector satisfying $P[|x_i'\beta_0(t)| \geq M] \leq w < 1$, for some $M > 0$, where x_i is the i^{th} row of X (this is a weaker version of the condition that all the log-odds are bounded as assumed in Bühlmann and van de Geer (2011)). Then, we have*

$$0 < \lambda \leq \min_{k:|k|\leq m_n+|t|} \lambda_{min}\left(n^{-1}H_n(\beta_0(k))\right).$$

Proof. Proof is given in the Appendix. □

Condition 3.3.2 (c) is not really a restriction, because such a $\delta^* > 0$ always exists due to the sub-Gaussianity of $\Sigma^{-\frac{1}{2}}(\mathbf{E} - \mu)$. Also note that for typical random designs, the variable $u'\Sigma^{-\frac{1}{2}}(\mathbf{E} - \mu)/||u||$ is asymptotically distributed as $N(0, 1)$, so Condition 3.3.2 (c) is expected to hold for a small positive constant δ^*. In Condition 3.3.3, the

upper bound on the true model size, and the minimal signal strength match with those for penalized methods such as Lasso when $d = 0$, but impose slightly stronger conditions when $d > 0$. For the screening property of Lasso to hold, Corollary 7.6 of Bühlmann and van de Geer (2011) assumes the minimum signal to be at least in the order of $\sqrt{|t| \log p_n / n}$ and the true model size $|t| = O(\sqrt{n / \log p_n})$.

Theorem 3.3.2. *Under Conditions 3.3.1 – 3.3.4, we have*

$$P\left[Z = t \mid (\mathbf{E}, \text{ and } |Z| \leq m_n)\right] \xrightarrow{\text{P}} 1, \text{ as } n \to \infty. \text{ Moreover,}$$

$$\sum_{k \neq t; |k| \leq m_n} \frac{P[Z = k \mid \mathbf{E}]}{P[Z = t \mid \mathbf{E}]} \leq C \exp\{-\epsilon \log p_n\} \to 0, \text{ for some } C, \epsilon > 0.$$

The strong selection consistency is a stronger property than the usual Bayes factor consistency for large p_n. As Johnson and Rossell (2012) argued (see proof of Theorem 2 there) that for large $p_n > n^{1/2+\epsilon}$, the posterior of the true model relative to models of a fixed size may also be very small, i.e., it is possible that $\sum_{k \neq t; |k| = |t|+1} \frac{P[Z=k|\mathbf{E}]}{P[Z=t|\mathbf{E}]} \to \infty$, even under the Bayes factor consistency. This will make it difficult to identify the active predictors based on a finite chain, because the posterior probability of the true model can be close to zero, so that the ratios $P[Z = k \mid \mathbf{E}]/P[Z = t \mid \mathbf{E}]$ are difficult to estimate.

Remark 9. For the sake of convenience, we assume that the true model representation t is unique. If multiple representations of the true model are available (due to the existence of linearly dependent predictors) the result of Theorem 3.3.2 holds if t represents the union of the true models representations.

Remark 10. Theorem 3.3.2 justifies the use of marginal posterior probabilities $P[Z_j \mid \mathbf{E}]$ for selecting the variables as long as our search of models is restricted to model size of m_n. This is useful in practice because we only need to estimate and store p_n

marginal posterior probabilities as opposed to dealing with posterior probabilities of $\binom{p_n}{m_n}$ models.

3.3.1 Connection with L_0 penalization

In this section, we provide a discussion about the connection between L_0 penalization and the variable selection from Skinny Gibbs. Due to Theorem 3.3.1 and Equation (3.54) in Section 3.8, the maximum a posteriori (MAP) estimate of the model corresponding to Skinny Gibbs is equivalent to minimizing the following objective function.

$$B(k) := \log \int_{\beta(k)} \exp\left\{L_n(\beta(k))\right\} \exp\left\{-\tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k)\right\} d\beta(k) + \log\left(Q_{n,k}/Q_{n,t}\right), \tag{3.18}$$

where $Q_{n,k} = (n + \tau_{0n}^{-2})^{|k|/2}\ v_n^{-|k|}$, where $v_n = \tau_{1n}(1-q_n)/(q_n\tau_{0n})$ as before.

Following the proof of Theorem 3.3.2 (see Equation (3.57) and a similar argument for the reverse inequality), we have the following inequality

$$c'(|k|-|t|)\log(\sqrt{n}\tau_{1,n}q_n^{-1}) \leq B(k) - L_n(\hat{\beta}(k)) \leq C'(|k|-|t|)\log(\sqrt{n}\tau_{1,n}q_n^{-1}),$$

for some $0 < c' \leq C' < \infty$. Therefore, we have

$$B(k) \quad = \quad L_n(\hat{\beta}(k)) + \psi_{n,k}(|k|-|t|), \tag{3.19}$$

where $c'\log(n \vee p_n) \leq \psi_{n,k} \leq C'\log(n \vee p_n)$, due to Condition 3.3.4 on τ_{1n} and q_n. This implies that the MAP estimate can be (asymptotically) described as the model corresponding to minimizing the following L_0-like penalized objective function.

$$m(\beta) := L_n(\beta) + \psi_{n,k}\left(\|\beta\|_0 - |t|\right), \text{ for } \beta \in R^{p_n}.$$

Due to the bounds on $\psi_{n,k}$, any inactive covariate will be penalized in the order of $\log(n \vee p_n)$ irrespective of the size of the coefficient, which is in the same spirit as the L_0 penalty. This is however not the case with the L_1 penalty or SCAD penalty, which are directly proportional to the magnitude of the coefficient in some interval around zero.

3.3.2 Comparisons with existing Bayesian methods

Chen and Chen (2012) proposed the extended Bayesian Information Criterion (EBIC)

$$EBIC(k) = -2L_n(\hat{\beta}(k)) + |k|(\log n + 2\gamma \log p_n) \tag{3.20}$$

for model selection, which is similar to the penalized log-likelihood given by (3.19). The model selection consistency under EBIC is established by Chen and Chen (2012) for $\gamma > (1 - \frac{1}{2\kappa})$, where $p_n = O(n^\kappa)$. For high dimensional problems, such an objective function cannot be applied to all possible models. Even if we restrict ourselves to a model of size m for a relatively small m, the number of possible models $\binom{p}{m}$ could be too large. The EBIC is typically used to choose models among a much smaller number of candidate models. In the simulation comparisons in the next section, we include the use of EBIC using an initial Lasso path as done in Chen and Chen (2012).

An alternative approach to Gaussian spike priors used in this paper is to take point mass spike priors, i.e., $\beta_j \mid Z_j = 0 \sim \delta_0$, the point-mass distribution at zero. An apparent attraction of the point mass prior is that we no longer have to deal with $p_n \times p_n$ matrix computations, if we can sample from the posterior of Z_j without β_j. This can indeed be done in linear regression models, as shown in Guan and Stephens (2011). Unfortunately, the posterior $P(Z = k \mid \mathbf{E})$ does not have a closed form for logistic regression, and approximations have to be used for sampling from the posterior. In this direction, Hans et al. (2007) proposed a shotgun stochastic search (SSS) algorithm based on a Laplace approximation to the posterior. Liang

et al. (2013) proposed Bayesian subset regression (BSR) modeling using a stochastic approximation Monte Carlo (SAMC) algorithm Liang et al. (2007) that aims to avoid the potential local-trap problem for SSS by sampling a specified sub-regions of the model space uniformly. Like Skinny Gibbs, these algorithms avoid p_n^2 operations in each step of the iteration, but they are analogous to stepwise variable selection whereas Skinny Gibbs allows more general updates of the model in every iteration. For the SAMC algorithm to be competitive, the number of sub-regions used in the method needs to increase with p_n, making it less computationally competitive. Some empirical comparisons of various methods are given in Section 3.4.

For the strong selection consistency we established here, the spike prior variance τ_{0n}^2 can be arbitrarily close to zero making the limiting case of $\tau_{0n}^2 = 0$ the same as the point-mass prior for $\beta_j \mid Z_j = 0$. We note that Liang et al. (2013) used a point-mass spike prior and a slab prior whose variance depends on the size of the model, and showed strong selection consistency. However, the consistency result of Liang et al. (2013) relied on a condition on the posterior distribution itself, which makes their result indicative rather than confirmatory. In this sense, we hope that our theoretical treatment also completes the strong selection consistency theory on point-mass priors in high dimensional models.

3.3.3 Unbiasedness of Skinny Gibbs

Even though the posteriors of Z for both the usual Gibbs and Skinny Gibbs algorithms concentrate at the true model asymptotically in a similar fashion, the posteriors of β are different for these algorithms. For simplicity, let us consider the linear regression case. Skinny Gibbs for linear regression would be obtained by treating Y as observed and taking W to be equal to identity matrix in the Skinny Gibbs algorithm of Section 3.2.3. Then, from the proof of Theorem 3.3.2, the posterior

of β from Skinny Gibbs converges to

$$\beta_t \mid Y \sim N(m_t, V_{t1}^{-1}), \qquad \beta_{t^c} \mid Y \sim N(0, V_{t0}^{-1}), \tag{3.21}$$

where $V_{t1} = (X_t'X_t + \tau_{1n}^{-2} I)$, $m_t = V_{t1}^{-1} X_t'Y$, and $V_{t0} = Diag\,(X_{t^c}'X_{t^c} + \tau_{0n}^{-2} I)$. On the other hand, for the usual Gibbs it is given by

$$\beta \mid Y \sim N\left((X'X + D_t)^{-1}X'Y, (X'X + D_t)^{-1}\right), \tag{3.22}$$

where $D_t = Diag(Z\tau_{1,n}^{-2} + (1 - Z)\tau_{0,n}^{-2})$ as defined in Equation (3.2.2), with Z corresponding to the true model. From the above equations, we can see that the posterior of β_t from Skinny Gibbs is (almost) equal to the distribution of the OLS estimator given the covariates X_t (because τ_{1n}^{-2} is negligible) and hence is (nearly) unbiased. Moreover, β_{t^c} is also unbiased for zero. On the other hand, a stronger condition is needed (such as $\tau_{0n}^2 \lambda_{max}(X'X) = o(1)$) for the posterior of usual Gibbs to have the unbiasedness property. For this reason, for small samples we expect that Skinny Gibbs would be more effective in identifying the true covariates whereas the usual Gibbs would be slightly better in controlling the false positives. Our simulation results in Section 3.4 also suggest the same.

3.4 Simulation study

In this section, we study the performance of the proposed method and compare them with several existing methods by simulation studies. Let X denote the design matrix whose first p_1 columns correspond to the active covariates for which we have nonzero coefficients, while the rest correspond to the inactive ones with zero coefficients. In all the simulations, we generate each row of X independently from a normal distribution with a p-dimensional covariance matrix such that the correlation between

any pair of active covariates is equal to ρ_1, the correlation between an active covariate and an inactive covariate is ρ_2, and the correlation between any pair of inactive covariates is ρ_3. Given X, we sample Y from a logistic model $P(Y_i = 1|x_i) = e^{x_i\beta}/(1+e^{x_i\beta})$, for $i = 1, \ldots, n$. We fix $n = 100$, $p_1 = 4$, and $\beta = (1.5, 2, 2.5, 3, 0, 0, \cdots, 0)$ in all our simulations. We will specify the number of covariates p and the correlations ρ_1, ρ_2, and ρ_3 in the tables.

We report the results from the usual Gibbs sampler described in Subsection 3.2.2 (BASAD), and Skinny Gibbs (as a simplified version of BASAD), along with the results from EBIC, Bayesian Subset Regression (BSR, Liang et al. (2013)) Adaptive Lasso, SCAD, as well as MCP. For EBIC, we use the penalty coefficient $\gamma = 1$ and initial path obtained using the package "glmpath" as suggested in Chen et al. (2008). For BSR as well, we set the hyperparameter $\gamma = 1$. We use the R package "glmnet" for Adaptive Lasso, and the package "ncvreg" for SCAD and MCP. For Adaptive Lasso, the initial estimate of β is obtained from Lasso with the penalty parameter of $\lambda = 10^{-4}$. For all the penalization methods, BIC is used to select the tuning parameters. For the BASAD and Skinny Gibbs, we have three parameters to choose: τ_{0n}^2, τ_{1n}^2 and q_n. In all our empirical work, we use

$$\tau_{0n}^2 = \tfrac{1}{n},\ \tau_{1n}^2 = \max\left(\tfrac{p_n^{2.1}}{100n}, 1\right),$$

and we choose $q_n = P[Z_i = 1]$ such that $P\left[\sum_{i=1}^{p_n} Z_i = 1 > K\right] = 0.1$, for a pre-specified value of K. Our default value is $K = \max(10, \log(n))$. These choices are very similar to the implementation of BASAD for linear models in Narisetty and He (2014). The models for these methods are obtained by thresholding the marginal posterior probabilities by 0.5. This model is referred to as the median probability model by Barbieri and Berger (2004). The choice of 0.5 is a natural choice for the threshold especially when the true model is unique. For the real data examples of

Section 3.5, we investigate models of different sizes.

We will present the following model selection performance measures using 200 randomly generated datasets. Average True Positive (TP) is the average number of active covariates chosen; Average False Positive (FP) is the average number of inactive covariates chosen; The column $Z = t$ gives the proportion of choosing the true model exactly, while $Z \supset t$ is the proportion of times the true model is included in the chosen model; Finally, the column $Z_4 = t$ gives the proportion of times the chosen model of size four is the true model, which gives an idea about how well a method can order the active covariates ahead of the inactive ones.

In Table 3.1, we have four cases corresponding to the number of covariates $p = 50$ or 250, with a common correlation of $\rho_1 = \rho_2 = \rho_3 = 0$ or 0.25. The results from these cases show that BASAD and Skinny Gibbs, like other Bayesian model selection methods, have much smaller false positives than non-Bayesian methods, and do not lose much in terms of true positives. Overall, our proposed methods have higher exact identification rate ($Z = t$) but none of the methods dominate others in all the measures.

In our next simulation settings, we consider different values for ρ_1, ρ_2 and ρ_3. We consider $(\rho_1, \rho_2, \rho_3) = (0.10, 0.25, 0.50)$ and $(\rho_1, \rho_2, \rho_3) = (0.20, 0.40, 0.60)$ to see the effects of higher correlations among inactive variables and between inactive and active variables. From these results shown in Table 3.2, we observe that the performance of all the methods deteriorates in comparison to the independent covariates case (Table 3.1) as expected. However, the effects of higher correlations on our methods are less substantial in comparison to the other methods. For example, in Table 3.2, $Z = t$ and $Z_4 = t$ rates are clearly higher for BASAD and Skinny Gibbs than for the competing methods. This can be attributed to the similarity of our methods with the L_0 penalty, whose performance would be less affected by the correlations between covariates, and to the ability of Skinny Gibbs to perform broader search of the model space than

EBIC.

In Figure 3.1, we plot the proportion of active covariates (out of the four active ones) that are selected as a function of model sizes. Note that this plot does not depend on tuning in the penalization methods and shows that Skinny Gibbs has the largest proportion of active covariates across settings for model sizes less than or equal to four. Adaptive Lasso and BSR are close competitors according to this measure. In some cases, Adaptive Lasso and MCP have higher proportions for larger models but Skinny Gibbs remains competitive. Though the average numbers of correctly chosen covariates are similar for Skinny Gibbs and Adaptive Lasso, Skinny Gibbs has higher values of $Z_4 = t$, indicating higher chance of selecting the true model. We exclude BASAD in Figure 3.1 simply because its performance is almost identical to that of Skinny Gibbs but with much more computational time involved. In the supplementary materials, we provide several plots of the marginal posterior probabilities along the Skinny Gibbs iterates for some simulated data to demonstrate the stability in the convergence of the Skinny Gibbs chains.

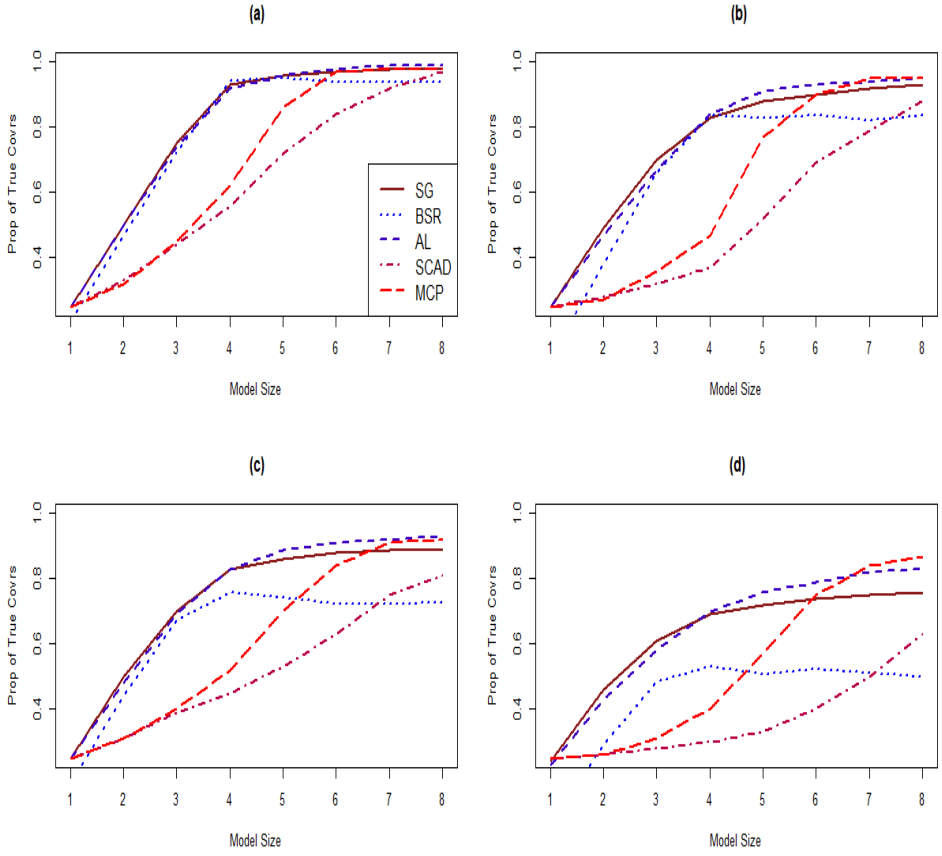
Table 3.1: Simulation results with low and moderate correlations among predictors: TP $\rightarrow$ True Positive; FP $\rightarrow$ False Positive; $Z = t \rightarrow$ Proportion of choosing the true model; $Z \supset t \rightarrow$ Proportion of the times true model is included in the chosen model; $Z_4 = t \rightarrow$ Proportion of times the chosen model of size $p_1 = 4$ is the true model.

(a) $n = 100, p = 50$; $\rho_1 = \rho_2 = \rho_3 = 0$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	3.77	0.16	0.70	0.78	0.86
Skinny Gibbs	3.87	0.36	0.64	0.86	0.88
EBIC	3.55	0.20	0.58	0.75	0.91
BSR	3.57	0.15	0.54	0.62	0.79
Alasso	3.93	3.00	0.10	0.93	0.79
SCAD	3.90	2.38	0.13	0.90	0.69
MCP	3.94	4.96	0.00	0.94	0.79
(b) $n = 100, p = 50$; $\rho_1 = \rho_2 = \rho_3 = 0.25$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	3.39	0.22	0.40	0.47	0.58
Skinny Gibbs	3.54	0.50	0.34	0.58	0.58
EBIC	3.17	0.28	0.39	0.57	0.67
BSR	3.02	0.13	0.26	0.31	0.56
Alasso	3.80	3.07	0.07	0.80	0.53
SCAD	3.68	2.85	0.05	0.68	0.32
MCP	3.83	4.32	0.01	0.84	0.45
(c) $n = 100, p = 250$ $\rho_1 = \rho_2 = \rho_3 = 0$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	3.59	0.49	0.43	0.63	0.64
Skinny Gibbs	3.64	1.19	0.26	0.67	0.61
EBIC	2.02	0.03	0.18	0.20	0.84
BSR	3.17	0.23	0.25	0.33	0.58
Alasso	3.76	4.00	0.01	0.78	0.48
SCAD	3.72	3.26	0.02	0.74	0.35
MCP	3.84	4.90	0.00	0.85	0.49
(d) $n = 100, p = 250$; $\rho_1 = \rho_2 = \rho_3 = 0.25$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	2.92	0.84	0.16	0.22	0.26
Skinny Gibbs	2.92	1.40	0.12	0.24	0.23
EBIC	1.54	0.02	0.03	0.04	0.44
BSR	2.59	0.15	0.05	0.05	0.23
Alasso	3.43	4.04	0.01	0.52	0.23
SCAD	3.15	3.30	0.04	0.34	0.12
MCP	3.58	5.26	0.01	0.64	0.20

Table 3.2: Simulation results with high correlations among predictors: TP $\rightarrow$ True Positive; FP $\rightarrow$ False Positive; $Z = t \rightarrow$ Proportion of choosing the true model; $Z \supset t$ $\rightarrow$ Proportion of the times true model is included in the chosen model; $Z_4 = t \rightarrow$ Proportion of times the chosen model of size $p_1 = 4$ is the true model.

(a) $n = 100, p = 50$; $\rho_1 = 0.10, \rho_2 = 0.25, \rho_3 = 0.50$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	3.51	0.15	0.48	0.55	0.72
Skinny Gibbs	3.65	0.32	0.49	0.66	0.71
EBIC	3.01	0.51	0.26	0.57	0.55
BSR	3.31	0.22	0.37	0.45	0.64
Alasso	3.88	3.18	0.09	0.89	0.52
SCAD	3.86	2.79	0.04	0.89	0.14
MCP	3.93	4.36	0.00	0.93	0.34
(b) $n = 100, p = 50$; $\rho_1 = 0.2, \rho_2 = 0.4, \rho_3 = 0.60$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	3.07	0.29	0.25	0.29	0.43
Skinny Gibbs	3.22	0.51	0.26	0.36	0.43
EBIC	2.60	0.72	0.08	0.40	0.23
BSR	2.56	0.29	0.10	0.13	0.27
Alasso	3.67	3.43	0.04	0.71	0.33
SCAD	3.57	3.18	0.03	0.75	0.03
MCP	3.79	4.15	0.01	0.81	0.10
(c) $n = 100, p = 250$; $\rho_1 = 0.10, \rho_2 = 0.25, \rho_3 = 0.50$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	3.31	0.99	0.31	0.46	0.44
Skinny Gibbs	3.41	1.17	0.24	0.52	0.44
EBIC	1.70	0.03	0.09	0.11	0.28
BSR	2.70	0.44	0.18	0.21	0.44
Alasso	3.66	4.03	0.02	0.70	0.30
SCAD	3.40	3.33	0.01	0.56	0.03
MCP	3.68	4.18	0.01	0.71	0.11
(d) $n = 100, p = 250$; $\rho_1 = 0.2, \rho_2 = 0.4, \rho_3 = 0.60$					
	TP	FP	$Z = t$	$Z \supset t$	$Z_4 = t$
BASAD	2.69	0.98	0.12	0.19	0.18
Skinny Gibbs	2.75	1.43	0.08	0.21	0.21
EBIC	1.31	0.04	0.02	0.03	0.08
BSR	1.47	0.80	0.01	0.01	0.03
Alasso	3.25	4.23	0.02	0.42	0.15
SCAD	3.18	3.52	0.00	0.56	0.00
MCP	3.46	4.20	0.00	0.56	0.02

Figure 3.1: Proportion of True Covariates included versus Model Size under the same settings of Table 3.2. The two curves that stay consistently on the top correspond to Skinny Gibbs (SG) and Adaptive Lasso (AL).



3.5 Real data examples

3.5.1 PCR dataset

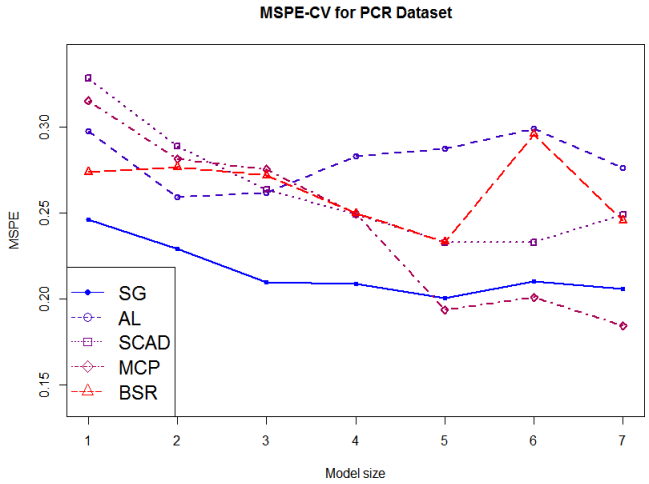
We consider the data from an experiment by Lan et al. (2006) to study the genetics of two inbred mouse populations B6 and BTBR. The data include expression levels of 22,575 genes from 31 female and 29 male mice, resulting in a total of 60 arrays. The physiological phenotype glycerol-3-phosphate acyltransferase (GPAT) was also measured by quantitative real-time PCR. The gene expression data and the phenotypic data are publicly available at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330). It is of importance to learn which genes are associated

with low levels of GPAT as low levels of GPAT are found to diminish Hepatic Steatosis, a disease commonly caused by obesity Wendel et al. (2010). For illustration, we obtain a binary response based on the variable GPAT as $\mathbf{E} = I(GPAT < Q(0.4))$, where $Q(0.4)$ is the 0.4th quantile of GPAT. The subsequent analysis will be made on the response variable $\mathbf{E}$. Due to the very large number of genes, we first perform a screening in this example but a larger p is considered in the Lymph dataset in Subsection 3.5.2. We use p-values obtained from the simple logistic regression of the response $\mathbf{E}$ against individual genes to select 99 marginally most significant genes, which along with the gender variable form $p = 100$ covariates. We apply Skinny Gibbs along the Bayesian Subset Regression method (BSR) of Liang et al. (2013)), Lasso, SCAD and MCP for selecting the covariates. The results for Skinny Gibbs are based on a chain of length 4×10^4 obtained after a burn-in of length 2×10^4. The initial value for β is the zero vector and the initialization of Z contains ones for the $K = 10$ marginally most significant covariates. For BSR, the results are based on an MCMC chain of length 2×10^5 after a burn-in chain of length 5×10^4.

