Bayesian model selection consistency for high-dimensional regression

by

Min Hua

B.A., Jiangxi University of Finance and Economics, China, 2009

M.S., Hangzhou Dianzi University, China, 2013

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Abstract

Bayesian model selection has enjoyed considerable prominence in high-dimensional variable selection in recent years. Despite its popularity, the asymptotic theory for high-dimensional variable selection has not been fully explored yet. In this study, we aim to identify prior conditions for Bayesian model selection consistency under high-dimensional regression settings. In a Bayesian framework, posterior model probabilities can be used to quantify the importance of models given the observed data. Hence, our focus is on the asymptotic behavior of posterior model probabilities when the number of the potential predictors grows with the sample size. This dissertation contains the following three projects.

In the first project, we investigate the asymptotic behavior of posterior model probabilities under the Zellner's g-prior, which is one of the most popular choices for model selection in Bayesian linear regression. We establish a simple and intuitive condition of the Zellner's g-prior under which the posterior model distribution tends to be concentrated at the true model as the sample size increases even if the number of predictors grows much faster than the sample size does. Simulation study results indicate that the satisfaction of our condition is essential for the success of Bayesian high-dimensional variable selection under the g-prior.

In the second project, we extend our framework to a general class of priors. The most pressing challenge in our generalization is that the marginal likelihood cannot be expressed in a closed form. To address this problem, we develop a general form of Laplace approximation under a high-dimensional setting. As a result, we establish general sufficient conditions for high-dimensional Bayesian model selection consistency. Our simulation study and real data analysis demonstrate that the proposed condition allows us to identify the true data generating model consistently.

In the last project, we extend our framework to Bayesian generalized linear regression models. The distinctive feature of our proposed framework is that we do not impose any

specific form of data distribution. In this project we develop a general condition under which the true model tends to maximize the marginal likelihood even when the number of predictors increases faster than the sample size. Our condition provides useful guidelines for the specification of priors including hyperparameter selection. Our simulation study demonstrates the validity of the proposed condition for Bayesian model selection consistency with non-Gaussian data.

Bayesian model selection consistency for high-dimensional regression

by

Min Hua

B.A., Jiangxi University of Finance and Economics, China, 2009

M.S., Hangzhou Dianzi University, China, 2013

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Major Professor
Gyuhyeong Goh

Copyright

Abstract

Bayesian model selection has enjoyed considerable prominence in high-dimensional variable selection in recent years. Despite its popularity, the asymptotic theory for high-dimensional variable selection has not been fully explored yet. In this study, we aim to identify prior conditions for Bayesian model selection consistency under high-dimensional regression settings. In a Bayesian framework, posterior model probabilities can be used to quantify the importance of models given the observed data. Hence, our focus is on the asymptotic behavior of posterior model probabilities when the number of the potential predictors grows with the sample size. This dissertation contains the following three projects.

In the first project, we investigate the asymptotic behavior of posterior model probabilities under the Zellner's g-prior, which is one of the most popular choices for model selection in Bayesian linear regression. We establish a simple and intuitive condition of the Zellner's g-prior under which the posterior model distribution tends to be concentrated at the true model as the sample size increases even if the number of predictors grows much faster than the sample size does. Simulation study results indicate that the satisfaction of our condition is essential for the success of Bayesian high-dimensional variable selection under the g-prior.

In the second project, we extend our framework to a general class of priors. The most pressing challenge in our generalization is that the marginal likelihood cannot be expressed in a closed form. To address this problem, we develop a general form of Laplace approximation under a high-dimensional setting. As a result, we establish general sufficient conditions for high-dimensional Bayesian model selection consistency. Our simulation study and real data analysis demonstrate that the proposed condition allows us to identify the true data generating model consistently.

In the last project, we extend our framework to Bayesian generalized linear regression models. The distinctive feature of our proposed framework is that we do not impose any

specific form of data distribution. In this project we develop a general condition under which the true model tends to maximize the marginal likelihood even when the number of predictors increases faster than the sample size. Our condition provides useful guidelines for the specification of priors including hyperparameter selection. Our simulation study demonstrates the validity of the proposed condition for Bayesian model selection consistency with non-Gaussian data.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The journey of the Ph.D. program is no doubt one of the most rewarding experiences I have had my whole life. The amount of things I learned through the ups and downs during the years of the program are priceless. I would love to take this opportunity to express my gratitude to people who had been there for me.

First, I would like to express my most sincere gratitude to my major advisor, Dr. Gyuhyeong Goh. I could not complete the research project without his continuous guidance and support throughout the whole process. Dr. Goh always responded to my questions with advises which were effective in solving the problems. His knowledge and expertise helped me overcome every obstacle in my research. I am also grateful for his patience and encouragement which kept me moving forward, especially during the most difficult time of the program.