In the real data applications, we consider 10-fold cross-validated prediction errors as a measure of performance of the variable selection methods. For obtaining these cross-validated errors, we divide the data D into 10 folds $D_1, \cdots, D_{10}$. For each $D_k, \{k \in 1, 2, \cdots, 10\}$, we perform variable selection using the data from $D \setminus D_k$ to obtain predicted probabilities for the responses in D_k. The cross validation error for the fold k is defined as $CV_k = \sum_{i \in D_k} (\hat{\pi}_i - E_i)^2$, where $\hat{\pi}_i$ is the predicted probability for the ith observation. The overall CV error is $CV = \sum_{k=1}^{10} CV_k/n$, where $n = \sum_{k=1}^{10} |D_k|$.

Figure 4.1 shows the 10-fold cross validation errors for different methods given the number of covariates chosen. The X-axis represents different model sizes, and the Y-axis shows CV-errors for different methods considered. We note that Skinny Gibbs performs well along with MCP. In particular, the CV error is the smallest for Skinny Gibbs if we use smaller model sizes. In Figure 3.3, we plot the marginal

Figure 3.2: PCR Dataset: Cross Validated Prediction Error versus Model Size for several model selection methods



posterior probabilities using the entire data D for two different Gibbs chains. We see that the top three genes from both the chains are the same and have higher inclusion probabilities than the rest. This is also consistent with Figure 4.1, which shows largest decrease in CV error for the first three covarites. The Affymetrix IDs of the top genes in descending order of marginal posterior probabilities are 1432002-at, 1441569-at, and 1438936-s-at. The genes 1438936-s-at and 1438937-x-at (which is among the top five genes in both the Gibbs chains) belong to the Angiogenin gene family, which is previously found to be associated with obesity (see Imai et al. (2008), Silha et al. (2005)).

3.5.2 Lymph data

We now consider the gene expression data set considered in Hans et al. (2007), and Liang et al. (2013). The dataset contains gene expressions of $n = 148$ individuals. The response of interest is positive (high risk) or negative (low risk) status of the lymph node that is related to human breast cancer. There are 100 low risk cases and 48 high risk cases. After prescreening in Hans et al. (2007), a total of 4512 genes are selected showing a variation above the noise levels. In addition, there are two

Figure 3.3: PCR Dataset: Marginal posterior probabilities from two different chains of Skinny Gibbs. The Affymetrix IDs of the top genes are given in the legend.

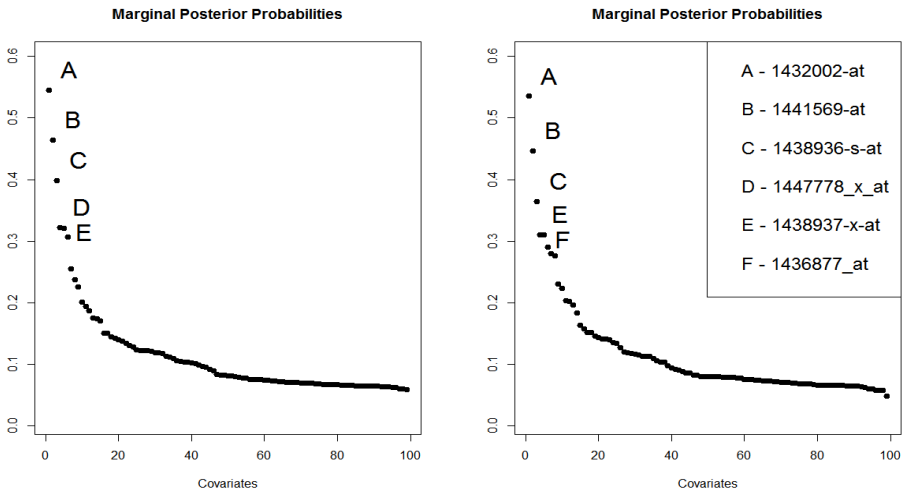
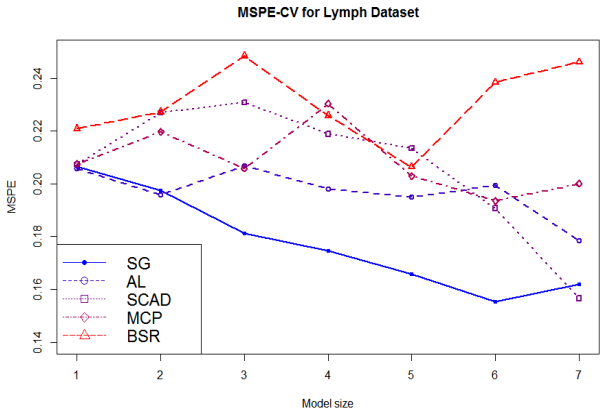


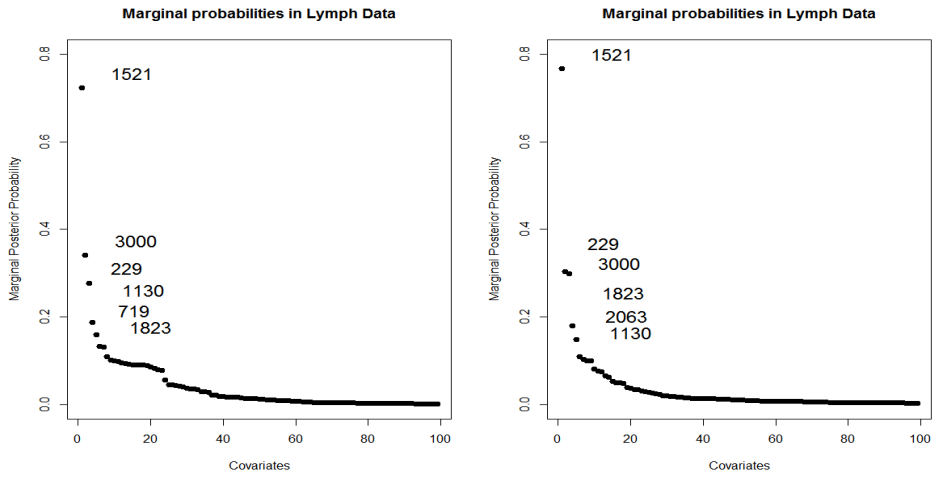
Figure 3.4: Lymph Dataset: Cross Validated Prediction Error versus Model Size for several model selection methods



clinical variables, including the tumor size in centimeters as well as the protein assay-based estrogen receptor status (coded as binary). Hence we have $p = 4514$ candidate covariates with a sample size of $n = 148$. Due to the large p in this example, the results for Skinny Gibbs are based on combining ten different chains with the length and initialization described in Subsection 3.5.1.

As in Subsection 3.5.1, we present the 10-fold cross validated prediction errors for the methods considered, see Figure 3.4. We reported the errors for the models of size smaller than or equal to seven as the larger models often lead to complete separation

Figure 3.5: Lymph Dataset: Marginal Posterior Probabilities from two different chains of Skinny Gibbs. The labels on the top six points correspond to the column numbers of the genes.



when the model is fit to the estimation data leading to unstable prediction for the testing data. All the methods considered have similar performance in terms of CV errors, with Skinny Gibbs having slightly lower errors. The CV errors from Skinny Gibbs suggest that the top six genes are important. Figure 3.5 shows the largest 100 posterior probabilities of $Z_j = 1$ from two different chains of Skinny Gibbs. The two chains lead to slightly different ordering of the top six genes, but there is only one non-overlapping gene in the two sets indicating the stability of the results. It is comforting to note that a few variables have substantially higher marginal probabilities than the rest in both the chains. However, some of the top variables do not have the marginal posterior probabilities close to 1, which can be attributed to the phenomenon that multiple sets of predictors in this problem can represent the model nearly equally well. In the supplementary materials, plots of the marginal posterior probabilities along the Skinny Gibbs iterates are provided for the two real data sets considered in this section.

3.6 Skinny Gibbs Chains

3.6.1 Simulated data settings

We shall look at the chains generated by Skinny Gibbs to check its stability and convergence. For this purpose, we used data from the simulation settings of Table 1 (c), (d) and Table 2 (c), (d) having $n = 100$; $p = 250$. We recall these correlation settings in Table 3.3.

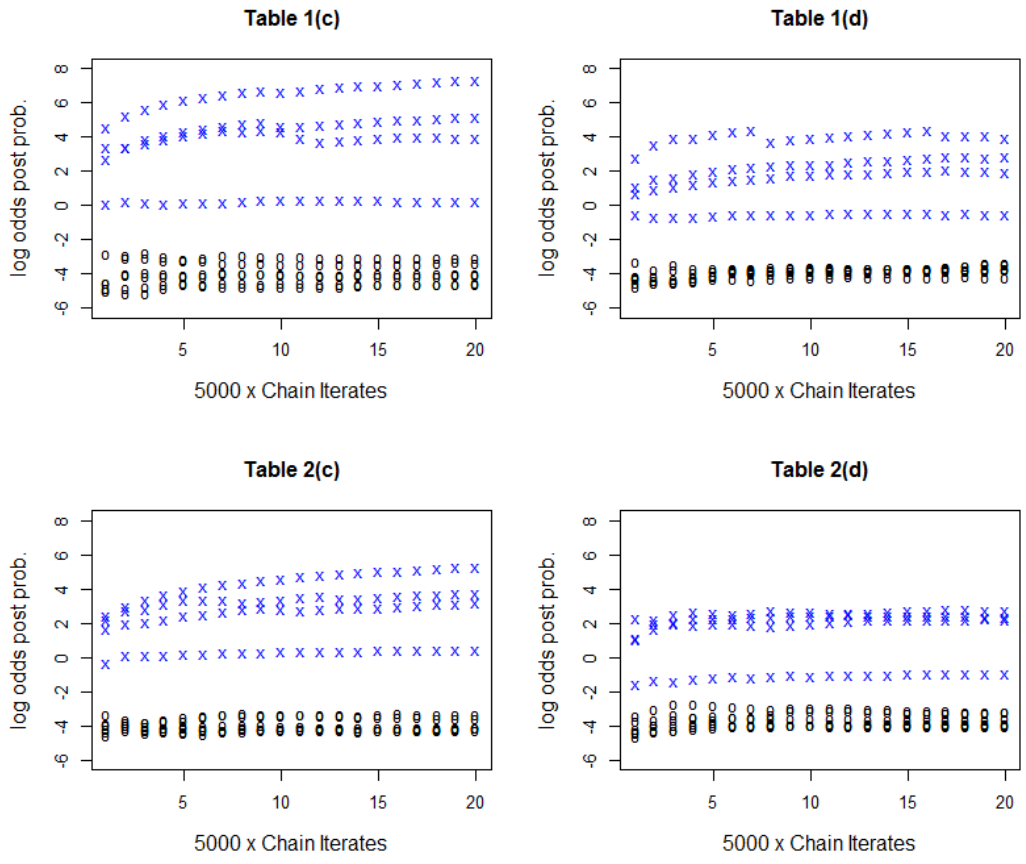
Table 3.3: Correlation settings for the chains in Figure 3.6. ρ_1: correlation between a pair of active covariates; ρ_2: correlation between a pair of active and inactive covariates; ρ_3: correlation between a pair of inactive covariates.

	ρ_1	ρ_2	ρ_3
Table 1 (c)	0.00	0.00	0.00
Table 1 (d)	0.25	0.25	0.25
Table 2 (c)	0.10	0.25	0.50
Table 2 (d)	0.20	0.40	0.60

For obtaining the Skinny Gibbs chains for each setting, we start from the null model (all Z_j are 0) and obtain marginal posterior probabilities computed at regular intervals of 5000 along the Skinny Gibbs chain. In Figure 3.6, the marginal posterior probabilities (averaged based on 10 datasets) of the four active covariates and a few inactive covarites are plotted along the chain.

It can be seen from Figure 3.6 that the Gibbs chains are stable and behaves well in general even in the settings with high correlations between covariates. The magnitudes of the posterior probabilities for the four active ones correspond to the magnitudes of their true coeffiecients. We note that the active coefficient with the least magnitude is difficult to identify explaining why the average number of true positives is closer to 3 than to 4 (Table 3.2).

Figure 3.6: Log odds of the posterior probabilities along the Skinny Gibbs chains for $n = 100; p = 250$ and different settings described in Table 3.3. The chains for the active variables are labelled with 'x' and those for inactive ones are labelled with 'o'.

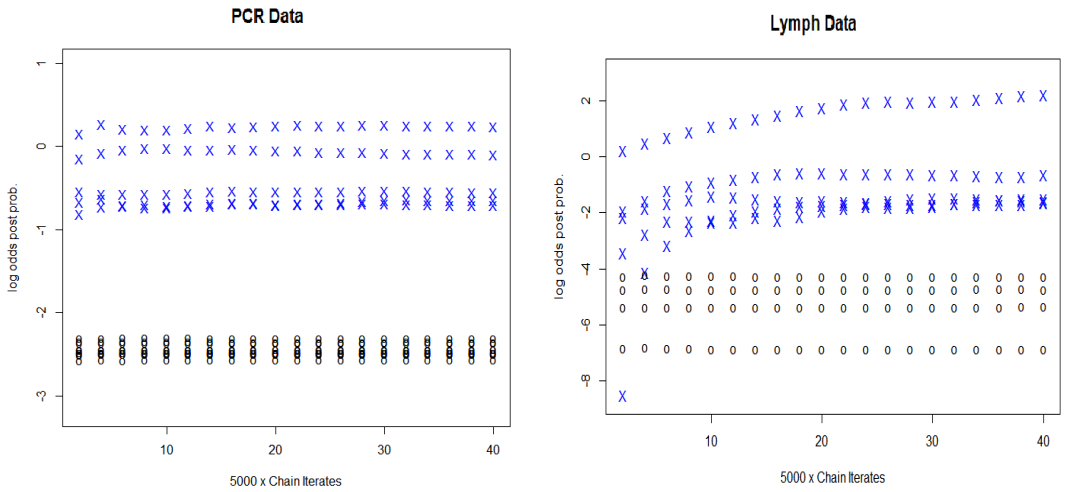


3.6.2 Real data examples

We now consider the Skinny Gibbs chains for PCR and Lymph data examples considered in Section 3.5 of the paper. We start with the null model and obtain the marginal posterior probabilities at multiples of 5000 iterations as before In Figure 3.7, we plot the log odds of the marginal posterior probabilities (averaged using 10 different chains) for some of the covariates. These covariates include five covariates having the highest marginal posterior probabilities which are labelled with 'x' and a few other covariates having low marginal posterior probabilities which are labelled with 'o'. It can be seen that the marginal posterior probabilities are stable overall

but the chain for Lymph data set takes more number of iterations to stabilize. This is reasonable given the size of the data set $(n, p) = (148, 4514)$. Nevertheless, the ordering of these variables remains mostly the same irrespective of where we stop the chain.

Figure 3.7: Log odds of the marginal posterior probabilities along the Skinny Gibbs chains for PCR data and Lymph data examples



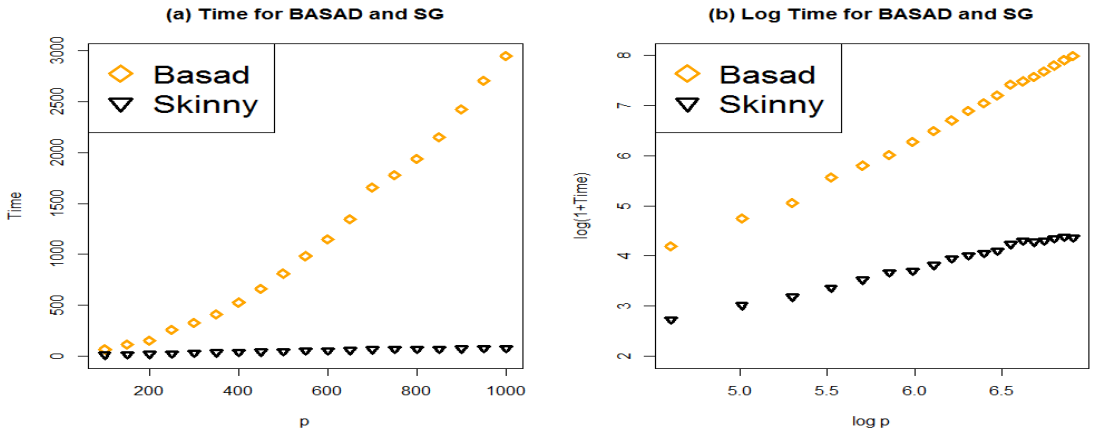
3.7 Time Improvement of Skinny Gibbs

We perform a small study for time comparison between BASAD and Skinny Gibbs. We present the CPU time for different methods based on 10 data sets with sample size $n = 100$ and varying number of variables as $p = 50, 100, \ldots, 1000$. We generate data using one of our simulation settings (as in Table 1 of the paper with $\rho_1 = \rho_2 = \rho_3 = 0$). That is, we generate each row of X independently from a normal distribution with a p-dimensional identity matrix I_p. Given X, we sample E from a logistic model $P(E_i = 1|x_i) = e^{x_i\beta}/(1 + e^{x_i\beta})$, for $i = 1, \ldots, n$. We use $p_1 = 4$, and $\beta = (1.5, 2, 2.5, 3, 0, 0, \cdots, 0)$. We use thinkpad s230u Twist with Intel(R) Core(TM) i7-3537U CPU@ 2.00GHz, 8.00GB memory, and Windows7 64bit. For both the methods, we use a burn-in of size 2000 and additional iteration 5000.

In Figure 3.8, we plot the time for BASAD and Skinny Gibbs. It can be seen that

the time for BASAD grows at a higher rate than Skinny Gibbs and look quadratic in p whereas the time for Skinny grows linearly in p.

Figure 3.8: CPU time (in seconds) for BASAD and Skinny on 10 data sets with $n = 100$ and p varies. (a) shows Time as a function of p, and (b) shows log (1+Time) as a function of $\log p$.



3.8 Proofs

Proof of Theorem 3.3.1:

We prove the theorem by checking that the conditionals corresponding to the posterior in Equation (3.16) are the same as those of Skinny Gibbs. Then the posterior of Z can be computed by integrating out the other variables, i.e., W,Y, and β. The conditional distribution of β under (3.16) is given by

$$\begin{aligned} &P(\beta \mid W, Y, Z = k) \\ &\propto \exp\left\{-\tfrac{1}{2}\left((\beta(k) - \tilde{\beta}(k))' V_{k1} (\beta(k) - \tilde{\beta}(k)) + \beta(k^c)' V_{k0} \beta(k^c)\right)\right\}, \end{aligned} \tag{3.23}$$

where $V_{k1} = (X_k' W X_k + \tau_{1n}^{-2} I)$, $\tilde{\beta}(k) = V_{k1}^{-1} X_k' W Y$, and $V_{k0} = (n + \tau_{0n}^{-2}) I$. Now, the conditional distribution of Z under (3.16) is given by

$$\begin{aligned} &P(Z = k \mid \beta, W, Y) \\ &\propto \exp\left\{-\tfrac{1}{2}\left(\beta(k)' V_{k1} \beta(k) - 2\beta(k)' X_k' W Y + \beta(k^c)' V_{k0} \beta(k^c)\right)\right\} v_n^{-|k|}, \end{aligned} \tag{3.24}$$

where $v_n = (1-q_n)\tau_{1n}/(q_n\tau_{0n})$. Furthermore, the conditionals of each Z_j based on (3.45) can be derived as:

$$R := \frac{P(Z_j = 1 \mid \beta, Z_{-j} = u, W, Y)}{P(Z_j = 0 \mid \beta, Z_{-j} = u, W, Y)} = \frac{P(Z_j = 1, Z_{-j} = u \mid \beta, W, Y)}{P(Z_j = 0, Z_{-j} = u \mid \beta, W, Y)}.$$

where Z_{-j} represents the components of Z excluding Z_j. Denote the model corresponding to $(Z_{-j} = u, Z_j = 0)$ by u. Then, due to (3.45), we have

$$\begin{aligned}
&-2\log R - 2\log v_n \\
&= (\beta(u)', \beta_j) \begin{bmatrix} \tau_{1n}^{-2} I + X_u' W X_u & X_u' W X_j \\ X_j' W X_u & \tau_{1n}^{-2} I + X_j' W X_j \end{bmatrix} \begin{pmatrix} \beta(u) \\ \beta_j \end{pmatrix} \\
&\quad -\beta(u)'(\tau_{1n}^{-2} I + X_u' W X_u)\beta(u) - (n + \tau_{0n}^{-2})\beta_j^2 \\
&\quad -2\left(\beta(u)' X_u' + \beta_j X_j'\right) WY + 2\beta(u)' X_u' WY \\
&= 2(\beta(u)' X_u' - Y') W X_j \beta_j + (\tau_{1n}^{-2} I + X_j' W X_j)\beta_j^2 - (n + \tau_{0n}^{-2})\beta_j^2 \\
&= 2(\beta(u)' X_u' - Y') W X_j \beta_j + X_j'(W - I) X_j \beta_j^2 + (\tau_{1n}^{-2} I - \tau_{0n}^{-2})\beta_j^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
R &= \frac{P(Z_j = 1 \mid Y, W, \beta, Z_{-j})}{P(Z_j = 0 \mid Y, W, \beta, Z_{-j})} \\
&= \exp\left\{-\tfrac{1}{2}\left(2(\beta(u)' X_u' - Y') W X_j \beta_j + X_j'(W - I) X_j \beta_j^2 + (\tau_{1n}^{-2} I - \tau_{0n}^{-2})\beta_j^2\right)\right\} v_n^{-1} \\
&= \frac{q_n \phi(\beta_j, 0, \tau_{1,n}^2)}{(1-q_n)\phi(\beta_j, 0, \tau_{0,n}^2)} \exp\left\{(Y' - \beta(u)' X_u')\, W X_j \beta_j + \frac{1}{2} X_j'(I - W) X_j \beta_j^2\right\}.
\end{aligned}$$

From (3.16), the conditionals of W and Y are clearly the same as those of Skinny Gibbs, which proves the theorem. □

Proof of Theorem 3.3.2:

We first define the ratio of posterior of models k and t as

$$PR(k, t) = \frac{P[Z = k \mid \mathbf{E}]}{P[Z = t \mid \mathbf{E}]}.$$

We shall prove the theorem by showing that $\sum_{k \neq t; |k| \leq m_n} PR(k,t) \xrightarrow{\text{P}} 1$. For a given $K > 1$ (to be chosen later), we divide the set of candidate models into

1. Over-fitted models: $M_1 = \{k : k \supset t, k \neq t, |k| \leq m_n\}$, i.e., the models of dimension smaller than m_n which include all the active covariates plus one or more inactive covariates.

2. Large models: $M_2 = \{k : K|t| < |k| \leq m_n\}$, the models with dimension greater than $K|t|$ but smaller than m_n.

3. Under-fitted models: $M_3 = \{k : k \not\supset t, |k| \leq K|t|\}$, the models of moderate dimension which miss an active covariate.

We shall prove that $\sum_{k \in M_u} PR(k,t) \xrightarrow{\text{P}} 1$ for $u = 1, 2, 3$.

Some preliminaries

We use the following additional notations. For any model k, $\hat{\beta}(k)$ denotes the maximum likelihood estimator (MLE) of $\beta(k)$ under the model k. Recall that $\beta_0(t)$ denotes the true regression vector (defined in Condition 3.3.3). For any model $k \supset t$, we use $\beta_0(k)$, to denote the $|k| \times 1$ vector including $\beta_0(t)$ for t and zeroes for $k \cap t^c$. We first prove the following lemma, which would be useful for the rest of the proof. We use c, c', c^* as generic constants that can take different values depending on the context.

Lemma 3.8.1. *Let $c > 0$ be any fixed constant. Under Conditions 3.3.1– 3.3.4, there exists $\epsilon_n \to 0$ such that*

$$(1 - \epsilon_n) H_n(\beta_0(s)) \leq H_n(\beta(s)) \leq (1 + \epsilon_n) H_n(\beta_0(s)), \tag{3.25}$$

for any model $s \in M_1$, and for all $\beta(s)$ such that $\|\beta(s) - \beta_0(s)\| \leq \sqrt{c|s|\Lambda_{|s|} \log p_n / n}$,

where

$$\Lambda_m := \max_{k:|k|\leq m} \lambda_{max}\left(n^{-1}X_k'X_k\right). \tag{3.26}$$

Proof: Recall that $H_n(\beta(s)) = X_s'\Sigma\left(\beta(s)\right)X_s$. Therefore, to prove the lemma, it is sufficient to show that

$$(1-\epsilon_n)\sigma_i^2(\beta_0(s)) \leq \sigma_i^2(\beta(s)) \leq (1+\epsilon_n)\sigma_i^2(\beta_0(s)),$$

for each $i = 1, \cdots, n$. By the fact that $(1+e^a)/(1+e^b) \leq e^{|a-b|}$, we have

$$\begin{aligned}
\sigma_i^2(\beta(s))\sigma_i^{-2}(\beta_0(s)) &= \frac{\exp\left\{x_i(\beta(s)-\beta_0(s))\right\}(1+e^{x_i\beta_0(s)})^2}{(1+e^{x_i\beta(s)})^2}, \\
&\leq \exp\left\{3|x_i(\beta(s)-\beta_0(s))|\right\} \\
&\to 1, \text{ as } n\to\infty.
\end{aligned}$$

because $u_n = |x_i(\beta(s)-\beta_0(s))| \leq \|x_i\|\|\beta(s)-\beta_0(s)\| \leq C\sqrt{c|s|^2\Lambda_{|s|}\log p_n/n} \preceq \sqrt{m_n^2\Lambda_{|s|}\log p_n/n} = o(1)$ by Condition 3.3.2. By interchanging $\sigma_i^2(\beta(s))$ and $\sigma_i^2(\beta_0(s))$, we would obtain the reverse inequality. □

Remark 11. Since we have $\epsilon_n \to 0$, we henceforth denote it by ϵ and treat it to be small enough.

For proving Theorem 3.3.2, we require deviation bounds of quadratic forms involving the logistic response vector E. We obtain them by using the following inequality for subgaussian random vectors.

Theorem 3.8.1 (Hsu, Kakade and Zhang (2012))**.** *Suppose $U = (U_1, \cdots, U_n)$ is a random vector such that for some $\sigma > 0$,*

$$\mathbb{E}\left[\exp(\alpha'U)\right] \leq \exp\left\{\frac{1}{2}\|\alpha\|^2\sigma^2\right\}, \tag{3.27}$$

for all $\alpha \in R^n$. Then, for any positive semidefinite matrix Q, we have

$$P\left[U'QU > \sigma^2(tr(Q) + 2\sqrt{tr(Q^2)c} + 2\|Q\|c)\right] \leq e^{-c},$$

where $tr(\cdot)$ denotes the trace of the matrix argument.

Proof: We refer to Theorem 2.1 of Hsu et al. (2012).

We apply the above theorem for $U = \mathbf{E} - \mu$. Let $\theta_i = \log(\mu_i) - \log(1 - \mu_i)$, which implies $\mu_i = e^{\theta_i}/(1 + e^{\theta_i})$. Also, define $b(\theta) = \log(1 + e^{\theta})$, which implies $b'(\theta_i) = \mu_i$ and $b''(\theta_i) = \sigma_i^2$. To check that the subgaussian inequality (3.51) holds, note that

$$\begin{aligned} \mathbb{E}\left[\exp\left\{\alpha'(\mathbf{E} - \mu)\right\}\right] &= \exp\left\{\sum_{i=1}^{n}[b(\theta_i + \alpha_i) - b(\theta_i) - \alpha_i\mu_i]\right\} \\ &= \exp\left[\tfrac{1}{2}\sum_{i=1}^{n}\alpha_i^2 b''(\theta_i + \tilde{\alpha}_i)\right] \\ &\leq \exp\left[\tfrac{1}{8}\sum_{i=1}^{n}\alpha_i^2\right], \end{aligned} \tag{3.28}$$

where $|\tilde{\alpha}_i| \leq |\alpha_i|$ and $b''(\cdot) = \mu(\cdot)(1 - \mu(\cdot)) \leq 1/4$. Therefore, (3.51) holds with $\sigma^2 = 1/4$. The following lemma provides an inequality for quadratic forms involving the projection matrices onto the column space of (scaled) design matrices.

Lemma 3.8.2. *Let $\tilde{U} = \Sigma^{-1/2}(\mathbf{E} - \mu)$ and P_k be the projection matrix onto the column space of $\Sigma^{1/2}X_k$, where k is such that $|k| \leq m_n$. Then, we have*

$$P\left[\tilde{U}'P_k\tilde{U} > (1 + \delta^*)(tr(P_k) + 2\sqrt{tr(P_k)t} + 2t)\right] \leq e^{-t},$$

for δ^* defined in Condition 3.3.2.

Proof: The proof is similar to that for Theorem 3.8.2 in Hsu et al. (2012), using Condition 3.3.2 (c). □

Lemma 3.8.3. *Under Conditions 3.3.1– 3.3.4, we have*

$$\sup_{k\supset t:|k|=m}\left\|\hat{\beta}(k)-\beta_0(k)\right\|=O_P\left(\sqrt{\frac{m\Lambda_m\log p_n}{n}}\right),$$

uniformly for all $m\le m_n$, where Λ_m is as defined in (3.50).

Proof: Let $\beta(k)=\beta_0(k)+c_nu$, where $u\in R^{|k|}$ and $u'u=1$, $c_n=\sqrt{\frac{5m\Lambda_m\log p_n}{n\lambda^2(1-\epsilon)^2}}$ and $m=|k|$. Then, for some $\tilde{\beta}(k)$ such that $\|\tilde{\beta}(k)-\beta_0(k)\|\le c_n$, we have

$$\begin{aligned}
&L_n(\beta(k))-L_n(\beta_0(k))\\
&=(\beta(k)-\beta_0(k))'s_n(\beta_0(k))-\tfrac{1}{2}(\beta(k)-\beta_0(k))'H_n(\tilde{\beta}(k))(\beta(k)-\beta_0(k))\\
&=c_nu's_n(\beta_0(k))-\tfrac{1}{2}c_n^2u'H_n(\tilde{\beta}(k))u\\
&\le c_nu's_n(\beta_0(k))-\tfrac{1}{2}c_n^2(1-\epsilon)n\lambda,
\end{aligned}$$

due to Lemma 3.8.4 and Condition 3.3.2. We obtain

$$\begin{aligned}
&P[L_n(\beta(k))-L_n(\beta_0(k))>0 \text{ for some } u]\\
&\quad\le P\left[u's_n(\beta_0(k))\ge\tfrac{1-\epsilon}{2}c_nn\lambda \text{ for some } u\right]\\
&\quad\le P\left[\|s_n(\beta_0(k))\|\ge\tfrac{1}{2}\sqrt{5m\Lambda_mn\log p_n}\right]\\
&\quad=P\left[\|X_k'(E-\mu)\|\ge\tfrac{1}{2}\sqrt{5m\Lambda_mn\log p_n}\right]\\
&\quad\le\exp\{-2m\log p_n\}=p_n^{-2m},
\end{aligned}$$

where we used that $s_n(\beta_0(k))=X_k'(\mathbf{E}-\mu)$, and applied Theorem 3.8.2 to the quadratic form $(\mathbf{E}-\mu)'X_kX_k'(\mathbf{E}-\mu)$. This implies that with probability at least $1-p_n^{-2m}$, we have $L_n(\beta(k))-L_n(\beta_0(k))<0$. The concavity of L_n implies that $\left\|\hat{\beta}(k)-\beta_0(k)\right\|\le c_n$ with probability at least $1-p_n^{-2m}$. By taking a union bound over all models $k\supset t$ with size at most m_n, we have

$$P\left[\sup_{k\supset t:|k|=m}\left\|\hat{\beta}(k)-\beta_0(k)\right\|>c_n,\ \text{for any } m\le m_n\right]\le\sum_{|t|\le m\le m_n}p_n^{-2m}p_n^m\to 0,$$

which proves the lemma. □

We are now ready to prove that $\sum_{k\in M_u} PR(k,t) \xrightarrow{\text{P}} 1$ for $u = 1, 2, 3$. Using the joint posterior given by Theorem 3.3.1, we obtain the posterior of Z by integrating out the other variables. That is,

$$P(Z = k \mid \mathbf{E}) = C^* Q_{n,k} \int_{\beta(k)} \exp\{L_n(\beta(k))\} \exp\left\{-\tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k)\right\} d\beta(k), \tag{3.29}$$

where $Q_{n,k} = (n + \tau_{0n}^{-2})^{|k|/2}\ v_n^{-|k|}$ and C^* is the normalizing constant.

For Overfitted models:

By Taylor's expansion of $L_n(\beta(k))$ around $\hat{\beta}(k)$ (the MLE of $\beta(k)$ under model k), we have

$$L_n(\beta(k)) = L_n(\hat{\beta}(k)) - \frac{1}{2}(\beta(k) - \hat{\beta}(k))' H_n(\tilde{\beta}(k))(\beta(k) - \hat{\beta}(k)), \tag{3.30}$$

for some $\tilde{\beta}(k)$ such that $\|\tilde{\beta}(k) - \hat{\beta}(k)\| \le \|\beta(k) - \hat{\beta}(k)\|$. Note that $\tilde{\beta}(k)$ may depend on $\beta(k)$. However, due to Lemmas 3.8.4 & 3.8.6, for any $k \in M_1$, we have

$$L_n(\beta(k)) - L_n(\hat{\beta}(k)) \le -\frac{(1-\epsilon)}{2}(\beta(k) - \hat{\beta}(k))' H_n(\beta_0(k))(\beta(k) - \hat{\beta}(k)),$$

for all $\beta(k)$ such that $\|\beta(k) - \beta_0(k)\| < c\sqrt{|k|\Lambda_{|k|}\log p_n/n} := cw_n$. Note that for $\beta(k)$ such that $\|\beta(k) - \hat{\beta}(k)\| = cw_n/2$,

$$L_n(\beta(k)) - L_n(\hat{\beta}(k)) \le -(1-\epsilon)n\lambda c^2 w_n^2/4 = -(1-\epsilon)c^2\lambda|k|\Lambda_{|k|}\log p_n/4 \to -\infty. \tag{3.31}$$

By concavity of $L_n(\cdot)$ and the fact that $\hat{\beta}(k)$ maximizes $L_n(\beta(k))$, (3.56) also holds for any $\|\beta(k) - \hat{\beta}(k)\| > cw_n/2$. Now due to Lemma 3.8.6, we have $B := \{\beta(k) : \|\beta(k) - \hat{\beta}(k)\| \le cw_n/2\} \subset \{\beta(k) : \|\beta(k) - \beta_0(k)\| \le cw_n\}$ with probability going to

one uniformly in M_1 (for c large enough). Therefore, we have for any $k \subset M_1$,

$$
\begin{aligned}
&P(Z = k \mid \mathbf{E}) \\
&\leq C^* \; Q_{n,k} \exp\{L_n(\hat{\beta}(k))\} \\
&\times \Big(\int_B \exp\left\{ -\tfrac{1}{2}(1-\epsilon)(\beta(k)-\hat{\beta}(k))' H_n(\beta_0(k))(\beta(k)-\hat{\beta}(k)) - \tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k) \right\} d\beta(k) \\
&\qquad + \exp(-c^2(1-\epsilon)\lambda\Lambda_{|k|}|k| \log p_n/4) \int_{B^c} \exp\{-\tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k)\} d\beta(k) \Big) \\
&\leq w' \; Q_{n,k} \;\; \exp\{L_n(\hat{\beta}(k))\} \; \tau_{1n}^{|k|} \;\; |H_n(\beta_0(k))(1-\epsilon)\tau_{1n}^2 + I|^{-1/2}(1+o(1)),
\end{aligned}
\tag{3.32}
$$

where w' is a constant. We used that for $A = (1-\epsilon)H_n(\beta_0(k))$,

$$
\begin{aligned}
&\int_B \exp\left\{ -\tfrac{1}{2}(\beta(k)-\hat{\beta}(k))' A (\beta(k)-\hat{\beta}(k)) - \tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k) \right\} d\beta(k) \\
&= \int_B \exp\left\{ -\tfrac{1}{2}\left((\beta(k)-\beta^*(k))'(A+\tau_{1n}^{-2}I)(\beta(k)-\beta^*(k)) \right) \right\} d\beta(k) \\
&\quad \times \exp\left\{ -\tfrac{1}{2}\hat{\beta}(k)' \left(A - A(A+\tau_{1n}^{-2})^{-1}A \right) \hat{\beta}(k) \right\} \\
&\leq w' |H_n(\beta_0(k))(1-\epsilon) + \tau_{1n}^{-2}I|^{-1/2} \exp\left\{ -\tfrac{1}{2}\hat{\beta}(k)' \left(A - A(A+\tau_{1n}^{-2})^{-1}A \right) \hat{\beta}(k) \right\} \\
&\leq w' |A + \tau_{1n}^{-2}I|^{-1/2},
\end{aligned}
$$

where $\beta^*(k) = (A+\tau_{1n}^{-2}I)^{-1}A\hat{\beta}(k)$. Following similar arguments, we obtain a lower bound for $P[Z = t \mid \mathbf{E}]$, i.e.,

$$
P(Z = t \mid \mathbf{E}) \geq w \;\; \exp\{L_n(\hat{\beta}(t))\} Q_{n,t}(\tau_{1n})^{|t|} |H_n(\beta_0(t))(1+\epsilon)\tau_{1n}^2 + I|^{-1/2}, \tag{3.33}
$$

for some constant w. Therefore, we have

$$
\begin{aligned}
&PR(k,t) \\
&\preceq \;\; \tfrac{Q_{n,k}}{Q_{n,t}} |H_n(\beta_0(t))(1+\epsilon) + \tau_{1n}^{-2}|^{1/2} |H_n(\beta_0(k)(1-\epsilon) + \tau_{1n}^{-2}|^{-1/2} \\
&\quad \times \exp\left\{ L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \right\} \\
&\preceq q_n^{(|k|-|t|)} (n\tau_{1n}^2)^{-(|k|-|t|)/2} (n\tau_{0n}^2+1)^{(|k|-|t|)/2} \exp\left\{ L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \right\} \\
&\preceq p_n^{-(2+\delta)(|k|-|t|)} \exp\left\{ L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \right\},
\end{aligned}
\tag{3.34}
$$

where δ is from Condition 3.3.4. By applying Taylor's expansion for $L_n(\beta(k))$ about $\beta_0(k)$ and evaluating at $\hat{\beta}(k)$, we would obtain

$$\begin{aligned} L_n(\hat{\beta}(k)) &= L_n(\beta_0(k)) + (\hat{\beta}(k) - \beta_0(k))' s_n(\beta_0(k)) \\ &\quad -\tfrac{1}{2}(\hat{\beta}(k) - \beta_0(k))' H_n(\tilde{\beta}(k))(\hat{\beta}(k) - \beta_0(k)), \end{aligned} \tag{3.35}$$

such that $\|\tilde{\beta}(k) - \beta_0(k)\| \leq \|\hat{\beta}(k) - \beta_0(k)\|$. Then due to Lemmas 3.8.4 & 3.8.6,

$$\begin{aligned} & L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \\ & \leq L_n(\hat{\beta}(k)) - L_n(\beta_0(k)) \\ & \leq (\hat{\beta}(k) - \beta_0(k))' s_n(\beta_0(k)) - \tfrac{1-\epsilon}{2}(\hat{\beta}(k) - \beta_0(k))' H_n(\beta_0(k))(\hat{\beta}(k) - \beta_0(k)) \\ & \leq \tfrac{1}{2(1-\epsilon)} s_n(\beta_0(k))' H_n(\beta_0(k))^{-1} s_n(\beta_0(k)) \\ & = \tfrac{1}{2(1-\epsilon)} (\mathbf{E} - \mu)' X_k H_n(\beta_0(k))^{-1} X_k' (\mathbf{E} - \mu) \\ & = \tfrac{1}{2(1-\epsilon)} \tilde{U}' P_k \tilde{U}, \end{aligned} \tag{3.36}$$

where $\Sigma = Diag(\sigma_1^2, \cdots, \sigma_n^2)$, $\tilde{U} = \Sigma^{-1/2}(\mathbf{E} - \mu)$, and $P_k = \Sigma^{1/2} X_k H_n(\beta_0(k))^{-1} X_k' \Sigma^{1/2}$ is the projection matrix onto the column space of $\Sigma^{1/2} X_k$.

We now use Lemma 3.8.5 to obtain a probability bound on the quadratic form obtained in Equation (3.61). Consider $b_n = (1 + \delta^*)(1 + 2w) \log p_n$, where w is small such that $(1 + \delta^*)(1 + 2w) < (1 + \delta)$ (this is possible due to Condition 3.3.4) and ϵ is small such that $\psi = (1 - \epsilon)(1 + w) - 1 > 0$. Then, we have

$$\begin{aligned} & P\left[\left|L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))\right| > b_n(|k| - |t|)\right] \\ & \leq P\left[\tilde{U}' P_k \tilde{U} > 2(1 - \epsilon) b_n(|k| - |t|)\right] \\ & \leq \exp\left\{-(1 - \epsilon)(1 + w)(|k| - |t|) \log p_n\right\} = p_n^{-(1+\psi)(|k|-|t|)}. \end{aligned} \tag{3.37}$$

Now, by taking a union bound we obtain

$$\begin{aligned} &P[|L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))| > b_n(|k| - |t|) \text{ for any } k \in M_1] \\ &\quad \leq \sum_{|k|=|t|+1}^{m_n} p_n^{-(1+\psi)(|k|-|t|)} p_n^{-(|k|-|t|)} \to 0. \end{aligned}$$

By restricting to the complement of the above small probability event, and using Equation (3.59), we have

$$\begin{aligned} &\sum_{k \in M_1} PR(k,t) \\ &\quad \preceq \sum_{k \in M_1} p_n^{-(2+\delta)(|k|-|t|)} \exp\{(1+\delta^*)(1+2w)(|k|-|t|)\log p_n\} \\ &\quad \preceq \sum_{d=|t|+1}^{m_n} p_n^{(d-|t|)} p_n^{-(2+\delta)(d-|t|)} \exp\{(1+\delta^*)(1+2w)(d-|t|)\log p_n\} \\ &\quad \preceq p_n^{-(1+\delta-(1+\delta^*)(1+2w))} = o_P(1). \end{aligned} \tag{3.38}$$

Large models

The proof for large models is similar to the overfitted ones. We only need to note that for a large underfitted model, we shall work with $k^* = k \cup t$. Similar to the previous case, consider Taylor's expansion (3.55) of $L_n(\beta(k^*))$ evaluated at $\beta(k^*) = (\beta(k) \cup 0)_{|k^*| \times 1}$. As before, we then have

$$\begin{aligned} L_n(\beta(k^*)) &= L_n(\hat{\beta}(k^*)) - \tfrac{1}{2}(\beta(k^*) - \hat{\beta}(k^*))' H_n(\tilde{\beta}(k^*))((\beta(k^*) - \hat{\beta}(k^*)) \\ &\leq L_n(\hat{\beta}(k^*)) - \tfrac{n(1-\epsilon)\lambda}{2}(\beta(k^*) - \hat{\beta}(k^*))'(\beta(k^*) - \hat{\beta}(k^*)), \end{aligned}$$

for all $\beta(k^*)$ such that $\|\beta(k^*) - \beta_0(k^*)\| < c\sqrt{|k^*|\Lambda_{|k^*|}\log p_n/n}$. Then, following the arguments in the previous case, we have

$$P(Z = k \mid \mathbf{E}) \leq w' \, Q_{n,k} \, \exp\{L_n(\hat{\beta}(k^*))\} \, \tau_{1n}^{|k|} \, (n\lambda(1-\epsilon)\tau_{1n}^2)^{-|k|/2}(1+o(1)),$$

which implies that

$$PR(k,t) \;\preceq\; p_n^{-(2+\delta)(|k|-|t|)} \exp\left\{L_n(\hat{\beta}(k^*)) - L_n(\hat{\beta}(t))\right\}. \tag{3.39}$$

Now, similar to Equation (3.62) we have

$$P\left[\cup_{k:|k|=d}\left|L_n(\hat{\beta}(k^*)) - L_n(\hat{\beta}(t))\right| > b_n(|k|-|t|)\right] \;\leq p_n^{-(1+\psi)(d-|t|)} p_n^d,$$

which implies that

$$\begin{aligned} &P[\left|L_n(\hat{\beta}(k^*)) - L_n(\hat{\beta}(t))\right| > b_n(|k|-|t|) \text{ for any } k \in M_2] \\ &\leq \sum_{d=K|t|+1}^{m_n} p_n^{-(1+\psi)(d-|t|)} p_n^d \\ &\preceq p_n^{-\psi K|t|} p_n^{(1+\psi)|t|} \to 0. \end{aligned}$$

if we take $K > (1+\psi)/\psi$, in the definition of M_2. Hence, we have the result for large models by observing that $\sum_{k\in M_2} PR(k,t) = o_P(1)$, as in the case of Equation (3.63).

Underfitted models

Now, we shall prove the same for under-fitted models, i.e., for models in M_3. From Equation (3.54), we have

$$\begin{aligned} P(Z=k \mid \mathbf{E}) \;&\leq C^* \exp\{L_n(\hat{\beta}(k))\}\; Q_{n,k} \int_{\beta(k)} \exp\left\{-\tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k)\right\} d\beta(k) \\ &\preceq \exp\{L_n(\hat{\beta}(k))\} Q_{n,k} (\tau_{1n})^{|k|}. \end{aligned} \tag{3.40}$$

We shall now show that

$$|L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))| \preceq -|t|\Lambda_{(K+1)|t|} \log p_n,$$

with probability going to one, uniformly over models in M_3. To see this, for any $k \in M_3$, let $k^* = k \cup t$. Consider $\beta(k^*)$ such that $\|\beta(k^*) - \beta_0(k^*)\| = a_n$, where $a_n = \min_{1 \le i \le |t|} |\beta_{0i}(t)| \ge \sqrt{c|t|\Lambda_{(K+1)|t|} \log p_n / n}$, for a large enough constant c due to Condition 3.3.3. Then, we have

$$\begin{aligned}
&L_n(\beta(k^*)) - L_n(\beta_0(k^*)) \\
&\le (\beta(k^*) - \beta_0(k^*))' s_n(\beta_0(k^*)) - \tfrac{1}{2}(\beta(k^*) - \beta_0(k^*))' H_n(\tilde{\beta}(k^*))(\beta(k^*) - \beta_0(k^*)) \\
&\le (\beta(k^*) - \beta_0(k^*))' s_n(\beta_0(k^*)) - \tfrac{1}{2} n\lambda \|\beta(k^*) - \beta_0(k^*)\|^2 \\
&\le a_n \|s_n(\beta_0(k^*))\| - \tfrac{1}{2} n\lambda a_n^2.
\end{aligned} \tag{3.41}$$

Next, we shall obtain a bound for $\|s_n(\beta_0(k^*))\|$. Note that, $\beta_0(k^*)$ is the true $\beta_0(t)$ appended with zeroes for $k \cap t^c$. From the proof of Lemma 3.8.6, we have $\|s_n(\beta_0(k^*))\| = O_P(\sqrt{n|t|\Lambda_{(K+1)|t|} \log p_n})$ uniformly in $k^* = \{k \cup t : k \in M_3\}$ as $|k^*|$ is bounded by $(K+1)|t|$. Therefore, from (3.66), we have

$$L_n(\beta(k^*)) - L_n(\beta_0(k^*)) \quad \le w' a_n \sqrt{n|t|\Lambda_{(K+1)|t|} \log p_n} - \tfrac{1}{2} n\lambda a_n^2 \quad \preceq -\tfrac{1}{2} n\lambda a_n^2,$$

Let $\tilde{\beta}(k^*)$ be the $|k^*| \times 1$ vector including $\hat{\beta}(k)$ for k and zeroes for $k \cap t^c$. Then, we have $L_n(\tilde{\beta}(k^*)) = L_n(\hat{\beta}(k))$ and $\|\tilde{\beta}(k^*) - \beta_0(k^*)\| \ge a_n$. By the concavity of $L_n(\cdot)$, we obtain

$$L_n(\tilde{\beta}(k^*)) - L_n(\beta_0(k^*)) \le L_n(\beta(k^*)) - L_n(\beta_0(k^*)) \preceq -\frac{1}{2} n\lambda a_n^2. \tag{3.42}$$

On the other hand, we have

$$\frac{Q_{n,k}(\tau_{1n})^{|k|}}{Q_{n,t}(\tau_{1n})^{|t|} (n\tau_{1n}^2)^{-|t|/2}} \le w p_n^{c'|t|},$$

which implies that

$$\begin{aligned}\sum_{k\in M_3} PR(k,t) &\preceq \sum_{k\in M_3} p_n^{c'|t|}\exp\{L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))\}\\ &\leq \sum_{k\in M_3} p_n^{c'|t|}\exp\{L_n(\tilde{\beta}(k^*)) - L_n(\beta_0(k^*))\}\\ &\preceq p_n^{c'|t|}p_n^{(K+1)|t|}\exp\left\{-\tfrac{1}{2}n\lambda a_n^2\right\}\\ &= \exp\left\{(K+c'+1)|t|\log p_n - \tfrac{1}{2}n\lambda a_n^2\right\} \to 0,\ as\ n\to\infty,\end{aligned}$$

with probability going to one. Therefore, we have $\sum_{k:k\neq t;|k|\leq m_n} \frac{P[Z=k|\mathbf{E}]}{P[Z=t|\mathbf{E}]} \xrightarrow{\text{P}} 0$. This implies that $\sum_{k:k\neq t;|k|\leq m_n} \frac{P[Z=k||Z|\leq m_n,\mathbf{E}]}{P[Z=t||Z|\leq m_n,\mathbf{E}]} \xrightarrow{\text{P}} 0$, which in turn implies that $P[Z = t \mid |Z| \leq m_n, \mathbf{E}] \xrightarrow{\text{P}} 1$. □

Proof of Lemma 3.8.7: Due to Theorem 5.39 of Vershynin (2012) and following the proof of Lemma 6.1 of Narisetty and He (2014), we have with probability going to one $0 < c' \leq \min_{k:|k|\leq m_n+|t|} \lambda_{min}\left(n^{-1}X_k'\Sigma_k^{-1}X_k\right)$, where Σ_k is the $|k|\times|k|$ submatrix of $\Sigma = Cov(x_i)$ corresponding to the covariates in model $|k|$. As the minimum eigenvalue of Σ_k is bounded away from zero, we also have $0 < \min_{k:|k|\leq m_n+|t|} \lambda_{min}\left(n^{-1}X_k'X_k\right)$ for some $c > 0$. Now, as $\beta_0(t)$ satisfies $P[|x_i'\beta_0(t)| \geq M] \leq w < 1$, for some $M > 0$, the set of indices $I = \{i : |X_i'\beta_0(t)| \leq M\}$ satisfies $|I| \geq n(1-w)/2$ with exponentially large probability. Then, with probability gping to one, we have

$$\begin{aligned}\lambda_{min}\left(n^{-1}H_n(\beta_0(k))\right) &= \lambda_{min}\left(n^{-1}X_k'Diag\left(\sigma_i^2(\beta_0(k))\,X_k\right)\right.\\ &\geq \lambda_{min}\left(n^{-1}X_k^{*\prime}D^*X_k^*\right)\\ &\geq d_M\lambda_{min}\left(n^{-1}X_k^{*\prime}X_k^*\right)\\ &\geq (1-w)cd_M/2 > 0,\end{aligned} \tag{3.43}$$

where X^* is the $|I|\times p$ matrix with rows from X indexed by I, $D^* = Diag\left(\sigma_i^2(\beta_0(k)) : i \in I\right)$, and $d_M = \exp\{M\}/(1+\exp\{M\})^2$. □