Second, I want to express my appreciation to the committee members, Dr. Weixing Song, Dr. Jingru Mu and Dr. Jisang Yu. I am grateful for their time to serve in my committee as well as their valuable comments on my research. I would like to extend my thanks to the Department of Statistics and the Lolafaye Coyne Statistics Graduate Scholarship for providing funding to support my research. I would like to say a special thank you to Dr. Christopher Vahl, Dr. Weixing Song and Dr. Jieun Lee for writing the reference letters for my job applications.

Next, I would like to express my deepest gratitude to my families and friends for their unconditional love and support. To my parents, thank you for believing me and standing by me every step of the way. To my friends, thank you for being the strongest emotional support every time I feel down and lost.

The last but not the least, thank you to everyone else who had helped and supported me. I could not have made it this far without anyone of you. Thank you all so much, and I appreciate everything.

Chapter 1

Introduction

As data become more available in different areas, the data structure diversifies. One of the most commonly observed data patterns is the so-called high-dimensional data. In high-dimensional data, the number of variables is greater than the number of observations. Examples are Biotech data, financial data, satellite imagery and so on. The high-dimensionality in data brings many challenges. For example, in a variable selection problem, the investigation of all possible models can be computationally infeasible due to the extremely large number of variables. In addition, for estimation, the existence of maximum likelihood estimate is not guaranteed in general when the number of variables is larger than the number of observations. Unfortunately, the traditional regression methods are not designed to cope with the "small n large p" nature of the high-dimensional data. Therefore, we need a better approach to solve the curse of high-dimensionality.

In the context of regression analysis, we usually deal with the data (X, y), where $y = (y_1 \ldots, y_n)^\top$ is the n-dimensional response vector and $X = (x_1, \ldots, x_p)$ is the $n \times p$ design matrix. We assume that $E(y|X)$ is a function of the combination of X through the parameter vector β in the following way:

$$E(y|X) = E(y|X, \beta) \quad = \quad g^{-1}(X\beta), \tag{1.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is *a* p-dimensional regression coefficient vector and $g^{-1}(\cdot)$ is the

inverse link function. Let $\gamma \subset \{1, \ldots, p\}$ be an index set such that

$$\gamma = \{j : \beta_j \neq 0, j = 1, \ldots, p\}.$$

In other words, γ indicates which variables are actually important in the model. Given γ, model (1.1) reduces to

$$E(y|X, \beta, \gamma) = g^{-1}(X_\gamma \beta_\gamma),$$

where X_γ is the $n \times p_\gamma$ submatrix of X and β_γ is the p_γ-dimensional vector of non-zero regression coefficients. The goal of variable selection is to estimate γ which contains all indices of non-zero coefficients, called the true model.

1.1 Review of variable selection consistency

In high-dimensional variable selection, an important research question is "Can we identify the true model consistently?" The model selection consistency is important to capture the true relationship between the response and the predictor variables as the sample size increases. The variable selection problem in high-dimensional regression has been studied intensively from both frequentist perspective and Bayesian perspective.

From a frequentist perspective, a variety of methods are developed based on minimizing the loss functions combined with a penalty on the complexity of the model. Knight and Fu (2000) have shown estimation consistency for Lasso (Tibshirani, 1996) for fixed p and fixed $\beta^0_{\gamma_*}$ as n increases. Zou (2006) proposes the adaptive lasso which is consistent in the estimation of the true model parameters for a fixed p_n. For $n \ll p_n$, Huang et al. (2008) show that under the partial orthogonal condition, the adaptive lasso achieves model selection consistency. However, it fails to be consistent in estimating the model parameters. The SCAD regression proposed by Fan and Li (2001) uses a penalty function that is nonconcave on $(0, \infty)$. Fan and Li (2001) show the oracle property of the SCAD estimator. Under certain regularity conditions, Wang et al. (2015) demonstrate the oracle property of SCAD

estimator in the high-dimensional least absolute deviation regression. Candes and Tao (2007) propose the Dantzig selector for the large p_n case (including $n < p_n$) in regression. Bickel et al. (2009) show that the asymptotic equivalence of Lasso estimator and Dantzig selector when the true model is sparse in both linear regression and nonparametric regression. In the sparsity scenario, James et al. (2009) also derive conditions under which the Dantzig selector produces identical solution as the Lasso estimator.