3.8.1 Proof of Theorem 3.3.1:

We prove the theorem by checking that the conditionals corresponding to the posterior in Equation (3.16) are the same as those of Skinny Gibbs. Then the posterior of Z can be computed by integrating out the other variables, i.e., W,Y, and β. The conditional distribution of β under (3.16) is given by

$$\begin{aligned} &P(\beta \mid W, Y, Z = k) \\ &\propto \exp\left\{-\tfrac{1}{2}\left((\beta(k) - \tilde{\beta}(k))' V_{k1} (\beta(k) - \tilde{\beta}(k)) + \beta(k^c)' V_{k0} \beta(k^c)\right)\right\}, \end{aligned} \tag{3.44}$$

where $V_{k1} = (X_k' W X_k + \tau_{1n}^{-2} I)$, $\tilde{\beta}(k) = V_{k1}^{-1} X_k' W Y$, and $V_{k0} = (n + \tau_{0n}^{-2}) I$. Now, the conditional distribution of Z under (3.16) is given by

$$\begin{aligned} &P(Z = k \mid \beta, W, Y) \\ &\propto \exp\left\{-\tfrac{1}{2}\left(\beta(k)' V_{k1} \beta(k) - 2\beta(k)' X_k' W Y + \beta(k^c)' V_{k0} \beta(k^c)\right)\right\} v_n^{-|k|}, \end{aligned} \tag{3.45}$$

where $v_n = (1 - q_n)\tau_{1n}/(q_n \tau_{0n})$. Furthermore, the conditionals of each Z_j based on (3.45) can be derived as:

$$R := \frac{P(Z_j = 1 \mid \beta, Z_{-j} = u, W, Y)}{P(Z_j = 0 \mid \beta, Z_{-j} = u, W, Y)} \quad = \quad \frac{P(Z_j = 1, Z_{-j} = u \mid \beta, W, Y)}{P(Z_j = 0, Z_{-j} = u \mid \beta, W, Y)}. \tag{3.46}$$

where Z_{-j} represents the components of Z excluding Z_j. Denote the model corresponding to $(Z_{-j} = u, Z_j = 0)$ by u. Then, due to (3.45), we have

$$\begin{aligned}
&-2\log R - 2\log v_n \\
&= (\beta(u)', \beta_j) \begin{bmatrix} \tau_{1n}^{-2} I + X_u' W X_u & X_u' W X_j \\ X_j' W X_u & \tau_{1n}^{-2} I + X_j' W X_j \end{bmatrix} \begin{pmatrix} \beta(u) \\ \beta_j \end{pmatrix} \\
&\quad -\beta(u)'(\tau_{1n}^{-2} I + X_u' W X_u)\beta(u) - (n + \tau_{0n}^{-2})\beta_j^2 \\
&\quad -2\left(\beta(u)' X_u' + \beta_j X_j'\right) WY + 2\beta(u)' X_u' WY \\
&= 2(\beta(u)' X_u' - Y')W X_j \beta_j + (\tau_{1n}^{-2} I + X_j' W X_j)\beta_j^2 - (n + \tau_{0n}^{-2})\beta_j^2 \\
&= 2(\beta(u)' X_u' - Y')W X_j \beta_j + X_j'(W - I)X_j \beta_j^2 + (\tau_{1n}^{-2} I - \tau_{0n}^{-2})\beta_j^2.
\end{aligned} \tag{3.47}$$

Therefore,

$$\begin{aligned}
R &= \frac{P(Z_j = 1 \mid Y, W, \beta, Z_{-j})}{P(Z_j = 0 \mid Y, W, \beta, Z_{-j})} \\
&= \exp\left\{-\tfrac{1}{2}\left(2(\beta(u)' X_u' - Y')W X_j \beta_j + X_j'(W - I)X_j \beta_j^2 + (\tau_{1n}^{-2} I - \tau_{0n}^{-2})\beta_j^2\right)\right\} v_n^{-1} \\
&= \frac{q_n \phi(\beta_j, 0, \tau_{1,n}^2)}{(1 - q_n)\phi(\beta_j, 0, \tau_{0,n}^2)} \exp\left\{(Y' - \beta(u)' X_u')\, W X_j \beta_j + \frac{1}{2} X_j'(I - W)X_j \beta_j^2\right\}.
\end{aligned} \tag{3.48}$$

From (3.16), the conditionals of W and Y are clearly the same as those of Skinny Gibbs, which proves the theorem. □

Proof of Theorem 3.3.2:

We first define the ratio of posterior of models k and t as

$$PR(k, t) = \frac{P[Z = k \mid \mathbf{E}]}{P[Z = t \mid \mathbf{E}]}.$$

We shall prove the theorem by showing that $\sum_{k \neq t; |k| \leq m_n} PR(k, t) \xrightarrow{\mathrm{P}} 1$. For a given $K > 1$ (to be chosen later), we divide the set of candidate models into

1. Over-fitted models: $M_1 = \{k : k \supset t, k \neq t, |k| \leq m_n\}$, i.e., the models of dimension smaller than m_n which include all the active covariates plus one or

more inactive covariates.

2. Large models: $M_2 = \{k : K|t| < |k| \leq m_n\}$, the models with dimension greater than $K|t|$ but smaller than m_n.

3. Under-fitted models: $M_3 = \{k : k \not\supset t, |k| \leq K|t|\}$, the models of moderate dimension which miss an active covariate.

We shall prove that $\sum_{k \in M_u} PR(k,t) \xrightarrow{\text{P}} 1$ for $u = 1, 2, 3$.

Some preliminaries

We use the following additional notations. For any model k, $\hat{\beta}(k)$ denotes the maximum likelihood estimator (MLE) of $\beta(k)$ under the model k. Recall that $\beta_0(t)$ denotes the true regression vector (defined in Condition 3.3.3). For any model $k \supset t$, we use $\beta_0(k)$, to denote the $|k| \times 1$ vector including $\beta_0(t)$ for t and zeroes for $k \cap t^c$. We first prove the following lemma, which would be useful for the rest of the proof. We use c, c', c^* as generic constants that can take different values depending on the context.

Lemma 3.8.4. *Let $c > 0$ be any fixed constant. Under Conditions 3.3.1– 3.3.4, there exists $\epsilon_n \to 0$ such that*

$$(1 - \epsilon_n) H_n(\beta_0(s)) \leq H_n(\beta(s)) \leq (1 + \epsilon_n) H_n(\beta_0(s)), \tag{3.49}$$

for any model $s \in M_1$, and for all $\beta(s)$ such that $\|\beta(s) - \beta_0(s)\| \leq \sqrt{c|s|\Lambda_{|s|} \log p_n / n}$, where

$$\Lambda_m := \max_{k:|k| \leq m} \lambda_{max} \left(n^{-1} X_k' X_k \right). \tag{3.50}$$

Proof: Recall that $H_n(\beta(s)) = X_s' \Sigma\left(\beta(s)\right) X_s$. Therefore, to prove the lemma, it is

sufficient to show that

$$(1-\epsilon_n)\sigma_i^2(\beta_0(s)) \leq \sigma_i^2(\beta(s)) \leq (1+\epsilon_n)\sigma_i^2(\beta_0(s)),$$

for each $i = 1, \cdots, n$. By the fact that $(1+e^a)/(1+e^b) \leq e^{|a-b|}$, we have

$$\begin{aligned}
\sigma_i^2(\beta(s))\sigma_i^{-2}(\beta_0(s)) &= \frac{\exp\{x_i(\beta(s)-\beta_0(s))\}(1+e^{x_i\beta_0(s)})^2}{(1+e^{x_i\beta(s)})^2}, \\
&\leq \exp\{3|x_i(\beta(s)-\beta_0(s))|\} \\
&\to 1, \text{ as } n \to \infty.
\end{aligned}$$

because $u_n = |x_i(\beta(s)-\beta_0(s))| \leq \|x_i\|\|\beta(s)-\beta_0(s)\| \leq C\sqrt{c|s|^2\Lambda_{|s|}\log p_n/n} \preceq \sqrt{m_n^2\Lambda_{|s|}\log p_n/n} = o(1)$ by Condition 3.3.2. By interchanging $\sigma_i^2(\beta(s))$ and $\sigma_i^2(\beta_0(s))$, we would obtain the reverse inequality. □

Remark 12. Since we have $\epsilon_n \to 0$, we henceforth denote it by ϵ and treat it to be small enough.

For proving Theorem 3.3.2, we require deviation bounds of quadratic forms involving the logistic response vector E. We obtain them by using the following inequality for subgaussian random vectors.

Theorem 3.8.2 (Hsu, Kakade and Zhang (2012)). *Suppose $U = (U_1, \cdots, U_n)$ is a random vector such that for some $\sigma > 0$,*

$$\mathbb{E}\left[\exp(\alpha' U)\right] \leq \exp\left\{\frac{1}{2}\|\alpha\|^2\sigma^2\right\}, \tag{3.51}$$

for all $\alpha \in R^n$. Then, for any positive semidefinite matrix Q, we have

$$P\left[U'QU > \sigma^2(tr(Q) + 2\sqrt{tr(Q^2)c} + 2\|Q\|c)\right] \leq e^{-c},$$

where $tr(\cdot)$ denotes the trace of the matrix argument.

Proof: We refer to Theorem 2.1 of Hsu et al. (2012).

We apply the above theorem for $U = \mathbf{E} - \mu$. Let $\theta_i = \log(\mu_i) - \log(1 - \mu_i)$, which implies $\mu_i = e^{\theta_i}/(1 + e^{\theta_i})$. Also, define $b(\theta) = \log(1 + e^{\theta})$, which implies $b'(\theta_i) = \mu_i$ and $b''(\theta_i) = \sigma_i^2$. To check that the subgaussian inequality (3.51) holds, note that

$$\begin{aligned} \mathbb{E}\left[\exp\left\{\alpha'(\mathbf{E} - \mu)\right\}\right] &= \exp\left\{\sum_{i=1}^{n}[b(\theta_i + \alpha_i) - b(\theta_i) - \alpha_i\mu_i]\right\} \\ &= \exp\left[\frac{1}{2}\sum_{i=1}^{n}\alpha_i^2 b''(\theta_i + \tilde{\alpha}_i)\right] \\ &\leq \exp\left[\frac{1}{8}\sum_{i=1}^{n}\alpha_i^2\right], \end{aligned} \tag{3.52}$$

where $|\tilde{\alpha}_i| \leq |\alpha_i|$ and $b''(\cdot) = \mu(\cdot)(1 - \mu(\cdot)) \leq 1/4$. Therefore, (3.51) holds with $\sigma^2 = 1/4$. The following lemma provides an inequality for quadratic forms involving the projection matrices onto the column space of (scaled) design matrices.

Lemma 3.8.5. *Let $\tilde{U} = \Sigma^{-1/2}(\mathbf{E} - \mu)$ and P_k be the projection matrix onto the column space of $\Sigma^{1/2}X_k$, where k is such that $|k| \leq m_n$. Then, we have*

$$P\left[\tilde{U}'P_k\tilde{U} > (1 + \delta^*)(tr(P_k) + 2\sqrt{tr(P_k)t} + 2t)\right] \leq e^{-t},$$

for δ^* defined in Condition 3.3.2.

Proof: The proof is similar to that for Theorem 3.8.2 in Hsu et al. (2012), using Condition 3.3.2 (c). □

Lemma 3.8.6. *Under Conditions 3.3.1– 3.3.4, we have*

$$\sup_{k \supset t : |k| = m} \left\| \hat{\beta}(k) - \beta_0(k) \right\| = O_P\left(\sqrt{\frac{m\Lambda_m \log p_n}{n}}\right),$$

uniformly for all $m \leq m_n$, where Λ_m is as defined in (3.50).

Proof: Let $\beta(k) = \beta_0(k) + c_n u$, where $u \in R^{|k|}$ and $u'u = 1$, $c_n = \sqrt{\frac{5m\Lambda_m \log p_n}{n\lambda^2(1-\epsilon)^2}}$ and

$m = |k|$. Then, for some $\tilde{\beta}(k)$ such that $\|\tilde{\beta}(k) - \beta_0(k)\| \leq c_n$, we have

$$\begin{aligned} & L_n(\beta(k)) - L_n(\beta_0(k)) \\ & = (\beta(k) - \beta_0(k))' s_n(\beta_0(k)) - \tfrac{1}{2}(\beta(k) - \beta_0(k))' H_n(\tilde{\beta}(k))(\beta(k) - \beta_0(k)) \\ & = c_n u' s_n(\beta_0(k)) - \tfrac{1}{2} c_n^2 u' H_n(\tilde{\beta}(k)) u \\ & \leq c_n u' s_n(\beta_0(k)) - \tfrac{1}{2} c_n^2 (1-\epsilon) n\lambda, \end{aligned} \tag{3.53}$$

due to Lemma 3.8.4 and Condition 3.3.2. We obtain

$$\begin{aligned} & P[L_n(\beta(k)) - L_n(\beta_0(k)) > 0 \text{ for some } u] \\ & \leq P\left[u' s_n(\beta_0(k)) \geq \tfrac{1-\epsilon}{2} c_n n\lambda \text{ for some } u\right] \\ & \leq P\left[\|s_n(\beta_0(k))\| \geq \tfrac{1}{2}\sqrt{5m\Lambda_m n \log p_n}\right] \\ & = P\left[\|X_k'(E-\mu)\| \geq \tfrac{1}{2}\sqrt{5m\Lambda_m n \log p_n}\right] \\ & \leq \exp\{-2m \log p_n\} = p_n^{-2m}, \end{aligned}$$

where we used that $s_n(\beta_0(k)) = X_k'(\mathbf{E} - \mu)$, and applied Theorem 3.8.2 to the quadratic form $(\mathbf{E} - \mu)' X_k X_k' (\mathbf{E} - \mu)$. This implies that with probability at least $1 - p_n^{-2m}$, we have $L_n(\beta(k)) - L_n(\beta_0(k)) < 0$. The concavity of L_n implies that $\left\|\hat{\beta}(k) - \beta_0(k)\right\| \leq c_n$ with probability at least $1 - p_n^{-2m}$. By taking a union bound over all models $k \supset t$ with size at most m_n, we have

$$P\left[\sup_{k \supset t : |k| = m} \left\|\hat{\beta}(k) - \beta_0(k)\right\| > c_n, \text{ for any } m \leq m_n\right] \leq \sum_{|t| \leq m \leq m_n} p_n^{-2m} p_n^m \to 0,$$

which proves the lemma. □

We are now ready to prove that $\sum_{k \in M_u} PR(k,t) \xrightarrow{\mathrm{P}} 1$ for $u = 1, 2, 3$. Using the joint posterior given by Theorem 3.3.1, we obtain the posterior of Z by integrating

out the other variables. That is,

$$P(Z = k \mid \mathbf{E}) = C^* Q_{n,k} \int_{\beta(k)} \exp\{L_n(\beta(k))\} \exp\left\{-\tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k)\right\} d\beta(k), \tag{3.54}$$

where $Q_{n,k} = (n + \tau_{0n}^{-2})^{|k|/2}\ v_n^{-|k|}$ and C^* is the normalizing constant.

For Overfitted models:

By Taylor's expansion of $L_n(\beta(k))$ around $\hat{\beta}(k)$ (the MLE of $\beta(k)$ under model k), we have

$$L_n(\beta(k)) = L_n(\hat{\beta}(k)) - \frac{1}{2}(\beta(k) - \hat{\beta}(k))' H_n(\tilde{\beta}(k))(\beta(k) - \hat{\beta}(k)), \tag{3.55}$$

for some $\tilde{\beta}(k)$ such that $\|\tilde{\beta}(k) - \hat{\beta}(k)\| \leq \|\beta(k) - \hat{\beta}(k)\|$. Note that $\tilde{\beta}(k)$ may depend on $\beta(k)$. However, due to Lemmas 3.8.4 & 3.8.6, for any $k \in M_1$, we have

$$L_n(\beta(k)) - L_n(\hat{\beta}(k)) \leq -\frac{(1-\epsilon)}{2}(\beta(k) - \hat{\beta}(k))' H_n(\beta_0(k))(\beta(k) - \hat{\beta}(k)),$$

for all $\beta(k)$ such that $\|\beta(k) - \beta_0(k)\| < c\sqrt{|k|\Lambda_{|k|}\log p_n/n} := cw_n$. Note that for $\beta(k)$ such that $\|\beta(k) - \hat{\beta}(k)\| = cw_n/2$,

$$L_n(\beta(k)) - L_n(\hat{\beta}(k)) \leq -(1-\epsilon)n\lambda c^2 w_n^2/4 = -(1-\epsilon)c^2\lambda|k|\Lambda_{|k|}\log p_n/4 \to -\infty. \tag{3.56}$$

By concavity of $L_n(\cdot)$ and the fact that $\hat{\beta}(k)$ maximizes $L_n(\beta(k))$, (3.56) also holds for any $\|\beta(k) - \hat{\beta}(k)\| > cw_n/2$. Now due to Lemma 3.8.6, we have $B := \{\beta(k) : \|\beta(k) - \hat{\beta}(k)\| \leq cw_n/2\} \subset \{\beta(k) : \|\beta(k) - \beta_0(k)\| \leq cw_n\}$ with probability going to

one uniformly in M_1 (for c large enough). Therefore, we have for any $k \subset M_1$,

$$
\begin{aligned}
&P(Z = k \mid \mathbf{E}) \\
&\leq C^* \; Q_{n,k} \exp\{L_n(\hat{\beta}(k))\} \\
&\times \Big(\int_B \exp\left\{ -\tfrac{1}{2}(1-\epsilon)(\beta(k)-\hat{\beta}(k))' H_n(\beta_0(k))(\beta(k)-\hat{\beta}(k)) - \tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k) \right\} d\beta(k) \\
&\qquad + \exp(-c^2(1-\epsilon)\lambda\Lambda_{|k|}|k| \log p_n/4) \int_{B^c} \exp\{-\tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k)\} d\beta(k) \Big) \\
&\leq w' \; Q_{n,k} \;\; \exp\{L_n(\hat{\beta}(k))\} \; \tau_{1n}^{|k|} \;\; |H_n(\beta_0(k))(1-\epsilon)\tau_{1n}^2 + I|^{-1/2}(1+o(1)),
\end{aligned}
\tag{3.57}
$$

where w' is a constant. We used that for $A = (1-\epsilon)H_n(\beta_0(k))$,

$$
\begin{aligned}
&\int_B \exp\left\{ -\tfrac{1}{2}(\beta(k)-\hat{\beta}(k))' A (\beta(k)-\hat{\beta}(k)) - \tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k) \right\} d\beta(k) \\
&= \int_B \exp\left\{ -\tfrac{1}{2}\left((\beta(k)-\beta^*(k))'(A+\tau_{1n}^{-2}I)(\beta(k)-\beta^*(k)) \right) \right\} d\beta(k) \\
&\quad \times \exp\left\{ -\tfrac{1}{2}\hat{\beta}(k)' \left(A - A(A+\tau_{1n}^{-2})^{-1}A \right) \hat{\beta}(k) \right\} \\
&\leq w' |H_n(\beta_0(k))(1-\epsilon) + \tau_{1n}^{-2}I|^{-1/2} \exp\left\{ -\tfrac{1}{2}\hat{\beta}(k)' \left(A - A(A+\tau_{1n}^{-2})^{-1}A \right) \hat{\beta}(k) \right\} \\
&\leq w' |A + \tau_{1n}^{-2}I|^{-1/2},
\end{aligned}
$$

where $\beta^*(k) = (A+\tau_{1n}^{-2}I)^{-1}A\hat{\beta}(k)$. Following similar arguments, we obtain a lower bound for $P[Z = t \mid \mathbf{E}]$, i.e.,

$$
P(Z = t \mid \mathbf{E}) \geq w \;\; \exp\{L_n(\hat{\beta}(t))\} Q_{n,t}(\tau_{1n})^{|t|} |H_n(\beta_0(t))(1+\epsilon)\tau_{1n}^2 + I|^{-1/2}, \tag{3.58}
$$

for some constant w. Therefore, we have

$$
\begin{aligned}
&PR(k,t) \\
&\preceq \;\; \tfrac{Q_{n,k}}{Q_{n,t}} |H_n(\beta_0(t))(1+\epsilon) + \tau_{1n}^{-2}|^{1/2} |H_n(\beta_0(k)(1-\epsilon) + \tau_{1n}^{-2}|^{-1/2} \\
&\quad \times \exp\left\{ L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \right\} \\
&\preceq q_n^{(|k|-|t|)} (n\tau_{1n}^2)^{-(|k|-|t|)/2} (n\tau_{0n}^2+1)^{(|k|-|t|)/2} \exp\left\{ L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \right\} \\
&\preceq p_n^{-(2+\delta)(|k|-|t|)} \exp\left\{ L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \right\},
\end{aligned}
\tag{3.59}
$$

where δ is from Condition 3.3.4. By applying Taylor's expansion for $L_n(\beta(k))$ about $\beta_0(k)$ and evaluating at $\hat{\beta}(k)$, we would obtain

$$\begin{aligned} L_n(\hat{\beta}(k)) &= L_n(\beta_0(k)) + (\hat{\beta}(k) - \beta_0(k))' s_n(\beta_0(k)) \\ &\quad -\tfrac{1}{2}(\hat{\beta}(k) - \beta_0(k))' H_n(\tilde{\beta}(k))(\hat{\beta}(k) - \beta_0(k)), \end{aligned} \tag{3.60}$$

such that $\|\tilde{\beta}(k) - \beta_0(k)\| \leq \|\hat{\beta}(k) - \beta_0(k)\|$. Then due to Lemmas 3.8.4 & 3.8.6, we have

$$\begin{aligned} & L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t)) \\ & \leq L_n(\hat{\beta}(k)) - L_n(\beta_0(k)) \\ & \leq (\hat{\beta}(k) - \beta_0(k))' s_n(\beta_0(k)) - \tfrac{1-\epsilon}{2}(\hat{\beta}(k) - \beta_0(k))' H_n(\beta_0(k))(\hat{\beta}(k) - \beta_0(k)) \\ & \leq \tfrac{1}{2(1-\epsilon)} s_n(\beta_0(k))' H_n(\beta_0(k))^{-1} s_n(\beta_0(k)) \\ & = \tfrac{1}{2(1-\epsilon)} (\mathbf{E} - \mu)' X_k H_n(\beta_0(k))^{-1} X_k' (\mathbf{E} - \mu) \\ & = \tfrac{1}{2(1-\epsilon)} \tilde{U}' P_k \tilde{U}, \end{aligned} \tag{3.61}$$

where $\Sigma = Diag(\sigma_1^2, \cdots, \sigma_n^2)$, $\tilde{U} = \Sigma^{-1/2}(\mathbf{E} - \mu)$, and $P_k = \Sigma^{1/2} X_k H_n(\beta_0(k))^{-1} X_k' \Sigma^{1/2}$ is the projection matrix onto the column space of $\Sigma^{1/2} X_k$.

We now use Lemma 3.8.5 to obtain a probability bound on the quadratic form obtained in Equation (3.61). Consider $b_n = (1 + \delta^*)(1 + 2w) \log p_n$, where w is small such that $(1 + \delta^*)(1 + 2w) < (1 + \delta)$ (this is possible due to Condition 3.3.4) and ϵ is small such that $\psi = (1 - \epsilon)(1 + w) - 1 > 0$. Then, we have

$$\begin{aligned} & P\left[\left|L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))\right| > b_n(|k| - |t|)\right] \\ & \leq P\left[\tilde{U}' P_k \tilde{U} > 2(1-\epsilon) b_n(|k| - |t|)\right] \\ & \leq \exp\left\{-(1-\epsilon)(1+w)(|k| - |t|) \log p_n\right\}, \\ & = p_n^{-(1+\psi)(|k|-|t|)}. \end{aligned} \tag{3.62}$$

Now, by taking a union bound we obtain

$$\begin{aligned} &P[|L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))| > b_n(|k| - |t|) \text{ for any } k \in M_1] \\ &\quad \leq \sum_{|k|=|t|+1}^{m_n} p_n^{-(1+\psi)(|k|-|t|)} p_n^{-(|k|-|t|)} \to 0. \end{aligned}$$

By restricting to the complement of the above small probability event, and using Equation (3.59), we have

$$\begin{aligned} &\sum_{k \in M_1} PR(k,t) \\ &\quad \preceq \sum_{k \in M_1} p_n^{-(2+\delta)(|k|-|t|)} \exp\{(1+\delta^*)(1+2w)(|k|-|t|)\log p_n\} \\ &\quad \preceq \sum_{d=|t|+1}^{m_n} p_n^{(d-|t|)} p_n^{-(2+\delta)(d-|t|)} \exp\{(1+\delta^*)(1+2w)(d-|t|)\log p_n\} \\ &\quad \preceq p_n^{-(1+\delta-(1+\delta^*)(1+2w))} = o_P(1). \end{aligned} \tag{3.63}$$

Large models

The proof for large models is similar to the overfitted ones. We only need to note that for a large underfitted model, we shall work with $k^* = k \cup t$. Similar to the previous case, consider Taylor's expansion (3.55) of $L_n(\beta(k^*))$ evaluated at $\beta(k^*) = (\beta(k) \cup 0)_{|k^*| \times 1}$. As before, we then have

$$\begin{aligned} L_n(\beta(k^*)) &= L_n(\hat{\beta}(k^*)) - \tfrac{1}{2}(\beta(k^*) - \hat{\beta}(k^*))' H_n(\tilde{\beta}(k^*))((\beta(k^*) - \hat{\beta}(k^*)) \\ &\leq L_n(\hat{\beta}(k^*)) - \tfrac{n(1-\epsilon)\lambda}{2}(\beta(k^*) - \hat{\beta}(k^*))'(\beta(k^*) - \hat{\beta}(k^*)), \end{aligned}$$

for all $\beta(k^*)$ such that $\|\beta(k^*) - \beta_0(k^*)\| < c\sqrt{|k^*|\Lambda_{|k^*|}\log p_n/n}$. Then, following the arguments in the previous case, we have

$$P(Z = k \mid \mathbf{E}) \leq w' \, Q_{n,k} \, \exp\{L_n(\hat{\beta}(k^*))\} \, \tau_{1n}^{|k|} \, (n\lambda(1-\epsilon)\tau_{1n}^2)^{-|k|/2}(1+o(1)),$$

which implies that

$$PR(k,t) \;\preceq\; p_n^{-(2+\delta)(|k|-|t|)} \exp\left\{L_n(\hat{\beta}(k^*)) - L_n(\hat{\beta}(t))\right\}. \tag{3.64}$$

Now, similar to Equation (3.62) we have

$$P\left[\cup_{k:|k|=d}\left|L_n(\hat{\beta}(k^*)) - L_n(\hat{\beta}(t))\right| > b_n(|k|-|t|)\right] \;\leq p_n^{-(1+\psi)(d-|t|)} p_n^d,$$

which implies that

$$\begin{aligned} &P[\left|L_n(\hat{\beta}(k^*)) - L_n(\hat{\beta}(t))\right| > b_n(|k|-|t|) \text{ for any } k \in M_2] \\ &\leq \sum_{d=K|t|+1}^{m_n} p_n^{-(1+\psi)(d-|t|)} p_n^d \\ &\preceq p_n^{-\psi K|t|} p_n^{(1+\psi)|t|} \to 0. \end{aligned}$$

if we take $K > (1+\psi)/\psi$, in the definition of M_2. Hence, we have the result for large models by observing that $\sum_{k\in M_2} PR(k,t) = o_P(1)$, as in the case of Equation (3.63).

Underfitted models

Now, we shall prove the same for under-fitted models, i.e., for models in M_3. From Equation (3.54), we have

$$\begin{aligned} P(Z=k \mid \mathbf{E}) \;&\leq C^* \exp\{L_n(\hat{\beta}(k))\}\; Q_{n,k} \int_{\beta(k)} \exp\left\{-\tfrac{1}{2\tau_{1n}^2}\beta(k)'\beta(k)\right\} d\beta(k) \\ &\preceq \exp\{L_n(\hat{\beta}(k))\} Q_{n,k} (\tau_{1n})^{|k|}. \end{aligned} \tag{3.65}$$

We shall now show that

$$|L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))| \preceq -|t|\Lambda_{(K+1)|t|} \log p_n,$$

with probability going to one, uniformly over models in M_3. To see this, for any $k \in M_3$, let $k^* = k \cup t$. Consider $\beta(k^*)$ such that $\|\beta(k^*) - \beta_0(k^*)\| = a_n$, where $a_n = \min_{1 \le i \le |t|} |\beta_{0i}(t)| \ge \sqrt{c|t|\Lambda_{(K+1)|t|} \log p_n / n}$, for a large enough constant c due to Condition 3.3.3. Then, we have

$$\begin{aligned}
&L_n(\beta(k^*)) - L_n(\beta_0(k^*)) \\
&\le (\beta(k^*) - \beta_0(k^*))' s_n(\beta_0(k^*)) - \tfrac{1}{2}(\beta(k^*) - \beta_0(k^*))' H_n(\tilde{\beta}(k^*))(\beta(k^*) - \beta_0(k^*)) \\
&\le (\beta(k^*) - \beta_0(k^*))' s_n(\beta_0(k^*)) - \tfrac{1}{2} n\lambda \|\beta(k^*) - \beta_0(k^*)\|^2 \\
&\le a_n \|s_n(\beta_0(k^*))\| - \tfrac{1}{2} n\lambda a_n^2.
\end{aligned} \tag{3.66}$$

Next, we shall obtain a bound for $\|s_n(\beta_0(k^*))\|$. Note that, $\beta_0(k^*)$ is the true $\beta_0(t)$ appended with zeroes for $k \cap t^c$. From the proof of Lemma 3.8.6, we have $\|s_n(\beta_0(k^*))\| = O_P(\sqrt{n|t|\Lambda_{(K+1)|t|} \log p_n})$ uniformly in $k^* = \{k \cup t : k \in M_3\}$ as $|k^*|$ is bounded by $(K+1)|t|$. Therefore, from (3.66), we have

$$L_n(\beta(k^*)) - L_n(\beta_0(k^*)) \quad \le w' a_n \sqrt{n|t|\Lambda_{(K+1)|t|} \log p_n} - \tfrac{1}{2} n\lambda a_n^2 \quad \preceq -\tfrac{1}{2} n\lambda a_n^2,$$

Let $\tilde{\beta}(k^*)$ be the $|k^*| \times 1$ vector including $\hat{\beta}(k)$ for k and zeroes for $k \cap t^c$. Then, we have $L_n(\tilde{\beta}(k^*)) = L_n(\hat{\beta}(k))$ and $\|\tilde{\beta}(k^*) - \beta_0(k^*)\| \ge a_n$. By the concavity of $L_n(\cdot)$, we obtain

$$L_n(\tilde{\beta}(k^*)) - L_n(\beta_0(k^*)) \le L_n(\beta(k^*)) - L_n(\beta_0(k^*)) \preceq -\frac{1}{2} n\lambda a_n^2. \tag{3.67}$$

On the other hand, we have

$$\frac{Q_{n,k}(\tau_{1n})^{|k|}}{Q_{n,t}(\tau_{1n})^{|t|} (n\tau_{1n}^2)^{-|t|/2}} \le w p_n^{c'|t|},$$

which implies that

$$\begin{aligned}\sum_{k\in M_3} PR(k,t) &\preceq \sum_{k\in M_3} p_n^{c'|t|}\exp\{L_n(\hat{\beta}(k)) - L_n(\hat{\beta}(t))\}\\ &\leq \sum_{k\in M_3} p_n^{c'|t|}\exp\{L_n(\tilde{\beta}(k^*)) - L_n(\beta_0(k^*))\}\\ &\preceq p_n^{c'|t|}p_n^{(K+1)|t|}\exp\left\{-\tfrac{1}{2}n\lambda a_n^2\right\}\\ &= \exp\left\{(K+c'+1)|t|\log p_n - \tfrac{1}{2}n\lambda a_n^2\right\}\\ &\to 0,\ as\ n\to\infty,\end{aligned}$$

with probability going to one. Therefore, we have $\sum_{k:k\neq t;|k|\leq m_n} \frac{P[Z=k|\mathbf{E}]}{P[Z=t|\mathbf{E}]} \xrightarrow{\text{P}} 0$. This implies that $\sum_{k:k\neq t;|k|\leq m_n} \frac{P[Z=k||Z|\leq m_n,\mathbf{E}]}{P[Z=t||Z|\leq m_n,\mathbf{E}]} \xrightarrow{\text{P}} 0$, which in turn implies that $P[Z = t \mid |Z| \leq m_n, \mathbf{E}] \xrightarrow{\text{P}} 1$. □

Lemma 3.8.7. *Let $X_{n\times p}$ be a random design matrix with rows i.i.d. from a sub-Gaussian distribution with covariance matrix Σ. Let the principal submatrices of Σ of order $m_n + |t|$ have minimum eigenvalues bounded (away from zero). Also, assume that $\beta_0(t)$ is a $|t| \times 1$ vector satisfying $P[|x_i'\beta_0(t)| \geq M] \leq w < 1$, for some $M > 0$, where x_i is the i^{th} row of X (this is a weaker version of the condition that all the log-odds are bounded as assumed in Bühlmann and van de Geer (2011)). Then, we have*

$$0 < \lambda \leq \min_{k:|k|\leq m_n+|t|} \lambda_{min}\left(n^{-1}H_n(\beta_0(k))\right).$$

Proof:

Due to Theorem 5.39 of Vershynin (2012) and following the proof of Lemma 6.1 of Narisetty and He (2014), we have with probability going to one

$$0 < c' \leq \min_{k:|k|\leq m_n+|t|} \lambda_{min}\left(n^{-1}X_k'\Sigma_k^{-1}X_k\right),$$

where Σ_k is the $|k| \times |k|$ submatrix of $\Sigma = Cov(x_i)$ corresponding to the covariates in

model $|k|$. As the minimum eigenvalue of Σ_k is bounded away from zero, we further have

$$0 < c \leq \min_{k:|k|\leq m_n+|t|} \lambda_{min}\left(n^{-1}X_k'X_k\right). \tag{3.68}$$

for some $c > 0$. Now, as $\beta_0(t)$ satisfies $P[|x_i'\beta_0(t)| \geq M] \leq w < 1$, for some $M > 0$, the set of indices $I = \{i : |X_i'\beta_0(t)| \leq M\}$ satisfies $|I| \geq n(1-w)/2$ with exponentially large probability. Then, with probability gping to one, we have

$$\begin{aligned}
\lambda_{min}\left(n^{-1}H_n(\beta_0(k))\right) &= \lambda_{min}\left(n^{-1}X_k'Diag\left(\sigma_i^2(\beta_0(k))\,X_k\right)\right. \\
&\geq \lambda_{min}\left(n^{-1}X_k^{*\prime}D^*X_k^*\right) \\
&\geq d_M\lambda_{min}\left(n^{-1}X_k^{*\prime}X_k^*\right) \\
&\geq (1-w)cd_M/2 > 0,
\end{aligned} \tag{3.69}$$

were, X^* is the $|I|\times p$ matrix with rows from X indexed by I, $D^* = Diag\left(\sigma_i^2(\beta_0(k)) : i \in I\right)$, and $d_M = \exp\{M\}/(1+\exp\{M\})^2$. We have applied Inequality (3.68) to X^* to get the last inequality in (3.69), which proves the lemma. □

3.9 Discussion and Conclusion

In this chapter, a novel Gibbs sampler is proposed for variable selection in logistic regression. The proposed Skinny Gibbs has desired theoretical and computational properties. The strong selection consistency of the Gibbs sampler is established, which guarantees that the posterior probability of the true model goes to one. Computationally, each iteration of Skinny Gibbs requires complexity linear in p_n. Empirical results presented in the paper illustrate the good performance of this approach for variable selection.

The theoretical and computational techniques developed in the paper can be extended to other models that have normal scale mixture representations. Skinny Gibbs can also be applicable to the case where the prior distribution has a normal scale

mixture representation such as the conjugate priors of Chen et al. (2008). Other extensions to cases such as fixed and random effects selection (Kinney and Dunson (2007)) under high dimensional covariates are potential future research directions.

There has been some recent advances on theory and computation for Bayesian high dimensional model selection. Ročková and George (2014) proposed an EM algorithm for identifying the posterior mode in the context of Gaussian spike and slab variable selection. Other recent approaches include Approximate Bayesian Computation (Bonassi and West, 2015), and Approximate Message Passing (Bonassi et al., 2015). There has been recent advances in understanding the computational complexity of Bayesian model selection algorithms including Román and Hobert (2012, 2015); Khare and Hobert (2013); Yang et al. (2016). Further research to understand the computational complexity and mixing properties of Skinny Gibbs remains to be an important futurre research direction.

CHAPTER IV

Extremal Notion of Depth for Functional Data and Applications

4.1 Introduction

Ranks, order-statistics, and quantiles have been used extensively for statistical inference with univariate data. Many authors have studied their generalizations for multivariate data using notions of "data depth". The classical measure based on Mahalanobis distance (Mahalanobis, 1936) is ideally suited for multivariate normal (or more generally elliptical) distributions. Tukey's half-space depth (Tukey, 1975) appears to be the first new notion for the multivariate case, and there has been a lot of work since then. Brown (1983) defined a 'median' for multivariate data using the L_1 metric, and Vardi and Zhang (2000) extended this to obtain a notion of multivariate depth. Other concepts include simplicial depth (Liu, 1990), geometric notion of quantiles (Chaudhuri, 1996), projection depth (Zuo and Serfling, 2000; Zuo, 2003), and spatial depth (Vardi and Zhang, 2000; Serfling, 2002). See Zuo and Serfling (2000) for a review. Various types of statistical inference have also been based on multivariate depth notions, including classification (Jörnsten, 2004; Ghosh and Chaudhuri, 2005; Li et al., 2012), outlier detection (Donoho and Gasko, 1992; Mosler, 2002), and hypothesis testing (Liu and Singh, 1997). Liu et al. (1999)

studied the use of depth-based methods for inference on distributional quantities such as location, scale, bias, skewness and kurtosis.

In comparison, there has been limited work on depth for functional data. Fraiman and Muniz (2001) proposed integrated data depth (ID); Lopéz-Pintado and Romo (2009) introduced band depth (BD) and modified band depth (MBD); and Lopéz-Pintado and Romo (2011) proposed a half-region depth (HRD) which was intended to be a generalization of the half-space depth for multivariate data. Several other notions of depth for multivariate data have also been extended to functional data. For instance, Chakraborty and Chaudhuri (2014a) developed spatial depth (SD) for functional data. One can also extend Zuo (2003)'s projection-based depth functions and multivariate medians to functional data. However, several of these notions and extensions suffer from a "degeneracy" problem pointed out in Chakraborty and Chaudhuri (2014a). Specifically, in infinite-dimensional function spaces, with probability one, all the functions will have zero depth (Chakraborty and Chaudhuri, 2014a,b).

As with multivariate data, functional depth can be used for many applications. Fraiman and Muniz (2001) used ID for constructing trimmed functional mean. Lopéz-Pintado and Romo (2006) used BD for classification of functional data, Sun and Genton (2011) proposed functional boxplots based on MBD, and Hubert et al. (2015) considered functional outlier detection based on some measures of depth and outlyingness. Depth notions can also be used to obtain central regions of data which, for instance, form the basis for constructing boxplots.

Both ID and MBD, which appear to be the most common, are based on some form of averaging of the depth at different points in the domain and, as a result, their depth level sets are not convex. This has important implications for corresponding central regions as discussed in later sections. In addition, they may not be resistant to functions that are outlying in small regions of the domain.

This paper develops a new notion called Extremal Depth (ED) for functional

data. We will show that ED and associated central regions possess several attractive features including:

- ED central regions achieve their nominal coverage *exactly* due to the convexity of the depth contours;
- There is a direct correspondence between the (simultaneous) ED central regions and the usual pointwise central regions based on quantiles; as a consequence, the width of the ED simultaneous central regions is, roughly speaking, proportional to a measure of variation at each point; and
- ED central regions are resistant to functions that are 'outlying' even in a small region of the domain.

These features lead to desirable properties for corresponding functional boxplots, simultaneous confidence regions for function estimation, and outlier detection.

The rest of the article is organized as follows. Section 4.2 introduces ED for a sample of functional data and illustrates it on a real dataset. Section 4.3 defines ED for general probability distributions and discusses its theoretical properties. Section 4.4 deals with construction of central regions of functional data and develops several results including exact coverage and correspondence to pointwise regions. Section 4.5 describes applications to functional boxplots and outlier detection, and the advantages of ED-based methods over others. Section 4.6 demonstrates how ED can be used to construct simultaneous confidence bands for functional parameters.

4.2 Extremal Depth

4.2.1 Depth distribution

Let $S := \{f_1(t), f_2(t), \cdots, f_n(t)\}$ be a collection of n functional observations with $t \in \mathcal{T}$. For ease of exposition, we assume throughout that the functions are continuous and infinite-dimensional and, without loss of generality, we take the domain $\mathcal{T}$ to be $[0, 1]$. However, as with other notions, ED can also be used for functional observations observed at a finite number of points.

Let $g(t)$ be a given function that may or may not be a member of S. For each fixed $t \in [0,1]$, define the pointwise depth of $g(t)$ with respect to S as

$$D_g(t,S) \;\; := 1 - \frac{|\sum_{i=1}^{n}[\mathbb{1}\{f_i(t) < g(t)\} - \mathbb{1}\{f_i(t) > g(t)\}]|}{n}. \tag{4.1}$$

Thus, any given function $g(\cdot)$ is mapped into the pointwise depth function $D_g(\cdot, S)$ whose range is $\mathbb{D}_g \subset \{0, 1/n, 2/n, \cdots, 1\}$. Let $\mathbb{D}$ be the union of $\mathbb{D}_g$ over all functions g. We call $\mathbb{D}$ the *set of depth values*.

Let $\Phi_g(\cdot)$ be the cumulative distributions function (CDF) of the distinct values taken by $D_g(t,S)$ as t varies in $[0,\ 1]$. This will be called the depth CDF or d-CDF and defined formally as

$$\Phi_g(r) = \int_0^1 \mathbb{1}\{D_g(t,S) \le r\} dt, \tag{4.2}$$

for each fixed $r \in \mathbb{D}$. Note that if Φ_g has most of its mass close to zero (or one), then g is away from (or close to) the center of the data. (See the illustrative example in Figure 4.1 for computation of d-CDFs.)

We need an appropriate way to order these d-CDFs (distributions) to get a one-dimensional notion of depth. (Clearly, there is no single approach that will dominate all others, so one has to decide on the appropriate one by examining its performance under different situations.) First-order stochastic dominance may appear to be the most natural way to order distributions, but it is not useful here except in the trivial case where the functions do not cross. Alternatively, one can use a simple functional of the d-CDFs such as the mean or median. In fact, the integrated depth (or ID) by Fraiman and Muniz (2001) is given (approximately) by $ID(g) = \int_0^1 D_g(t,S)\ dt$. (It is approximate because, in the definition of ID, $D_g(t,S)$ is based on the term $\sum_{i=1}^{n}\{\mathbb{1}\{f_i(t) \le g(t)\}$ rather than $\sum_{i=1}^{n}\{\mathbb{1}\{f_i(t) < g(t)\}$ which appears in equation 4.1.) The other notions, such as Band Depth (BD) or Modified Band Depth (MBD), do not depend directly on the d-CDFs. We will provide a comparison of various

functional depths in Section 4.3.

4.2.2 Definition of Extremal Depth

Our notion of extremal depth will be based on a comparison of $\Phi_g(r)$, the d-CDFs, for r near zero. It focuses on the left tail of the distribution and can be viewed as left-tail stochastic ordering. The idea can be explained simply as follows. Consider two functions g and h with corresponding d-CDFs Φ_g and Φ_h. Let $0 \leq d_1 < d_2 < \cdots < d_M \leq 1$ be the ordered elements of their combined depth levels. If $\Phi_h(d_1) > \Phi_g(d_1)$, then $h \prec g$ (or equivalently $g \succ h$, and is read as h is more extreme then g); if $\Phi_g(d_1) > \Phi_h(d_1)$, then $h \succ g$. If $\Phi_g(d_1) = \Phi_h(d_1)$, we move to d_2 and make a similar comparison based on their values at d_2. The comparison is repeated until the tie is broken. If $\Phi_g(d_i) = \Phi_h(d_i)$ for all $i = 1, ...M$, the two functions are equivalent in terms of depth and are denoted as $g \sim h$. (This ordering is defined formally in Section 4.3 when we consider a more general context with arbitrary function spaces S and distributions.)

The extremal depth (ED) of a function g with respect to the sample $S = \{f_1, \cdots, f_n\}$ can now be defined as

$$ED(g, S) = \frac{\#\{i : g \succeq f_i\}}{n}, \tag{4.3}$$

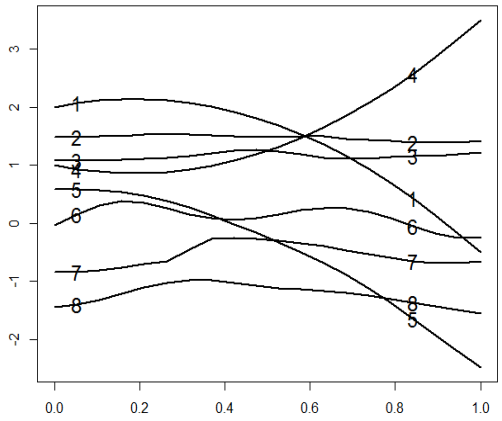
where $g \succeq f_i$ if either $g \succ f_i$ or $g \sim f_i$. If $g \in S$, then this is just the normalized rank of g; i.e., $ED(g, S) = R(g, S)/n$ where $R(g, S) = \{i : g \succeq f_i\}$ is the rank of g. This relationship between ED and its rank is similar to corresponding relationships of normalized rank functions for some other depth notions in the literature (Liu and Singh, 1993; Lopéz-Pintado and Romo, 2009). The distinguishing feature of ED is the nature of the ordering, i.e., left-tail stochastic ordering of the depth distributions.

The ED median of a set of functional observations S can be defined (in an obvious manner) as the function (or functions) in S that has (or have) the largest depth. ED median also has the following min-max interpretation. For a function $g \in S$, let

$d_{\min}(g) = \inf_{t\in[0,1]} D_g(t, S)$, the pointwise depth in Equation (4.1). Then, if g is an ED median, $d_{\min}(g)$ attains the maximum: $d_{\min}(g) = \max_{1\le k\le n} d_{\min}(f_k)$; i.e., an ED median maximizes the minimum pointwise depth over $t \in [0,\ 1]$.

Figure 4.1: An illustrative example: (a) eight sample functions and (b) their depth CDF's. The columns correspond to each of four depth levels $\{1/8, 3/8, 5/8, 7/8\}$ and the rows correspond to different sample functions.

(a)



(b)

	$r = 1/8$	$r = 3/8$	$r = 5/8$	$r = 7/8$
$\Phi_1(r)$	**0.60**	0.60	0.70	1.00
$\Phi_2(r)$	0.00	**1.00**	1.00	1.00
$\Phi_3(r)$	0.00	0.00	**0.80**	1.00
$\Phi_4(r)$	**0.40**	0.40	0.50	1.00
$\Phi_5(r)$	**0.25**	0.50	**0.60**	1.00
$\Phi_6(r)$	0.00	0.00	0.40	1.00
$\Phi_7(r)$	0.00	**0.50**	1.00	1.00
$\Phi_8(r)$	**0.75**	1.00	1.00	1.00

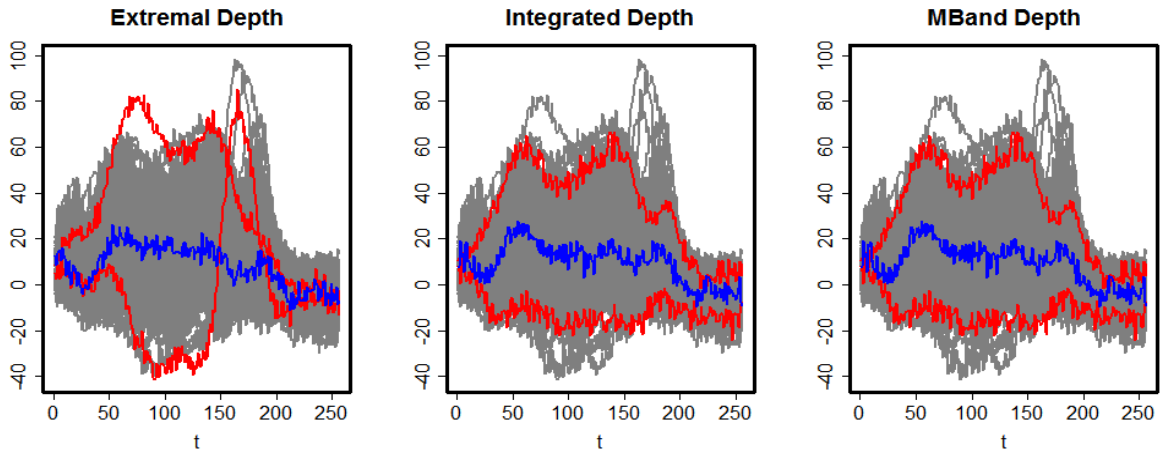
We now consider the illustrative example in Figure 4.1 (a) with eight sample functions. The d-CDF's of all the functions are shown as a table in Figure 4.1 (b). ED gives the ordering $f_8 \prec f_1 \prec f_4 \prec f_5 \prec f_2 \prec f_7 \prec f_3 \prec f_6$. So f_8 is the most extreme observation and f_6 is the deepest (median). Note that the ordering $f_8 \prec f_1 \prec f_4 \prec f_5$ is based on a comparison of the d-CDF values at $r = 1/8$ (these values are in bold); the ordering $f_2 \prec f_7$ is based on their d-CDF values at $r = 3/8$; and the ordering $f_3 \prec f_6$ is based on their values at $r = 5/8$. From this, we get the extremal depths of these functions as: $ED(f_8) = 1/8, ED(f_1) = 2/8$ and so on.

We now use the orthosis dataset (Cahouet et al., 2002) to illustrate ED and visually compare its performance with ID and MBD. This dataset consists of moment of force measured at the knee under four different experimental conditions, measured at 256 equally-spaced time points for seven subjects with ten replications per subject. Figure 4.2 shows the results for 240 functional observations from six subjects who have similar range of moment of force values. The x-axis represents time when the measurement

is taken and the y-axis shows the resultant moment of force at the knee. The sample functions are plotted in gray, while the deepest function is in blue and the two least deep functions are in red.

The three panels in Figure 4.2 correspond to ED, ID and MBD respectively. We restrict attention to ID and MBD in our empirical comparisons because these notions are commonly used, non-degenerate and invariant to monotone-transformations. (These properties are discussed in Section 4.3.2). The medians for all three notions are qualitatively similar. However, the two extreme functions based on ID and MBD fall well within the boundaries of the entire data cloud while the two for ED are most extreme in at least some part of the domain. As we shall see, this is due to the non-convexity of the depth level sets of ID and MBD.

Figure 4.2: Orthosis data example: The three panels show the 240 functional observations (in gray) along with their two most outlying functions (in red) and the median (in blue) using ED, ID and MBD, respectively.



4.3 ED for Theoretical (Population) Distributions and Its Properties

There has been discussion of the desirable properties for depth notions in the literature (Liu, 1990; Zuo and Serfling, 2000; Mosler and Polyakov, 2012). We will

examine the performance of ED with respect to these properties and compare it with existing notions. To do this, we first have to extend the notion of ED from sample data to theoretical (population) distributions.

4.3.1 Definition

Let $\mathbb{P}$ be a distribution on $C[0,1]$ and $X \sim \mathbb{P}$ be a random function. We denote F_t to be the CDF of the random variable $X(t)$, and $\bar{F}_t(\cdot) = 1 - F_t(\cdot)$. For any function g, define the depth of g at t as

$$\begin{aligned} D_g(t, X) &:= 1 - |\mathbb{P}[X(t) > g(t)] - \mathbb{P}[X(t) < g(t)]| \\ &= 1 - |\bar{F}_t(g(t)) - F_t(g(t)-)|. \end{aligned} \tag{4.4}$$

When the univariate distributions F_t are continuous, $D_g(t, X) = 1 - |1 - 2F_t(g(t))|$.

The d-CDF of the function g is defined, similar to the finite-sample case, as

$$\Phi_g(r) = \int_{[0,1]} \mathbb{1}\{D_g(t, X) \leq r\}\, dt, \tag{4.5}$$

for $r \in [0,1]$. Note that, if necessary, one can replace the uniform weight distribution in the definition of $\Phi_g(r)$ by a weighted measure to give higher or lower importance to certain regions of the domain.