In a Bayesian framework, high-dimensional variable selection methods include Bayesian Lasso (Park and Casella, 2008), stochastic search variable selection (George and McCulloch, 1995), and the spike-and-slab prior approach (Mitchell and Beauchamp, 1988). Using the hierarchical representation of the Laplace prior, Park and Casella (2008) propose a full Bayesian approach to Lasso. Under the orthogonal design and fixed variance settings, Dasgupta (2016) develops conditions for Bayesian lasso to achieve posterior consistency in parameter estimation when the number of predictor variables grows with the sample size. Castillo et al. (2015) show that the posterior distribution of model parameters can be concentrated around the true parameters values when assigning a Laplace-like prior for β_γ. Mitchell and Beauchamp (1988) originally propose the notion of the spike-and-slab prior that employs a mixture of the point mass distribution at 0 and a flat uniform distribution. For easy implementation via Gibbs sampling, George and McCulloch (1995) propose an approximation method for the spike-and-slab prior by mixing two normal distributions. Kuo and Mallick (1998) adopt the mixture of a point mass distribution at 0 and a normal distribution.

In addition, there are several studies that focus on theoretical aspects of Bayesian model selection methods. Jiang (2007) show that the Bayesian method is consistent in estimating the true density of the data under a carefully chosen prior. Bondell and Reich (2012) propose a model selection method via penalized credible regions and show its model selection consistency. Under several popular shrinkage priors, Armagan et al. (2013) show consistency in parameter estimation based on the posterior probability of the model parameters under the general conditions that hold when the dimension grows with the sample size. Sparks et al. (2015) derive necessary and sufficient conditions to achieve high-dimensional posterior consistency in parameters estimation under g-priors.

1.2 Problem statement and set-up

The definition of consistency for variable selection can be classified into two categories: i) Consistency in estimating the true model parameters; ii) Consistency in identifying the true model. In this study, the consistency we focus on is the consistency in identifying the true model, often referred to as model selection consistency. Our approach to model selection consistency is coming from a Bayesian perspective. Under the Bayesian framework, both model γ and parameter β are treated as random variables. By the Bayes theorem, the posterior probability of model γ is computed as

$$\text{pr}(\gamma|y) = \frac{p(y|\gamma)p(\gamma)}{\sum_{\gamma\in\mathcal{M}} p(y|\gamma)p(\gamma)},$$

where

$$p(y|\gamma) = \int p(y|\beta_\gamma, \gamma)p(\beta_\gamma|\gamma)d\beta_\gamma,$$

and $\mathcal{M}$ is the set of candidate models under consideration. The posterior model probability represents the probability that γ is the true model given the observed data. If we compute the posterior model probability for each candidate model, naturally, the model with the greatest posterior probability will be selected as the true model. Thus, we define the Bayesian variable selection consistency as:

$$\text{pr}(\gamma_*|y) \to 1,$$

in probability as $n \to \infty$, where γ_* is the true model. In the Bayesian variable selection context, this implies that the chance of selecting the true model increases as we accumulate more data, while the chance of selecting models other than the true model decreases.

One of practical challenges in Bayesian variable selection is the choice of priors. Note that the posterior model probability depends on the priors of both γ and β_γ. For the prior of γ, we naturally assume an uniform distribution to avoid preference over any candidate models. Thus, the real challenge lies in the choice of the prior for β_γ. Since the success of model selection depends on the choice of the prior, we are required to use a valid prior

that leads to the asymptotic concentration of the posterior model distribution at the true model. However, we do not have a clear picture of how the prior of β_γ affects the asymptotic behavior of the posterior model probability. Thus, our objective in this study is largely centered at uncovering the conditions for the prior of β_γ to achieve posterior model probability consistency in the high-dimensional regression analysis. Considering the "small n, large p" nature of the high-dimensional data, we allow the size of the full model to grow with the sample size, i.e. we assume that $p = O(n^\alpha)$ for $\alpha \in (0, \infty)$ which is a distinctive feature of this dissertation.

1.3 Outline

In Chapter 2, we investigate the model selection consistency under a specific prior, namely the Zellner's g-prior (Zellner, 1986). The g-prior is one of the popular choices for model selection in Bayesian linear regression. The hyperparameter g plays a crucial role in model selection. The success of model selection hinges on the choice of the g value. It is not yet well understood how the choice of g values affect the results of model selection. In our investigation, we focus on deriving conditions under which the posterior model distribution tends to concentrate at the true model as the sample size increases. The setting of our investigation is different from the existing literatures in that we allow the size of the full model to grow with the sample size. Under this setting, we find simple and intuitive conditions for the Zellner's g-prior to yield model selection consistency. The sufficient conditions provide guidelines for the specification of the hyperparameter g. The results of our simulation study confirm that our proposed conditions are crucial to the success of the model selection.