As in the finite-sample case, we use the d-CDFs to obtain an ordering of functions. Because the d-CDFs now can be continuous, we need a slightly more general definition. Consider a pair of functions $g, h \in C[0,1]$, and define

$$r^* = \inf\{r \in [0,1] : \Phi_g(r) \neq \Phi_h(r)\}, \tag{4.6}$$

the infimum of values at which d-CDFs of g and h differ. Then, we say $h \prec g$ (h more extreme than g) if there exists $\delta > 0$ such that $\Phi_h(r) > \Phi_g(r)$ for all $r \in (r^*,\ r^* + \delta)$. If $r^* < 1$, such a δ exists as long as Φ_g and Φ_h have finitely many crossings. If $r^* = 1$, we say that $g \sim h$.

Extremal depth of a function g w.r.t. the distribution $\mathbb{P}$ is now defined as

$$ED(g, \mathbb{P}) := 1 - \mathbb{P}\left[g \prec X\right] = \mathbb{P}\left[g \succeq X\right], \text{ where } X \sim \mathbb{P}. \tag{4.7}$$

4.3.2 Properties

Liu (1990); Zuo and Serfling (2000) proposed several desirable properties for multivariate depth notions, and Mosler and Polyakov (2012) extended them for functional depth. The first four properties below are satisfied by ED, ID, BD and MBD but not by some others. The next two concepts discussed below, convexity and 'null at the boundary' (NAB), are satisfied by ED but not by ID and MBD. The convexity property leads to desirable shapes for central regions as shown in the next section. The NAB property is also important and is related to being resistant to outliers.

Transitivity (if $f_1 \prec f_2$ and $f_2 \prec f_3$, then $f_1 \prec f_3$) and **invariance** under monotone transformations (order preserving as well as order reversing) are two well-known properties. It can be easily shown that ED satisfies them, as do ID, BD and MBD (where the ordering $f_1 \prec f_2$ for ID and MBD is interpreted as f_2 deeper than f_1). However, spatial depth (SD) (Chakraborty and Chaudhuri, 2014a) does not satisfy the invariance property. The details are omitted.

Maximality of the center property requires that if there exists a natural center for the distribution of interest, such as a center of symmetry, then it should have the highest depth. This holds for ED and that point is the ED median. When a center of symmetry exists, it has the highest depth for ED, ID and MBD; this is not necessarily true for SD. While BD also has the center of symmetry as its median under some conditions (Lopéz-Pintado and Romo (2009)), Chakraborty and Chaudhuri (2014c) showed that, for many common stochastic processes, BD assigns a depth of zero to the center of symmetry, making it not deeper than any other function.

Monotonicity from the center requires that, if m is a median and two functions f and g are such that either $m(t) \leq g(t) \leq f(t)$ or $m(t) \geq g(t) \geq f(t)$ for all t, then g should be at least as deep as f. ED, ID, BD and MBD all satisfy monotonicity from the center of symmetry, when it exists. The proof is omitted.

Convex depth level sets: For a given function h and fixed $\alpha \in (0,\ 1)$, define the ED level set as $\{h : ED(h, \mathbb{P}) \geq \alpha\}$.

Proposition 4.3.1. *Under a mild condition (Condition 4.7.1(b) in Appendix), the ED level sets are convex for each* $\alpha \in (0, 1)$.

This property is highly desirable for constructing central regions of a desired coverage $(1 - \alpha)$ (developed in the next section). Neither ID nor MBD is guaranteed to have convex depth level sets, which was already suggested by Figure 4.2. The proof is provided in the Appendix.

Null at the Boundary: (Mosler and Polyakov, 2012) considered a depth notion to satisfy the'null at infinity' (NAI) property if $D(h, \mathbb{P}) \to 0$ as $\|h\| \to \infty$. It is shown in the Appendix that ED satisfies the NAI property. Neither ID nor MBD satisfies the NAI property. This can be seen, for instance, by taking functions that go to infinity in a small interval but are near the center in the rest of the domain.

The NAI notion is not very informative if $\|h\|$ is bounded with $\mathbb{P}-$probability one. Therefore, we generalize it to the concept of 'null at the boundary' (NAB) which is defined in terms of quantiles rather than norms of the functional observations. The formal definition is given in Appendix where it is also shown that ED satisfies NAB property. Although BD may satisfy convexity and NAB properties, it may do so trivially due to the degeneracy problem noted earlier. ID and MBD do not satisfy NAB, since they do not satisfy the weaker NAI property.

4.3.3 Convergence of Sample ED

Fraiman and Muniz (2001) showed that, under suitable regularity conditions, the finite-sample versions of ID converge to the population quantity. The following proposition establishes the analogous consistency result for ED under suitable regularity conditions. The conditions and proof are given in the Appendix.

Proposition 4.3.2. *Let $\mathbb{P}$ be a stochastic process satisfying the regularity conditions 4.7.1 - 4.7.3 in the Appendix. Let $\mathbb{P}_n$ be the empirical distribution based on n samples from $\mathbb{P}$. Then,*

$$\lim_{n\to\infty} \sup_{f\in C[0,1]} |ED(f,\mathbb{P}_n) - ED(f,\mathbb{P})| \to 0,$$

4.3.4 Non-Degeneracy of ED

Chakraborty and Chaudhuri (2014a) showed that several existing notions of functional depth suffer from the following degeneracy problem. For a general class of continuous time Gaussian processes, with probability one, the depth of every function is zero. This is true for BD and the extensions of projection depth and half-region depth to functional data in the literature. ID, MBD, and SD do not suffer from these problems. Proposition 4.3.3 shows that extremal depth is non-degenerate for a general class of stochastic processes.

Consider $X = \{h(t, Y_t)\}, t \in [0,1]$, where: i) Y_t is a mean zero Gaussian process having continuous sample paths, bounded variance function $0 < \sigma^2(t) := E(Y^2(t)) < \infty$, and $\sup\{Y_t/\sigma(t), t \in [0,1]\}$ has a continuous distribution, ii) the function $h : [0,1] \times \mathbb{R}$ is continuous, and iii) $h(t,.)$ is strictly increasing with $h(t,s) \to \infty$ as $s \to \infty$ for each $t \in [0,1]$. Let $X \sim \mathbb{P}$, and define the *range of ED* for X as $R := \{\alpha \in [0,1] : ED(f,\mathbb{P}) = \alpha, \text{ for some } f \in C[0,\ 1]\}$. Then:

Proposition 4.3.3. *The range of ED for X is $(0,1]$.*

The result is proved in the Appendix.

4.4 Central Regions Based on ED

This section deals with construction of ED-based central regions, their theoretical properties and comparison with central regions based on other depth notions.

4.4.1 Definition and Properties

Consider a function space S of interest (such as $C[0,1]$ or a sample of n functional observations), and let $\mathbb{P}$ be the associated distribution of interest. Let $(1-\alpha)$ be the desired coverage level. Define the lower and upper $\alpha-$envelope functions as

$$\begin{aligned} f_L(t) &:= \inf\{f(t) : f \in S,\ ED(f,\mathbb{P}) > \alpha\}, \\ f_U(t) &:= \sup\{f(t) : f \in S,\ ED(f,\mathbb{P}) > \alpha\}, \end{aligned} \tag{4.8}$$

respectively. Then, the $(1-\alpha)$ ED central region is given by

$$C_{1-\alpha} = \{f \in S : f_L(t) \le f(t) \le f_U(t), \forall t \in [0,1]\}. \tag{4.9}$$

When S is a finite set of functions and $\mathbb{P}$ is the empirical distribution, then $C_{1-\alpha}$ is just the convex hull formed by all the sample functions having depth larger than α. When S is $C[0,1]$, and the marginal distribution of $\mathbb{P}$ at t has zero mass to the right of $f_L(t)$ or to the left of $f_U(t)$, we take $f_L(t)$ and $f_U(t)$ to be the largest and smallest possible values (which retain the marginal probability of the interval $[f_L(t), f_U(t)]$), respectively.

The following proposition shows that the central region of level α contains at least the desired amount of coverage $(1-\alpha)$. Further, when the boundary of the central region does not have any mass, the actual coverage equals the desired coverage exactly. This property is not shared by ID or MBD, and they often tend to have over-coverage problem. The proof is provided in the Appendix.

Fix α in the range of ED. Define the boundary set of $C_{1-\alpha}$ as $\partial C_{1-\alpha} = \{f \in C_{1-\alpha} : f(t) = f_L(t) \text{ or } f_U(t) \text{ for some } t \in [0,1]\}$. Then:

Proposition 4.4.1. *We have*

$$1-\alpha \leq \mathbb{P}\left[f \in C_{1-\alpha}\right] \leq (1-\alpha) + \mathbb{P}\left[f \in \partial C_{1-\alpha}\right]. \tag{4.10}$$

In particular, if $\mathbb{P}\left[f \in \partial C_{1-\alpha}\right] = 0$*, we have* $\mathbb{P}\left[C_{1-\alpha}\right] = 1-\alpha$*.*

As we shall see in Section 4.6, this property is very useful in achieving desired coverage in simultaneous inference problems. When S is the set of n sample functions, the boundary set $\partial C_{1-\alpha}$ is the same as the set of functions in $C_{1-\alpha}$ that equal f_L or f_U (defined in Equation (4.9)) for a part of the domain. The probability $\mathbb{P}\left[f \in \partial C_{1-\alpha}\right]$ may not be exactly zero in finite samples if there are one or more functions $f_i(t)$ which coincide with the upper or lower envelopes of the central region over an interval. However, in most situations of interest, this probability goes to zero as $n \to \infty$.

ED central regions have another interesting and attractive property: there is a close relationship between the ED (simultaneous) regions and the usual pointwise central regions. Specifically, for a fixed $\gamma \in (0,1)$, let $Q_{1-\gamma}$ be the $(1-\gamma)$- pointwise central region given by

$$Q_{1-\gamma} = \{f \in S : q_{\gamma/2}(t) \leq f(t) \leq q_{1-\gamma/2}(t), \forall t \in [0,1]. \tag{4.11}$$

Here $q_\eta(t)$ is the $\eta-$th quantile of the univariate distribution of $\mathbb{P}$ at t. Then, it is shown below that for every $\gamma \in [0,1]$, $Q_{1-\gamma}$ corresponds to an ED central region for some α. Thus, every pointwise central region is an ED central region.

Proposition 4.4.2. *Let* $\mathbb{P}$ *be the stochastic process of interest. For any* $\gamma \in [0,1]$*, there exists an ED central region* $C_{1-\alpha}$ *for some* α *such that* $\mathbb{P}\left[Q_{1-\gamma} \,\Delta\, C_{1-\alpha}\right] = 0$*, where* Δ *denotes set difference. That is, up to a set of* $\mathbb{P}-$*measure zero, the two sets are the same.*

Note that while $Q_{1-\gamma}$ corresponds to an ED central region for each γ, the converse may not be true in general. However, there is indeed a one-to-one correspondence for most continuous stochastic processes. For example, let $X = \{h(t, Y_t)\}, t \in [0,1]$, where Y_t and $h(\cdot,\cdot)$ satisfy the conditions in Proposition 4.3.3.

Corollary 4.4.1. *For every ED central region $C_{1-\alpha}$ of X, there exists $\gamma \in [0,1]$ such that $\mathbb{P}\left[C_{1-\alpha} \,\Delta\, Q_{1-\gamma}\right] = 0$. In particular, all the ED central regions for $Y := \{Y_t, t \in [0,1]\}$ take the form $\{f : -w\sigma(t) \leq f(t) \leq w\sigma(t), \forall t \in [0,1]\}$, for some $w > 0$.*

The last statement of Corollary 4.4.1 implies that ED central regions for the Gaussian process Y have width proportional to the standard deviation, which are perhaps the most natural central regions.

4.4.2 Comparison of Central Regions

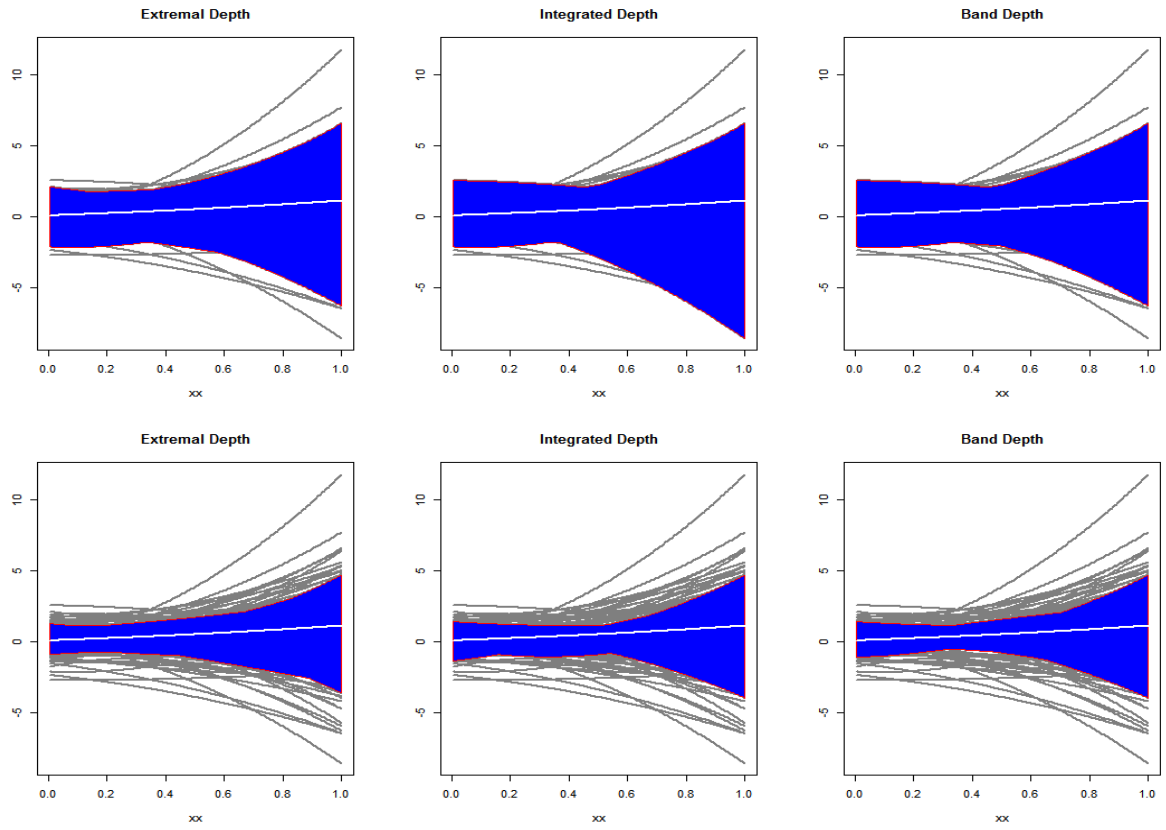
We examine the central regions formed by ED for a simulated dataset and the orthosis dataset considered earlier.

We consider a simple set-up where the functions are random quadratic functions with coefficients drawn from i.i.d. standard normal distributions. That is, each function is randomly generated as $f(t) = c_0 + c_1 t + c_2 t^2$, where $c_0, c_1, c_2 \sim N(0,1)$. We generate $n = 200$ functions at $p = 100$ equally spaced points in $[0,1]$. The central 90% and 50% regions formed by ED, ID and BD are shown in Figure 4.3. Figure 4.4 shows the widths of these regions as function of the pointwise standard deviation. We can see from these figures that the central regions formed by ED represent the central part of the data more appropriately. More specifically, ED central regions are central in the entire domain and more importantly represent the pointwise variability well. The ID and BD central regions however tend to be wider in less variable parts of the domain.

We use the orthosis dataset considered earlier to compare the central regions formed by ED with those from other functional depths. Figure 4.5 compares the 90% (upper panel) and 50% (lower panel) regions formed by ED, ID and MBD. The ID and MBD central regions are defined in a similar way as the ED central regions: the convex hull formed by the deepest $(1-\alpha) \times 100\%$ of the sample functions.

In the upper panel, both ID and MBD regions include the peak (at the top) at

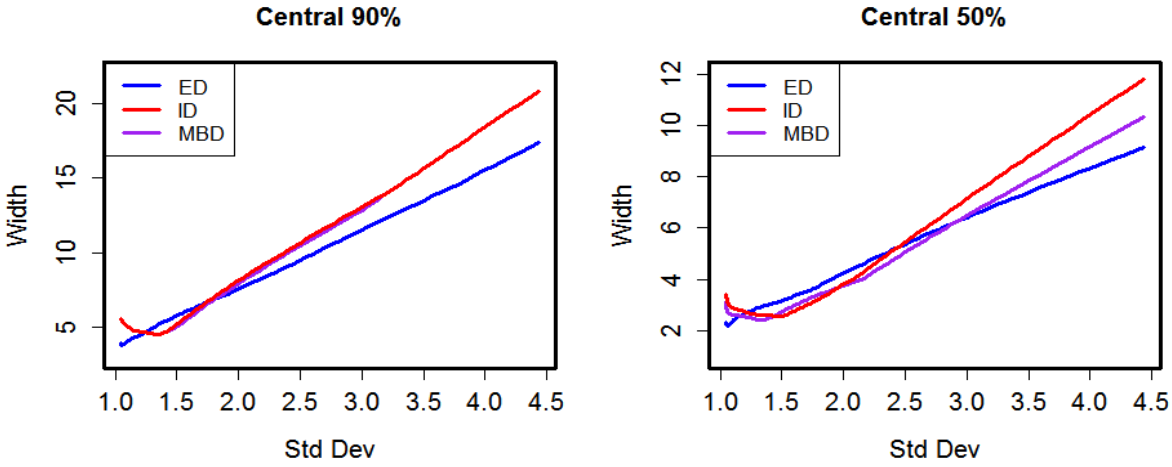
Figure 4.3: Central 90 % and 50 % central regions for the quadratic functions setting



around the value of 180 on x-axis while ED does not. The ID region in the lower panel (50%) also includes some of this peak. Of course, one does not know the "right" answer in this case. However, the connection with pointwise intervals would suggest that behavior of the ED regions is more reasonable.

Figure 4.6 is a plot of the widths of the ED, ID, and MBD central regions against the pointwise standard deviations of the data. We see that the ED central regions scale (approximately) proportionally to the pointwise standard deviations. This is not the case for the regions based on ID or MBD.

Figure 4.4: Central 90 % and 50 % central regions for the quadratic functions setting



4.5 Functional Boxplots and Outlier Detection

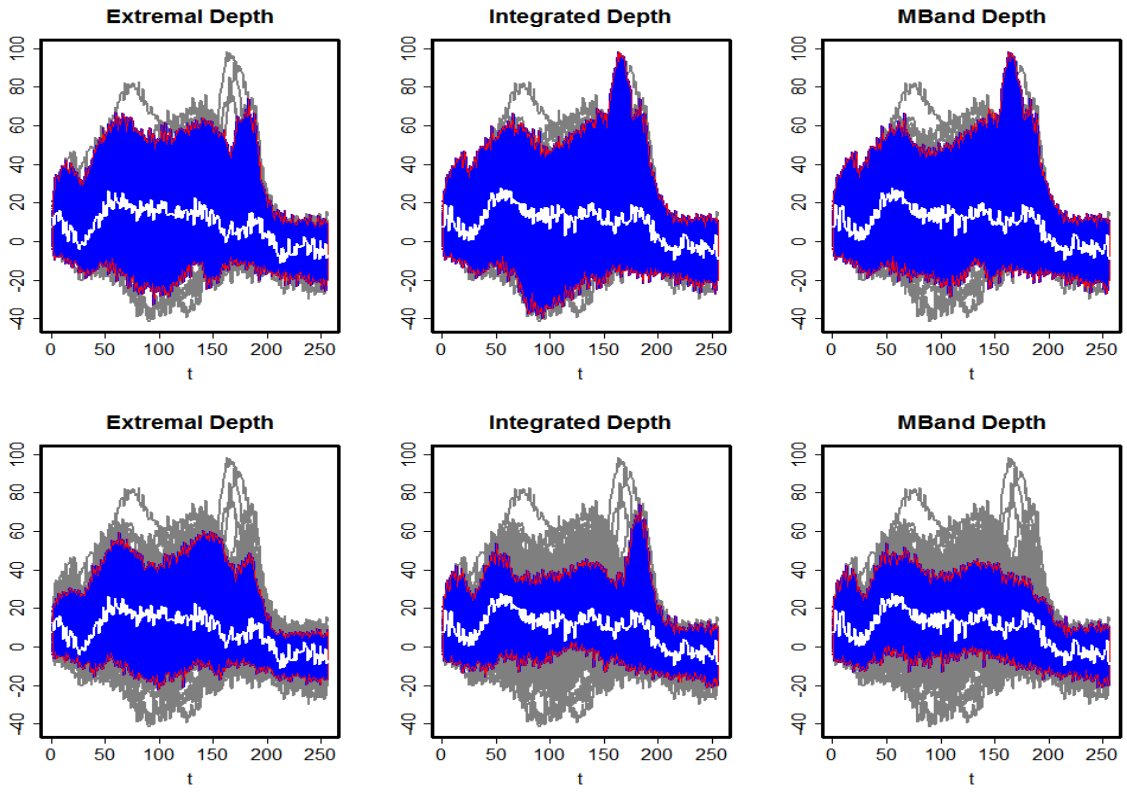
4.5.1 Boxplots

Central regions can be readily used to construct functional boxplots that provide a summary of the data. Sun and Genton (2011) used MBD to develop functional boxplots that are analogous to classical boxplots for univariate data. The plot includes middle 50% central region (the 'box') and an envelope obtained by inflating the central 50% central region by 1.5 times its pointwise range, the boundaries of which are referred to as 'whiskers'. Functions outside this envelope are considered potential candidates for outliers.

We use a simulation study to compare the performance of ED-based functional boxplots to those based on MBD (Sun and Genton (2011)) and ID. The models considered below in our analysis are the same as those in Sun and Genton (2011).

Model 1: *Baseline*: $X_i(t) = 4t + e_i(t), 1 \leq i \leq n$, where $e_i(t)$ is a Gaussian process with mean zero and covariance function $\gamma(s,t) = \exp\{-|t-s|\}$. This is the baseline

Figure 4.5: Central regions of Orthosis data set: 90 % and 50 % central regions in the upper and lowe panels, respectively.



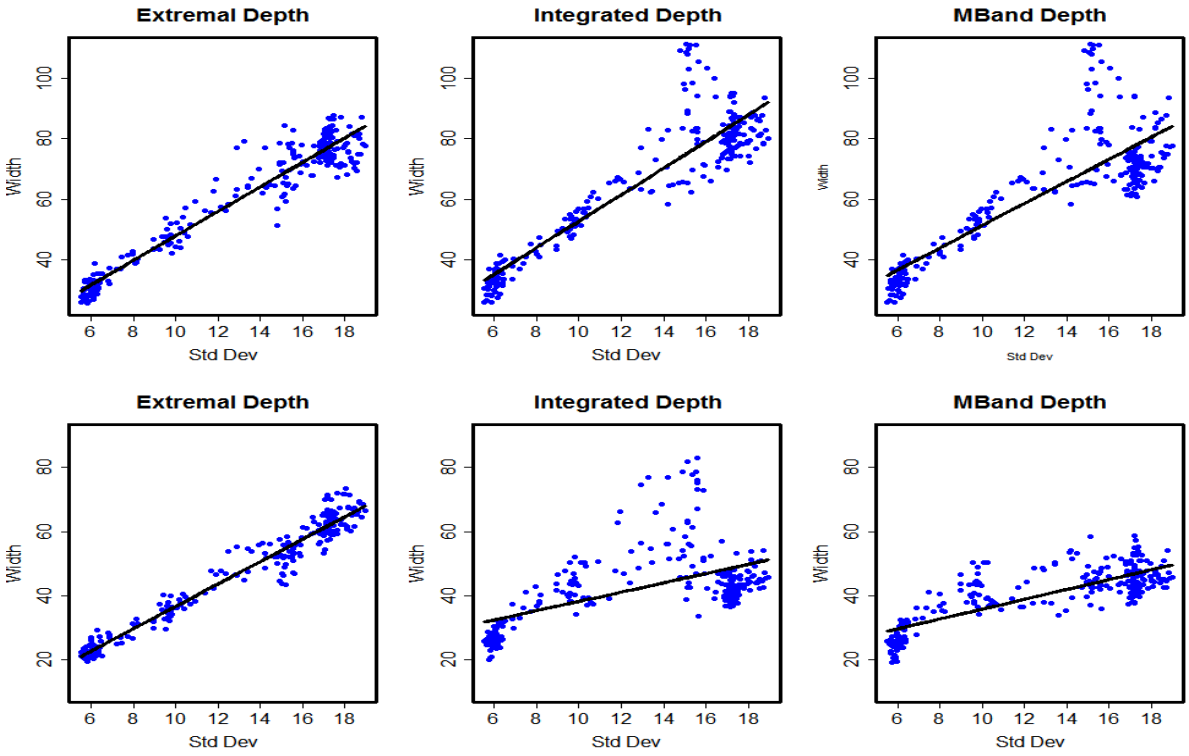
model for the subsequent models.

Models $2-5$ include outliers. Here $\{c_i, i = 1 \leq i \leq n\}$ are indicator functions of outliers and are $i.i.d$ Bernoulli with $p = 0.1$. That is, on average 10% of the observations are outliers. $\{\sigma_i, i = 1 \leq i \leq n\}$ are variables that take on values ± 1 with equal probability and indicate the direction of the outliers; $K = 6$ is the magnitude of the outlier.

Model 2: *Symmetric contamination*: $Y_i(t) = X_i(t) + c_i\sigma_i K$.

Model 3: *Partial contamination*: Let T_i be randomly generated from uniform distribution on $[0, 1]$. Then, $Y_i(t) = X_i(t) + c_i\sigma_i K$, if $t \geq T_i$, and $Y_i(t) = X_i(t)$, if $t < T_i$.

Figure 4.6: Width of the 90 % and 50 % central regions using different approaches: The blue dots are the widths versus standard deviation and the solid black line is the least squares line. It can be seen that the ED has width mostly proportional to the standard deviation while having relatively smaller or comparable width.



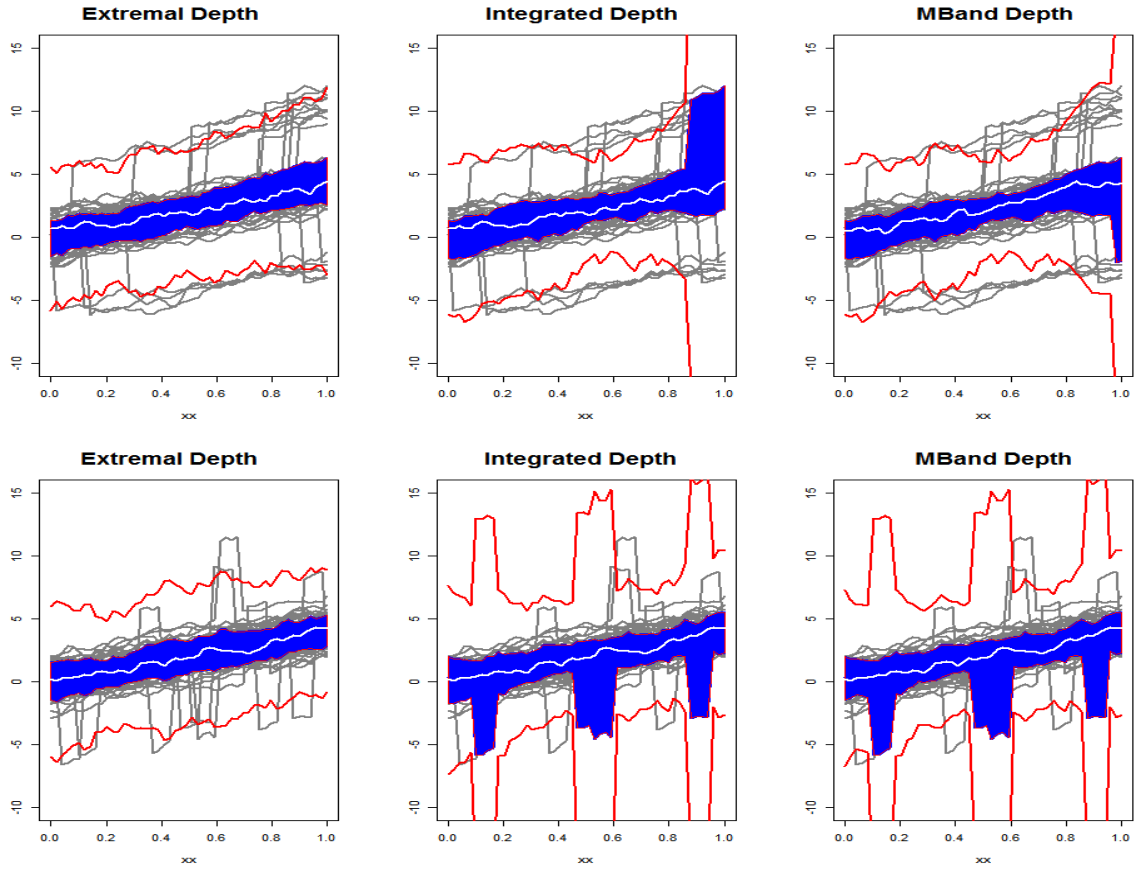
Model 4: *Contaminated by peaks:* Let T_i be randomly generated from uniform distribution on $[0, 1-\ell]$. Then, $Y_i(t) = X_i(t)+c_i\sigma_i K$, if $T_i \leq t \leq T_i+\ell$, and $Y_i(t) = X_i(t)$ otherwise. In the simulation, we fixed $\ell = 0.08$.

Model 5: *Shape contamination with different parameters in the covariance function*: $Y_i(t) = 4t + \tilde{e}_i(t)$, where $\tilde{e}_i$ is a mean zero Gaussian process with covariance $\gamma(s,t) = k\exp\{-|t-s|^{\mu}\}$, with $k = 8, \mu = 0.1$.

For the simulation, we generated $n = 100$ functional observations from the above models and evaluated them on a grid of size 50. Only a summary of the results is given here. For the baseline model with no outliers, all of the depths lead to 'well-behaved'

boxplots. With outliers, ID and MBD-based boxplots exhibited undesirable features, and this was most evident for Models 3 and 4. Figure 4.7 shows a sample dataset. For Model 3 (upper panel), the middle 50% of the central region is affected by the 10% contamination. The problem is less so for MBD but it is still evident. The issue is more serious for Model 4 where the performances of both ID and MBD are badly affected. As noted, part of the reason is that both ID and MBD rely on some type of averaging. The ED plots, which rely on the extremal property, are unaffected by the outliers in these examples.

Figure 4.7: Functional boxplots: The top and bottom panels correspond to data from Models 3 and 4, respectively. In each plot, the region in blue is the central 50% region and the lines in red are the whiskers.



4.5.2 Outlier Detection

This section provides a formal comparison of the performance of boxplots as outlier-detection tools. We use the same measures in Sun and Genton (2011) for comparison:

i) p_c: percentage of correctly identified outliers, and

ii) p_f: percentage of incorrectly detected outliers (equals the number of incorrectly identified outliers divided by total number of non-outlying functions). The standard errors of the percentages are given in parenthesis.

Table 4.1 shows the results based on 100 data sets simulated using the Models 1-5 described above. We see that p_f-values of ED are much lower across all models. The values of p_c are generally similar for the different depth notions except for model 4, where ED outperforms by a clear margin. This is not surprising as model 4 is contaminated by peaks; ID and MBD fail to find the outliers due to their "averaging" property as was evident in Figure 4.7.

These results suggest that when the outlying functions are consistently outlying in the whole domain, all three notions – ED, ID and MBD – perform well. However, when there are functions that are outlying in a subset of the domain as in Models 3 and 4, ED performs better while ID and MBD can do poorly.

Table 4.1: Outlier detection using Functional Box-Plots: p_c is the percentage of correctly identified outliers; p_f is the proportion of incorrectly identified outliers. Numbers in brackets indicate their standard errors.

		ED	ID	MBD
Model 1	p_f	**0.03** (0.17)	0.06 (0.25)	0.07 (0.27)
Model 2	p_c	98.52(4.42)	98.89(3.49)	**99.15**(3.03)
	p_f	**0.01** (0.10)	0.03 (0.20)	0.04 (0.21)
Model 3	p_c	**86.43**(13.64)	77.24 (16.72)	83.17(13.77)
	p_f	**0.01** (0.12)	0.03 (0.18)	0.03 (0.21)
Model 4	p_c	**84.42** (13.29)	41.06 (18.90)	45.94(18.99)
	p_f	**0.01** (0.17)	0.04 (0.21)	0.04 (0.22)
Model 5	p_c	75.74(16.15)	74.97 (16.91)	**78.17** (15.79)
	p_f	**0.01** (0.11)	0.03 (0.19)	0.04 (0.24)

The above discussion indicates that the corresponding estimators, such as functional trimmed means, based on ED will be more resistant to outliers. Specifically, let $m(\alpha)$ is the trimmed mean based on the sample functions in $(1-\alpha)$ ED central region. Then, the simulation results suggest that $m(\alpha)$ may remain bounded even as the outliers increase in magnitude while the corresponding trimmed means for ID and MBD can be unbounded. This result can be established formally and we plan to pursue this in the future.

4.6 Simultaneous Inference

In problems involving functional inference, such as regression and density estimation, it is often difficult to obtain exact simultaneous confidence bands. In such cases, one can combine resampling methods, such as the bootstrap (Efron, 1979), with central regions using functional depth to obtain simultaneous confidence regions. Under the asymptotic validity of the resampling technique, we can get approximate simultaneous confidence regions of desired coverage. This section demonstrates the application for the case of polynomial regression and compares it with other methods.

4.6.1 Polynomial and Other Parametric Regression

Consider the polynomial regression problem $Y(x_i) = \mu(x_i) + \epsilon_i$ with $\mu(x_i) = \theta_0 + \theta_1 x_i + \cdots + \theta_q x_i^q$. The covariates x_i's are fixed and the error terms ϵ_i's are $i.i.d$ with the standard regression assumptions. The goal is to get a simultaneous confidence region for $\mu(x)$ for all x.

It is known that there is no 'exact' method for this general problem. Scheffe's method (Scheffe, 1959) leads to conservative regions since a polynomial of a variable x of degree q does not span the full $(q+1)-$dimensional Euclidean space. The level of conservatism gets higher as the degree q increases. Exact methods have been developed in special cases. Piegorsch (1986) considered quadratic regression and provided

confidence bands sharper than Scheffe's bands. Liu et al. (2014) proposed exact bands for quadratic and cubic polynomial regressions. Wynn (1984) developed exact bands when the errors are normally distributed using special properties of normality. We describe here a general re-sampling based approach using ED central regions.

Let $\theta = (\theta_0, \theta_1, ...\theta_q)$ denote the vector of parameters, $\hat{\theta}$ denote the usual least-squares estimator, and $\hat{\mu}(x)$ be the corresponding predictor. Consider the residuals $r_i = Y(x_i) - \hat{\mu}(x_i)$ and $\hat{s}$, the residual standard error. Generate B bootstrap samples from the residuals to obtain bootstrap estimates $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*$ of θ, and $\hat{s}_1^*, \hat{s}_2^*, \cdots, \hat{s}_B^*$, of σ. Define an estimate of the polynomial mean function $\hat{\mu}(x|\hat{\theta}^*)$ in the obvious manner and the normalized (centered and scaled) version of this function as

$$m_j^*(x) = \frac{\hat{\mu}(x|\hat{\theta}_j^*) - \hat{\mu}(x|\hat{\theta})}{\hat{s}_j^*}, \tag{4.12}$$

for $j = 1, 2, \cdots B$. These are pivotal quantities: their distribution is free of θ and σ. The set of normalized bootstrapped functions $S^* := \{m_1^*, m_2^*, \cdots, m_B^*\}$ can now be treated as our functional data, and they can be used to construct the ED central region. Specifically, let $f_L^*(x)$ and $f_U^*(x)$ be the lower and upper envelopes of this region. Then, the $(1-\alpha)-$level simultaneous confidence band for $\mu(x)$ is given by

$$C_n^\alpha = \{\mu(x) : \hat{\mu}(x) + \hat{s} f_L^*(x) \leq \mu(x) \leq \hat{\mu}(x) + \hat{s} f_U^*(x), \forall x\}. \tag{4.13}$$

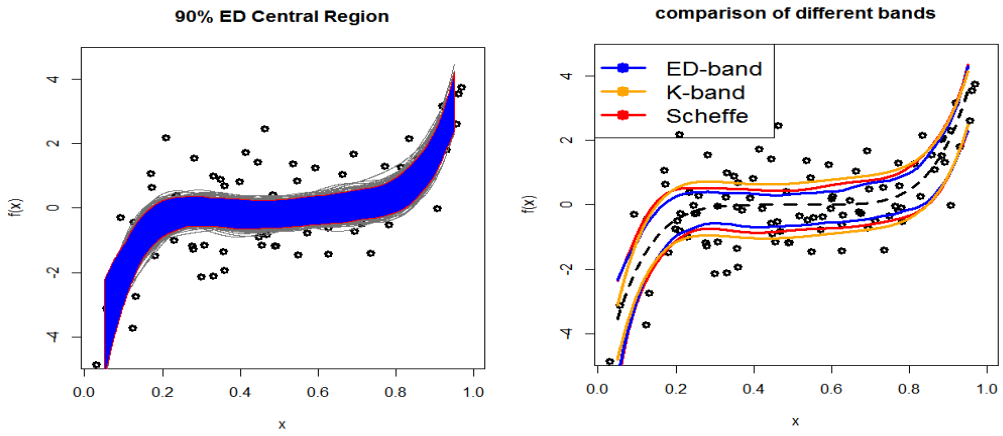
Based on the results in Section 4.4, and the bootstrap validity for parametric regression models (Freedman, 1981), we get $P[\mu(x) \in C_n^\alpha \ \forall \ x] \to (1-\alpha)$ as $n \to \infty$.

We use a limited simulation study to examine the finite sample performance of this band and compare it with bands based on Scheffe's method and a Kolmogorov-like sup-norm statistic. The sup-norm statistic is $K_j^* = \sup_x(|\hat{\mu}_j^*(x) - \hat{\mu}(x)|)/\hat{s}_j^*$. The Scheffe's band is obtained in the usual manner assuming normality. The simulation was done for a degree five polynomial $\mu(x) = 192(x - 0.5)^5$; the coefficient 192 was chosen so that the absolute mean function integrates to one. This is the dashed

function in the right panel of Figure 4.8. We simulated $n = 100$ observations with $i.i.d.$ normal error terms having standard deviation 5; the covariate x was randomly generated from $U[0, 1]$. We used $B = 2000$ bootsrap samples for obtain ED confidence bands.

Figure 4.8 shows the confidence bands and the true mean function (dashed line) for one data set. The confidence band based on ED are tighter than both Scheffe's and K- bands (the band using K_j^*'s). Table 4.2 gives the numerical results from the

Figure 4.8: Simultaneous confidence bands: The figure on the left plots all the bootstrapped functions along with 90 % ED central region and the plot on the right gives confidence bands from the three different methods



simulation study. The first row is the coverage probability and the next five rows show the power values for five different alternative polynomials. The first two alternatives are given by $P_k = C_k\ sign(x-0.5)\ (x-0.5)^k$ for degrees $k = 4, 6$, and C_k is a constant such that $|P_k|$ integrates to one. The next three alternatives are additive shifts from the original mean function P_5.

As expected, the Scheffe-band is very conservative (actual coverage is 99% while the nominal coverage is only 90%). The ED-band has coverage very close to 90% as desired. The coverage proportion of the K-band is close to the nominal. However, the band is wide in the middle and narrow in the tails. This leads to lower power than the ED-band for a large class of alternatives which have shift in the middle of the domain.

This can be seen in last three rows of Table 4.2, where K-band has substantially lower power for the three shift alternatives. Power of ED-bands for the P_4, P_6 alternatives, which mostly differ from K-band at the tails, also remains competitive.

Table 4.2: Level (row 1) and Power (rows 2 - 6) for 90 % simultaneous confidence bands using different methods

	Scheffe	K-band	ED
Level (P_5)	0.01	0.10	0.10
P_4	0.02	0.14	0.16
P_6	0.03	0.17	0.19
$0.2 + P_5$	0.08	0.21	0.32
$0.2 + 0.2x + P_5$	0.31	0.32	0.66
$0.2\ sign(x - 0.5) + P_5$	0.09	0.22	0.38

This application to polynomial regression can be readily extended to more general models of the form $Y_i = \theta_0 + \phi_1(\mathbf{x}_i)\theta_1 + \cdots + \phi_q(\mathbf{x}_i)\theta_q + \epsilon_i$, where $\phi_1, \cdots, \phi_q$ are splines or other known basis functions. The covariates can also be multidimensional in this framework.

4.6.2 Other applications to testing for a distribution, acceptance bands for Q-Q plots and confidence bands for empirical CDF

In this section, we provide a short outline of some potnetial applications of ED in general settings. Extensive studies of these ideas will be considered elsewhere due to space limitation. We first consider another application for obtaining confidence bands for cumulative distribution functions (CDFs). Inference for cumulative distribution functions (CDFs) has a long history. For complete data (no censoring), there are good methods for obtaining simultaneous confidence bands for the CDF in the literature. The ED approach, however, can be useful in situations with complex censoring, and this topic is pursued elsewhere. Our primary goal in this section is to establish the relationship between ED bands for this problem and a class of weighted Kolmogorov bands.

As the distribution of the Kolmogorov statistic $K = \sup_{t \in R^1} |F_N(t) - F(t)|$, under the distribution $F(t)$ does not depend on the underlying distribution and that it can be easily inverted to obtain simultaneous $(1-\alpha)-$ level confidence bands of the following form for F:

$$(F_N(t) - c_N(t), F_N(t) + C_N(t)) \quad \forall\ t \in R^1.$$

Here $C_N(t)$ is the upper $\alpha-$level quantile of the Kolmogorov statistic. The values of $C_N(t)$ have been extensively tabulated in the literature and they can also be obtained through simulation due to the distribution-free property of the statistic. Often, one uses the asymptotic approximation that is also well studied.

The Kolmogorov band has constant width $\pm C_N(t)$ while the variance of the $F_N(t)$ is proportional to $\sqrt{F(t)(1-F(t))}$. So alternative statistics based the use of weighted Kolmogorov statistics have been studied in the literature. These are of the form $K^* = \sup_{t \in E_\delta} |F_N(t) - F(t)|\psi(F(t))$ for some wight function $psi(\cdot)$. The most common examples are: i) $\psi(t) = \frac{1}{\sqrt{t(1-t)}}$, which corresponds to the standard deviation and has been called the Equal Precision (EP) band in the context of censored data by Nair (1984); and ii) $\psi(t) = 1/\sqrt{t}$ or $1/\sqrt{(1-t)}$ are called Renyi bands and give more weight to the lower and upper tails respectively. For technical reasons, E_δ, the set over which the supremum is taken, has to be restricted. For example, for the Equal Precision band with weight $1/\sqrt{t(1-t)}$, in general, the range has to be restricted to $E = t : \delta < F(t) < 1 - \delta$ for some $\delta > 0$. This condition is needed in general to ensure the asymptotic convergence of the statistic. The restriction of E at the tails implies that the statistic is not strictly distribution-free. In practice, it is implemented by applying it to the empirical region $E_N = t : \delta < F_N(t) < 1 - \delta$, i.e., one clips a small proportion of the extreme observations and computes the statistic on the remaining region.

The EP band is attractive because its width is proportional to the pointwise standard error and hence is structurally similar to the pointwise intervals. In this section, we discuss the connections between the EP-band and our ED band based on bootstrapped data. Specifically, let $F_N(t)$ be the empirical CDF (ECDFs) based on N iid observations $X_1, ...X_N$ and let $\mathscr{P}_N$ denote the corresponding distribution. Generate B bootstrap samples from $\mathscr{P}_N$ and denote the estimated ECDFs as $F^*_{N,j},\ j = 1, ..., B$. We now treat the bootstrapped ECDFs as functional data and obtain the $(1-\alpha)-$level ED central region. Let $C^*_{N,\alpha}$ denote this central region based on bootstrapped data. Then, from our results, we have that $P_N(C^*_{N,\alpha}) \geq 1-\alpha$. But we are interested in coverage probability for the underlying CDF $F(t)$. The following lemma shows that it asymptotically has the desired coverage.

Lemma 4.6.1. *We have that* $P[F(t) \in \hat{C}_\alpha(t), \forall t \in [t_0, t_1]] \to (1-\alpha)$, *as* $n \to \infty$.

Proof. We first denote the lower and upper boundaries of $\hat{C}_\alpha(t)$ by L_n and U_n. Due to Theorem 4.4.2 and the fact that conditional on $X_1, \cdots, X_n$, for $t \in [1/n < t < 1-1/n]$ the pointwise distribution of the bootstrap CDF is normal with mean $F_n(t)$ and variance $nF_n(t)(1-F_n(t))$, the ED central region of the bootstrap distribution for $t \in [t_0, t_1]$ is given by $\hat{C}_\alpha = \{F^*_n : \sup_{t\in[t_0,t_1]} \frac{|F^*_n - F_n|}{\sqrt{F_n(1-F_n)}} \leq c_n(F_n)\}$, which has coverage equal to $(1-\alpha)$. Therefore, we have

$$P^*\left[\hat{C}_\alpha = \sup_{t\in[t_0,t_1]} \frac{|F^*_n - F_n|}{\sqrt{F_n(1-F_n)}} \leq c_n(F_n) \mid X_1, \cdots, X_n\right] = 1-\alpha. \tag{4.14}$$

Then following Bickel and Freedman (1981), we also have that

$$P\left[\sup_{t\in[t_0,t_1]} \frac{|F_n - F|}{\sqrt{F_n(1-F_n)}} \leq c_n(F_n)\right] \to 1-\alpha,\ \text{as } n \to \infty.$$

Therefore, the region $F_n \mp c_n(F_n)\sqrt{F_n(1-F_n)}$ is an asymptotically valid simultaneous confidence band for F in $t \in [t_0, t_1]$. Let us now consider the width function of the ED

central region constructed using the bootsrap distribution, i.e., $w_n(t) = U_n - L_n$. By the equal precision lemma, we have that for some $\beta_m(\alpha)$, $w(t) = Q_{\beta(\alpha)}(t) - Q_{-\beta(\alpha)}(t)$. Therefore, for any fixed t, we have $w(t) \to Q_{\beta(\alpha)} - Q_{-\beta(\alpha)} = c(\beta(\alpha))\sqrt{F_n(1-F_n)}$. Therefore, the Bootstrap central region provides $\beta(\alpha)$ that satisfies Equation (4.14). and hence the above argument yields its asymptotic validity. □

Recently Wang et al (2013) proved uniform closeness of a smooth estimator $\tilde{F}_N$ to the true CDF. This result together with the previous arguments provide the validity of bootstrap based on such estimators as well. In future work, we plan to perform extensive empirical studies to understand the finite sample properties of confidence bands using depth based approaches.

We now consider the problem of testing for a distribution of interest. In practice, when it is common to assume a parametric distribution for the data such as the normal distribution. On the other hand, Quantile-Quantile plots are very commonly used as a visual tool to check for deviations from the distributional assumptions. However, Q-Q plots have been used joined with a subjective decision of whether the plot "looks" linear or not. We here propose a quantitative approach for making such a decision using the proposed extremal notion of depth. The idea is to use the central region of level $(1-\alpha)$ from the distribution of Q-Q plots of the hypothesized distribution. The complement of this central region would serve as the critical region with level α, i.e., if a Q-Q plot lies outside this region, we would reject the hypothesized distributional assumption. To obtain this central region for a given sample size n , we would generate B independent data sets of size n from the normal distribution to obtain B Q-Q plots under the null hypothesis. We then use the extremal depth to obtain the $B(1-\alpha)$ deepest functions to form the central region of level $(1-\alpha)$. In the following table, we show the actual level of a 90% acceptance band for normal and GIG distributions using ED and ID based on 100 data sets each with $n = 100$ observations based on boostrap sizes of $B = 2000$. We do not consider BD here as it is computationally

slow if at all infeasible for such problems.

Table 4.3: Level for 90 % Acceptance Bands for Normal Q-Q plot

$n = 100$				
	EP	ED	DID	SD
Level	0.101	0.099	0.007	0.022

Table 4.4: Power for 90 % Acceptance Bands under different alternatives

$n = 100$				
	EP	ED	DID	SD
t_2	0.982	0.982	0.919	0.950
t_5	0.508	0.516	0.192	0.276
Laplace	0.803	0.809	0.462	0.622
Uniform	0.942	0.958	0.498	0.587

As can be seen from Table 4.3, ED achieves desired level whereas ID has level much smaller than the nominal level. ED provides an alternative to the classical Kolmogorov test, which can be viewed as having a constant width acceptance region based on the empirical CDF. As we argued earlier, ED acceptance regions will have width depending on the pointwise variation. For this reason, we see that ED based bands have better performance in terms of having higher power compared to alternative methods.

In summary, we have developed a new notion of functional depth, studied its properties, and demonstrated its usefulness through several applications. While no single notion of functional depth will do uniformly better than others, we hope that the results here suggest that the extremal-depth concept has many attractive properties and is a useful tool for exploratory analysis of functional data and has potential for applications in many problems including estimation and simultaneous inference for functional parameters.

4.7 Proofs

Condition and proof for Proposition 4.3.1:

Condition 4.7.1. *Assume that (a)* $\mathbb{P}[d_f = 0] = 0$, *and (b)* $\mathbb{P}[d_f = d_g, f \neq g] = 0$, *where* f, g *are independent random functions from* $\mathbb{P}$ *and* $d_f := \inf\{r \in [0,1] : \Phi_f(r) > 0\}$.

Proof of Proposition 4.3.1: We shall show that, if $ED(f_1, \mathbb{P}), ED(f_2, \mathbb{P}) \geq \alpha$, and $f_1(t) \leq f(t) \leq f_2(t)\ \forall t \in [0,1]$, then $ED(f, \mathbb{P}) \geq \alpha$. Note that $\forall t$, $D_f(t, \mathbb{P}) \geq \min(D_{f_1}(t, \mathbb{P}), D_{f_2}(t, \mathbb{P}))$, and hence $d_f \geq \min(d_{f_1}, d_{f_2})$. Therefore either $f \succeq f_1$ or $f \succeq f_2$ w.p.1 and $f \in C_\alpha$ due to Condition 4.7.1 (b). □

Condition 4.7.1 is a mild condition on $\mathbb{P}$. For instance, this holds if $\mathbb{P}$ is the distribution of X in Proposition 4.3.3.

NAB property:

Denote the pointwise quantile functions of $\mathbb{P}$ as q_α, i.e., for each t, $\mathbb{P}[X(t) < q_\alpha(t)] \leq \alpha$, and $\mathbb{P}[X(t) \leq q_\alpha(t)] \geq \alpha$ (for uniqueness we take the smallest one). Let for $\alpha_n \downarrow 0$, $f_n(t) \leq q_{\alpha_n}(t), \forall t \in U$, and for $\beta_n \uparrow 1$, $g_n(t) \geq q_{\beta_n}(t), \forall t \in U$, where U is some open interval in $[0,1]$. Then we say the depth notion D to have NAB property if $D(f_n, X) \to 0$ and $D(g_n, X) \to 0$.

We now show that ED satisfies NAB under Condition 4.7.1 (a). Since $f_n(t) \leq q_{\alpha_n}(t), \forall t \in U$, we have $\forall n \geq N$,

$$\begin{aligned} \mathbb{P}[f_{n+1} \succeq X] &\leq 1 - \mathbb{P}[q_{\alpha_n} < X < q_{1-\alpha_n}] \\ &= 1 - \mathbb{P}[\cup_{k \leq n}\{q_{\alpha_k} < X < q_{1-\alpha_k}\}]. \end{aligned}$$

Therefore, $\limsup \mathbb{P}[f_{n+1} \succeq X] \leq 1 - \mathbb{P}[\Omega := \cup_{1<k<\infty}\{q_{\alpha_k} < X < q_{1-\alpha_k}\}] = 0$, as the set Ω has probability one due to Condition 4.7.1 (a). Therefore, $D(f_n, X) \to 0$ and similarly $D(g_n, X) \to 0$.

Conditions and proof of Proposition 4.3.2:

Condition 4.7.2. *Let C_n be the total number of functional crossings by any pair of functions, where n is the number of sample functions. We assume that $C_n = \exp\{o_P(n)\}$.*

Condition 4.7.3. *Let $\mathbb{P}$ be a stochastic process on $C[0,1]$ whose univariate CDF at $t \in [0,1]$ is denoted by F_t. Define $R(\delta, u) = \sup_{|t-s|<\delta} |F_t(u) - F_s(u)|$. Then we assume that for any u_0, there is a neighborhood $B(u_0, \epsilon)$ such that $R(\delta_n, u) \to 0$ uniformly in $u \in B(u_0, \epsilon)$ as $\delta_n \to 0$. Further, we assume $\mathbb{P}$ to have Glivenko-Cantelli (GC) property uniformly over convex sets.*

Condition 4.7.2 assumes the number of crossings is at most exponential in sample size, and is related to the smoothness of the process. Condition 4.7.3 assumes that the CDF's of neighboring points in the domain are close. The GC property of $\mathbb{P}$ requires that the empirical distributions corresponding to $\mathbb{P}$ converge uniformly over convex sets. GC property for convex sets holds under general conditions for finite dimensional distributions (Eddy and Hartigan, 1977). Chakraborty and Chaudhuri (2014b) provides a GC type result for spatial distributions of infinite-dimensional spaces. For our result, we assume this as a technical condition.

Let $f \succeq_n g$ and $f \succeq g$ denote that f is deeper than or equal to g using ED w.r.t. the empirical distribution $\mathbb{P}_n$ and the true distribution $\mathbb{P}$, respectively. Then,

$$\begin{aligned}
\sup_{f \in C[0,1]} |ED(f, \mathbb{P}_n) - ED(f, \mathbb{P})| &= \sup_{f \in C[0,1]} |\mathbb{P}_n [f \succeq_n X_n] - \mathbb{P} [f \succeq X] | \\
&= \sup_{f \in C[0,1]} |\mathbb{P}_n [f \succeq_n X_n] - \mathbb{P} [f \succeq_n X] | \\
&\quad + \sup_{f \in C[0,1]} |\mathbb{P} [f \succeq_n X] - \mathbb{P} [f \succeq X] |, \qquad (4.15)
\end{aligned}$$

where $X_n \sim \mathbb{P}_n$ and $X \sim \mathbb{P}$.

The first term in RHS of (4.15) can be shown to go to zero because of the Glivenko-Cantelli (GC) property assumed by Condition 4.7.3. That is because the sets $\{f \succeq_n X\}$ are convex and the GC type result holds over all convex subsets. We then only need to show that $\sup_f |\mathbb{P} [f \succeq_n X] - \mathbb{P} [f \succeq X] | \to 0$.

We will now show that the second term in RHS of (4.15) goes to zero. Define $d_f := \inf_{y\in[0,1]}\{\Phi_f(y,\mathbb{P})>0\}$, and $d_f^n := \inf_{y\in[0,1]}\{\Phi_f(y,\mathbb{P}_n)>0\}$. We shall first show that $\sup_f |d_f^n - d_f| \xrightarrow{\text{P}} 0$ as $n\to\infty$.

Due to the rate of Glivenko-Cantelli of empirical distributions (Pollard, 1991), we have for any t,

$$\mathbb{P}\left[\sup_u |F_t^n(u) - F_t(u)| > \epsilon\right] \leq \exp\{-c\epsilon^2 n\}. \tag{4.16}$$

Let $D = \{d_1, d_2, \cdots,\}$ be a countable dense subset of $[0,1]$. Define $T_n = \{t_1, \cdots, t_{k_n}\}$ be the set containg all the points in $[0,1]$ where n sample functions cross and along with $\{d_1, d_2, \cdots, d_n\}$. As $n\to\infty$ we have $T := \cup_n T_n$ is the union of all the crossing points and D. Due to Condition 4.7.2, we have $\log|k_n| = o_P(n)$.

Due to Equation (4.16), we have

$$\mathbb{P}\left[\sup_{t\in T_n}\sup_u |F_t^n(u) - F_t(u)| > \epsilon\right] \leq \exp\{-c'\epsilon^2 n + \log k_n\}. \tag{4.17}$$

Now, note that $d_f^n = \inf_{t\in[0,1]} D_f(t,\mathbb{P}_n) = \min_{t\in T_n} D_f(t,\mathbb{P}_n)$, because the univariate depths in T_n have the same range as that of the whole interval $[0,1]$. We shall first show that

$$d_f = \inf_{t\in T} D_f(t,\mathbb{P}) = \lim_n \inf_{t\in T_n} D_f(t,\mathbb{P}), \tag{4.18}$$

using the facts that $\cup_n T_n = T$, T is dense and Condition 4.7.3. To see this, first note that $d_f \leq \inf_{t\in T} D_f(t,\mathbb{P})$. For the reverse inequality, consider a y_0 such that $D_{y_0}(f,\mathbb{P}) = d_f$ (this exists due to continuity of F_t in t). Since T is dense, we have a sequence $y_n \in T$ such that $y_n \to y_0$. Due to continuity of F and Condition 4.7.3, we have

$$\begin{aligned}
|D_{y_n}(f,\mathbb{P}) - D_{y_0}(f,\mathbb{P})| &= ||1-2F_{y_n}(f(y_n))| - |1-2F_{y_0}(f(y_0))|| \\
&\leq 2|F_{y_n}(f(y_n)) - F_{y_0}(f(y_0))| \\
&\leq 2|F_{y_n}(f(y_n)) - F_{y_0}(f(y_n))| + 2|F_{y_0}(f(y_n)) - F_{y_0}(f(y_0))| \to 0,
\end{aligned}$$

which implies (4.18). Now, using (4.17), we have

$$\mathbb{P}[\sup_f |d_f^n - d_f| > \epsilon_n] \leq \mathbb{P}\left[\sup_{t\in T_n}\sup_u |F_t^n(u) - F_t(u)| > \epsilon_n/4\right] \to 0,$$

if $\epsilon_n \to 0$ and $c'\epsilon_n^2 n - \log k_n \to \infty$. In particular, when $\epsilon_n = 4\max((3\log k_n/c'n)^{1/2}, 1/\sqrt{\log n})$, $\mathbb{P}[\sup_f |d_f^n - d_f| > \epsilon_n] < Cn^{-1-\epsilon}$, for some $C, \epsilon > 0$. Then using Borel-Cantelli lemma, we obtain $\sup_f |d_f^n - d_f| \to 0$ almost surely. Now, consider the events $A_n = \{d_f^n \geq d_g^n\}$ and $B_m = \{d_f < d_g - \delta_m\}$, where $\delta_m \to 0$ as $m \to \infty$. Note that A_n and B_m depend on the functions f and g. Then, $\mathbb{P}[\cup_{f,g} A_n \cap B_m] \leq \mathbb{P}[\sup_h |d_h^n - d_h| > \epsilon_m] \to 0$ as $n \to \infty$. Therefore, we have

$$\begin{aligned} \limsup_n \sup_f |\mathbb{P}[f \succeq_n X] - \mathbb{P}[f \succeq X]| &\leq \limsup_n \mathbb{P}[\cup_f \{f \succeq_n X\} \Delta \{f \succeq X\}] \\ &\leq \limsup_n \lim_m \mathbb{P}[\cup_{f,g} A_n \cap B_m] \\ &\leq \lim_m \limsup_n \mathbb{P}[\cup_{f,g} A_n \cap B_m] = 0. \quad \square \end{aligned}$$

Proof of Proposition 4.3.3:

As the process $Y = \{Y_t\}, t \in [0,1]$ has continuous sample paths, the sample paths of the process $X = \{X_t\}, t \in [0,1]$ also lie in $C[0,1]$ almost surely. Due to the monotone invariance property of ED, we only need to show that ED of Y takes all the values in $[0,1]$. Consider the sets $Q_{1-\gamma} := \{f : q_{\gamma/2}(t) \leq f(t) \leq q_{1-\gamma/2}(t), \forall t\}$, for $\gamma \in [0,1]$, where q_α is the α-th pointwise quantile of Y. Note that $q_{\gamma/2} \preceq f$, for any $f \in Q_{1-\gamma}$ and $q_{\gamma/2} \succ g$, for $g \in Q_{1-\gamma}^c$. Therefore, $ED(q_{\gamma/2}, \mathbb{P}) = \mathbb{P}[Q_{1-\gamma}]$. By noting that $Q_{1-\gamma} = \{f : \sup_t |f(t)/\sigma(t)| \leq c\}$ and that $\sup_t |f(t)/\sigma(t)|$ has a continuous distribution, $\mathbb{P}[Q_{1-\gamma}]$ takes all the values in $(0,1]$. $\square$

Proof of Proposition 4.4.1:

To prove the lower bound, consider a function g having ED equal to α. Then, $f \succeq g$, for $g \in C_{1-\alpha}^C$ with $\mathbb{P}$-probability one.This implies that $\alpha = \mathbb{P}[X : g \succeq X] \geq \mathbb{P}[C_{1-\alpha}^C]$, and hence $\mathbb{P}[C_{1-\alpha}] \geq (1-\alpha)$.

To prove the upper bound, we first note that the set $\{f : g \succeq f\}$ is contained in the union of the sets $C_{1-\alpha}^C$ and $\partial C_{1-\alpha}$. This is because, for any function $h \in C_{1-\alpha} - \partial C_{1-\alpha}$, $d_{min}(h, \mathbb{P}) > d_{min}(g, \mathbb{P})$, where $d_{\min}(h, \mathbb{P}) = \inf_{t \in [0,1]} D_h(t, \mathbb{P})$ as in Section 4.2. Otherwise, we have a function f with ED larger than α but $d_{min}(f, \mathbb{P}) < d_{min}(g, \mathbb{P})$, which is a contradiction. Therefore, $\alpha = \mathbb{P}[X : g \succeq X] \leq \mathbb{P}[C_{1-\alpha}^C \cup \partial C_{1-\alpha}]$. This

implies that $\mathbb{P}[C_{1-\alpha} - \partial C_{1-\alpha}] \leq (1-\alpha)$ and $\mathbb{P}[C_{1-\alpha}] \leq (1-\alpha) + \mathbb{P}[\partial C_{1-\alpha}]$, and the result follows. □

Proof of Proposition 4.4.2 & Corollary 4.4.1:

We shall show that the ED central region C^* formed by the functions $\{f : ED(f, \mathbb{P}) \geq ED(q_{\gamma/2}, \mathbb{P})\}$ proves the proposition. Although this central region is not in the form defined by Equation (4.8) (due to "$\geq$" instead of a "$>$"), this does not make a difference when $\mathbb{P}$ is a continuous stochastic process, and this same set can be written with a "$>$" when $\mathbb{P}$ is an empirical distribution. We have $f \succeq q_{\gamma/2} \sim q_{1-\gamma/2} \succ g$, for any $f \in Q_{1-\gamma}$, and $g \in Q^C_{1-\gamma}$. Therefore, $Q_{1-\gamma} \subset C^*$ and it remains to show that $\mathbb{P}[C^* - Q_{1-\gamma}] = 0$. However, $C^* - Q_{1-\gamma} \subset B := \{f \notin Q_{1-\gamma} : ED(f, \mathbb{P}) = ED(q_{\gamma/2}, \mathbb{P})\}$. As all the functions in B have the same ED, $\mathbb{P}[f \in B] = \mathbb{P}[f \in B : f \sim q_{\gamma/2}] = 0$. Therefore, $\mathbb{P}[C^* \Delta Q_{1-\gamma}] = 0$. The corollary follows directly because ED is a decreasing function of $\sup_t |f(t)|/\sigma(t)$. □

CHAPTER V

Future Work

5.1 Future Work on Bayesian Methods, Computation, and Inference for High Dimensional Data

In spite of the rapid developments in statistical methods for high dimensional data, some fundamental questions still remain open, particularly in the understanding of Bayesian methods and algorithms. A few topics of future interest are:

(i) inference on model parameters after selection is a much needed step ahead for high dimensional data. Although a Bayesian method gives a posterior distribution that may in principle be used for inference, coverage properties of such posterior intervals need to be studied together with new methods to deal with potential incorrect coverage issues.

(ii) optimal prior choice: the results from this thesis also suggest that model selection consistency holds for a large class of prior distributions if the prior parameters are sample size dependent. However, the mixing properties and complexity of the corresponding sampling algorithms depend on the prior choice. This motivates study of computational complexity of sampling algorithms, and optimal prior choices in the sense of minimizing statistical and sampling error together in the context of Bayesian modeling.

(iiii) nonconvex objective function: when the objective function under consideration is non-convex, optimization methods tend to have severe difficulties where sampling approaches in the Bayesian framework have a natural advantage. As an example, non-vexity of the objective function occurs in censored quantile regression using Powell's (Powell, 1984, 1986) objective function. Although conceptually appealing, quantile regression for censored data is challenging due to both computational and theoretical difficulties arising from non-convexity of the objective function involved. By considering exponentiated Powell's objective function as a working likelihood, the theoretical and computational strategies proposed in the current thesis have the potential to provide a better alternative in this case and more broadly for dealing with nonconvex objective functions.

5.2 Future Work on Applications of Extremal Depth to Simultaneous Inference for Functional Data

Functional depth notions can be very useful for estimation and inference of functional parameters. More specifically, extremal depth approach can be used to obtain simultaneous confidence bands together with resampling methods in problems such as density estimation, survival function estimation, and regression. However, the justification of the ED-based regions in general function estimation problems will depend on the limiting distributions of the functional estimators and the asymptotic validity of the bootstrap. Simulation results in finite samples suggest that the convergence of the actual level to the nominal one might be slow in fully nonparametric inference problems. A more extensive study is needed to understand the behavior, both theoretically and empirically.

This approach of data depth together with resampling can be also extended to other problems. These include the goodness-of-fit testing problem where one wants to

determine if the generative model belongs to a certain parametric family of distributions. One can combine the bootstrapping technique (parametric or nonparametric) with ED central regions to construct acceptance or confidence regions. (See Cuevas et al. (2006), Yeh (1996) and Yeh and Singh (1997) for some related discussion.) While this is a classical problem, our initial studies suggest that the ED-based approach has some advantages over methods based on weighted Kolmogorov statistics.

Another important class of problems deal with the case where the underlying functions of interest are observed with error. In other words, instead of observing random functions $X_i(t)$ from a generative model of interest, we observe $Y_i(t) = X_i(t) + \epsilon_i(t),\ i = 1, \cdots, n$. A natural approach is to use some type of smoother to 'recover' $X_i(t)$ and then use the techniques discussed so far. If there is some information of the error structure in $\epsilon(t)$, this can be used to guide the smoothing algorithm or the 'reconstruction' methods for $X_i(t)$. We plan to conduct further research in these directions.

BIBLIOGRAPHY

BIBLIOGRAPHY

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32:870–897, 2004.

A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson. Dirichlet laplace priors for optimal shrinkage. *Journal of the American Statistical Association, To appear*, 2014.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

F. V. Bonassi and M. West. Sequential monte carlo with adaptive weights for approximate bayesian computation. *Bayesian Analysis*, 10:171–187, 2015.

W. Bonassi, G. Reeves, and D.B. Dunson. Scalable approximations of marginal posteriors in variable selection. *http://arxiv.org/pdf/1506.06629v1.pdf*, 2015.

H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. *Biometrics*, 64:115–123, 2008.

H. D. Bondell and B. J. Reich. Consistent high dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107: 1610–1624, 2012.

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

B. M. Brown. Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B*, 45:25–30, 1983.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg, 2011.

V. Cahouet, L. Martin, and D. Amarantini. Static optimal estimation of joint accelerations for inverse dynamic problem solution. *Journal of Biomechanics*, 35: 1507–1513, 2002.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, 35:2313–2351, 2007.

A. Chakraborty and P. Chaudhuri. On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66:303–324, 2014a.

A. Chakraborty and P. Chaudhuri. The spatial distribution in infinite dimensional spaces and related quantiles and depths. *Annals of Statistics*, 42(3):1203–1231, 2014b.

A. Chakraborty and P. Chaudhuri. The deepest point for distributions in infinite dimensional spaces. *Statistical Methodology*, 20:27–39, 2014c.

P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91:862–872, 1996.

J. Chen and Z. Chen. Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*, 22:555–574, 2012.

M. H. Chen, L. Huang, J. G. Ibrahim, and Sungduk. Kim. Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis*, 3:585–614, 2008.

A. Cuevas, M. Febrero, and R. Fraiman. On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51: 1063–1074, 2006.

T. Dey, H. Ishwaran, and J. S. Rao. An in-depth look at highest posterior model selection. *Econometric Theory*, 24:377–403, 2008.

L. Dicker, B. Huang, and X. Lin. Variable selection and estimation with the seamless-l_0 penalty. *Statistica Sinica*, 23:929–962, 2013.

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 20:page, 1992.

W. F. Eddy and J.A. Hartigan. Uniform convergence of the empirical distribution function over convex sets. *Annals of Statistics*, 5:370–374, 1977.

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7: 1–26, 1979.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70:849–911, 2008.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928–961, 2004.

R. Fraiman and G. Muniz. Trimmed means for functional data. *Test*, 10:419–440, 2001.

D. A. Freedman. Bootstrapping regression models. *Annals of Statistics*, 9:1218–1228, 1981.

J. H. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 2008.

E. I. George and D. P. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87:731–747, 2000.

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

A. K. Ghosh and P. Chaudhuri. On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, 11:1–27, 2005.

Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5: 1780–1815, 2011.

C. Hans, A. Dobra, and M. West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102:507–516, 2007.

C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1:145–168, 2006.

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

J. Huang and H. Xie. Asymptotic oracle properties of scad-penalized least squares estimators. *IMS Lecture Notes -Monograph Series*, 55:149–166, 2007.

J. Huang and C. H. Zhang. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13:1839–1864, 2012.

M. Hubert, P. Rousseeuw, and P. Segaert. Multivariate functional outlier detection. *Statistical Methods and Applications*, page To appear, 2015.

Y. Imai, H. R. Patel, N. M. Doliba, F. M. Matschinsky, J. W. Tobias, and R. S. Ahima. Analysis of gene expression in pancreatic islets from diet-induced obese mice. *Physiol Genomics*, 36:43–51, 2008.

H. Ishwaran and J.S. Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *Annals of Statistics*, 33:730–773, 2005.

H. Ishwaran and J.S. Rao. Consistency of spike and slab regression. *Statistics and Probability Letters*, 81:1920–1928, 2011.

H. Ishwaran, U.B. Kogalur, and J.S. Rao. spikeslab: Prediction and variable selection using spike and slab regression. *The R Journal*, 2:68–73, 2010.

G. M. James, P. Radchenko, and J. Lv. Dasso: connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society, Series B*, 71:127–142, 2009.

W. Jiang. Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Annals of Statistics*, 35:1487–1511, 2007.

V. E. Johnson and D. Rossell. On numerical aspects of bayesian model selection in high and ultrahigh-dimensional settings. *Journal of the American Statistical Association*, 107:649–660, 2012.

R. Jörnsten. Clustering and classification based on the l_1 data depth. *Journal of Multivariate Analysis*, 90:67–89, 2004.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

K. Khare and J.P. Hobert. Geometric ergodicity of the bayesian lasso. *Electron. J. Statist.*, 7:2150–2163, 2013.

Y. Kim, S. Kwon, and H. Choi. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13:1037–1057, 2012.

S. K. Kinney and D. B. Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63:690–698, 2007.

H. Lan, M. Chen, J. B. Flowers, B. S. Yandell, D. S. Stapleton, C. M. Mata, E. T. Mui, M. T. Flowers, K. L. Schueler, K. F. Manly, R. W. Williams, K. Kendziorski, and A. D. Attie. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2:e6, 2006.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28:1302–1338, 2000.

J. Li, J. A. Cuesta-Albertos, and R. Y. Liu. Dd-classifier: Nonparametric classification procedure based on dd-plot. *Journal of the American Statistical Association*, 107: 737–753, 2012.

F. Liang, C. Liu, and R.J. Carroll. Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association*, 102:305–320, 2007.

F. Liang, Q. Song, and K. Yu. Bayesian subset modeling for high dimensional generalized linear models. *Journal of the American Statistical Association*, 108:589–606, 2013.

R. Y. Liu. On a notion of data depth based on random simplices. *Annals of Statistics*, 18:405–414, 1990.

R. Y. Liu and K. Singh. A quality index based on data depth and multivariate rank test. *Journal of the American Statistical Association*, 88:257–260, 1993.

R. Y. Liu and K. Singh. Notions of limiting p values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92:266 – 277, 1997.

R. Y. Liu, J. M. Parelius, and K. Singh. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, 27:783–858, 1999.

W. Liu, S. Zhou, and F. Bretz. Exact simultaneous confidence bands for quadratic and cubic polynomial regression with applications in dose response study. *Australian and New Zealand Journal of Statistics*, 55:421–434, 2014.

Y. Liu and Y. Wu. Variable selection via a combination of the l_0 and l_1 penalties. *Journal of Computational and Graphical Statistics*, 16:782–798, 2007.

S. Lopéz-Pintado and J. Romo. Depth-based classification for functional data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. American Mathematical Society.*, 72:103 – 120, 2006.

S. Lopéz-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104:718–734, 2009.

S. Lopéz-Pintado and J. Romo. A half-region depth for functional data. *Computational Statistics and Data Analysis*, 55:1679–1695, 2011.

P.C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83:1023–1036, 1988.

E. Moreno, F.J. Giron, and G. Casella. Consistency of objective bayes factors as the model dimension grows. *Annals of Statistics*, 38:1937–1952, 2010.

K Mosler. *Multivariate Dispersion, Central Regions and Depth.* Springer, New York, 2002.

K. Mosler and Y. Polyakov. General notions of functional depth. *Discussion Papers in Statistics and Econometrics 2/2012, University of Cambridge*, page arXiv:1208.1981v1, 2012.

N. Nair, V. Confidence bands for survival functions with censored data: A comparative study. *Technometrics*, 26:265–275, 1984.

N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics*, 42:789–817, 2014.

N. N. Narisetty and V. N. Nair. Extremal depth for functional data and applications. *Journal of the American Statistical Association (Theory & Methods), to appear*, 2016.

N. N. Narisetty, J. Shen, and X. He. Scalable and consistent variable selection for high dimensional logistic regression. *Submitted*, 2016.

S. M. O'Brien and D. B. Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60:739–746, 2004.

R. B. O'hara and M. J. Sillanpaa. A review of bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4:85–117, 2009.

M. Y. Park and T. Hastie. L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B*, 69:659–677, 2007.

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

W.W. Piegorsch. Confidence bands for polynomial regression with fixed intercepts. *Technometrics*, 28:241–246, 1986.

D. Pollard. *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, 1991.

J. L. Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25:303 – 325, 1984.

J. L. Powell. Censored quantile regression. *Journal of Econometrics*, 32:143 – 155, 1986.

J. C. Román and J. P. Hobert. Convergence analysis of the gibbs sampler for bayesian general linear mixed models with improper priors. *Annals of Statistics*, 40:2823–2849, 2012.

J. C. Román and J. P. Hobert. Geometric ergodicity of gibbs samplers for bayesian general linear mixed models with proper priors. *Linear Algebra and its Applications*, 2015.

V. Ročková and E. I. George. EMVS: The em approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109:828–846, 2014.

H. Scheffe. *The Analysis of Variance*. Wiley, New York, 1959.

Gideon E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978.

R. Serfling. A depth function and a scale curve based on spatial quantiles. *Statistics for Industry and Technology (ed. Y.Dodge), Birkhaeuser*, pages 25–38, 2002.

X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107:223–232, 2012.

J. V. Silha, M. Krsek, P. Sucharda, and L. J. Murphy. Angiogenic factors are elevated in overweight and obese individuals. *International Journal of Obesity*, 29:1308–1314, 2005.

L. A. Stefanski. A normal scale mixture representation of the logistic distribution. *Statistics and Probability Letters*, pages 69–70, 1991.

Y. Sun and M.C. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20:316–334, 2011.

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

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

J. Tukey. Mathematics and the picturing of data. *In Proc. 1975 Inter. Cong. Math., Vancouver; Montreal: Canad. Math. Congress*, pages 523–531, 1975.

van de Geer S. A. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36:614–645, 2008.

Y. Vardi and C. Zhang. The multivariate l_1 median and associated data depth. *Proceedings of the National Academy of Science USA*, 97:1423–1426, 2000.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed sensing, Cambridge University Press, Cambridge*, pages 210–268, 2012.

A. A. Wendel, L. O. Li, Li Y., G. W. Cline, G. I. Shulman, and R. A. Coleman. Glycerol-3-phosphate acyltransferase 1 deficiency in ob/ob mice diminishes hepatic steatosis but does not protect against insulin resistance or obesity. *Diabetes*, 59: 1321–1329, 2010.

H. P. Wynn. An exact confidence band for one-dimensional polynomial regression. *Biometrika*, 71:375–379, 1984.

Y. Yang and X. He. Bayesian empirical likelihood for quantile regression. *Annals of Statistics*, 40:1102–1131, 2012.

Y. Yang, M. J. Wainwright, and M. I. Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *Annals of Statistics*, 2016.

A. B. Yeh. Bootstrap percentile confidence bands based on the concept of curve depth. *Communications in Statistics - Simulation and Computation*, 25:905–922, 1996.

A. B. Yeh and K. Singh. Balanced confidence regions based on tukey's depth and the bootstrap. *Journal of the Royal Statistical Society. Series B*, 59:639–652, 1997.

M. Yuan and Y. Lin. Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100:1215–1225, 2005.

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

D. Zhang, Y. Lin, and M. Zhang. Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3:781–796, 2009.

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

Y. Zuo. Projection-based depth functions and associated medians. *Annals of Statistics*, 31:1460–1490, 2003.

Y. Zuo and R. Serfling. General notions of statistical depth function. *Annals of Statistics*, 28:461–482, 2000.