In Chapter 3, we extend the framework of Chapter 2 to a general class of priors. Under the general prior framework, we do not impose any specific distribution on the prior of β_γ. Without knowing the prior distribution, the marginal likelihood is not available in closed form. We propose a general form of Laplace approximation. The proposed Laplace approximation holds even when the model dimension grows with the sample size. We derive general sufficient conditions under which the posterior model distribution concentrates at

the true model asymptotically. The sufficient conditions suggest that the prior needs to be noninformative to avoid prior dominates the likelihood. At the same time, the prior should not be too noninformative to avoid the model selection favors the null model regardless of the information contained in the data. The sufficient conditions are consistent with the results in Chapter 2. Since the conditions are general conditions, we expect them to be applicable to various priors. To make our theoretical findings more practical, we discuss several examples from the shrinkage prior family such as Gaussian prior, Laplace prior, scaled Student's t prior, generalized double Pareto prior, and Horseshoe prior. In our examples, the sufficient conditions provide useful guidelines for hyperparameter specification. The simulation study demonstrates that the satisfaction of the sufficient conditions is essential in the success of model selection. The real data study also shows that specifying the hyperparameters according to our sufficient conditions leads to better model selection results.

In Chapter 4, we extend our framework to the generalized linear regression model by relaxing the normal likelihood assumption. Under this framework, we do not impose any specific distribution on the likelihood of the data. We derive sufficient conditions of the prior of β_γ for maximizing the marginal likelihood of the true model. The conditions we derived under this setting can be applied in a wide range of model selection cases. By investigating several commonly-seen likelihood as well as several frequently used priors in variable selection, we find that the sufficient conditions provide useful guidelines for prior hyperparameter specification. The results in Chapter 4 are consistent with the results in both Chapter 2 and Chapter 3. The simulation study demonstrates the validity of our proposed sufficient conditions.

Chapter 2

The consistency of Bayesian high-dimensional model selection under Zellner's g-prior

2.1 Introduction

Zellner's g-prior (Zellner, 1986) is one of the popular choices of priors in Bayesian variable selection. Under the Zellner's g-prior, the marginal likelihood is available in closed form which significantly reduces computational time. The marginal likelihood plays a key role in deriving the posterior model probability, thus the use of Zellner's g-prior increases the efficiency of the model selection. Despite the popularity of the Zellner's g-prior in the model selection studies, there is an important question we need to address which is "Can we consistently select the true model under the Zellner's g-prior?"

Our study focuses on the consistency of the posterior model probability derived under the g-prior. We investigate the posterior model probability consistency under the high-dimensional settings. We allow the model dimension to grow with the sample size, and even faster than the sample size in some cases. A Growing model space can be challenging, since the number of candidate models can grow immensely as the sample size gets larger. The

large number of candidate models make it infeasible to investigate every candidate model for model selection.

One of the demanding task in using the Zellner's g-prior is the choice of the hyperparameter g in the prior. The hyperparameter g is crucial to the success of model selection, there are many studies addressing the challenge of choosing the g value. Kass and Wasserman (1995) suggest to choose $g = n$, and the resulting prior is known as the "Unit information prior". Foster and George (1994) recommend to choose $g = p_n^2$ according the risk inflation criterion(RIC) for variable selection. Fernandez et al. (2001) suggest to use $g = \max(n, p^2)$ which is referred as "Benchmark prior". All these choices of g are based on the consistency of Bayes factor which does not guarantee the consistency of posterior model probability. We find in our study that, in order to achieve posterior model probability consistency, we need to choose a large value for g, however, g can not be too large. In other words, the g value needs to be confined within a certain range.

In this chapter, we derive sufficient conditions to achieve the posterior model probability consistency under the Zellner's g-prior. The sufficient conditions propose a theoretical boundary for the g values, which provides useful guidelines for the specification of the g value. The simulation study demonstrates the sufficient conditions are essential in establishing posterior model probability consistency.

2.2 Model set-up and assumptions

Consider a linear regression model,

$$y = X\beta + \epsilon, \tag{2.1}$$

where $y = (y_1, \ldots, y_n)^\top$ is the n-dimensional response vector, $X = (x_1, \ldots, x_p)$ is the $n \times p$ design matrix with $x_j = (x_{1j}, \ldots, x_{nj})^\top$, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional regression coefficient vector, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a n-dimensional random error vector with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Without loss of generality, we assume both y and X are centered so that the

intercept is omitted in our model.

Let γ be an index set representing a subset of the predictor variables. Given model γ, model (2.1) reduces to

$$y = X_\gamma \beta_\gamma + \epsilon,$$

where X_γ is the submatrix of X and β_γ is the subvector of β corresponding to the active predictors in model γ. The goal of model selection is to identify the subset γ_* which contains all the true predictor variables, namely, the true model.

Let $\mathcal{M} = \{\gamma : p_\gamma \leq K\}$ be the set of candidate models under our consideration, where p_γ denotes the size of the model γ and K is the upper bound of the candidate model size. Since we adopt the Bayesian approach to the model selection, given model γ, we consider the following priors for β_γ and σ^2:

$$\begin{aligned} \pi(\beta_\gamma|\gamma, \sigma^2) &= \phi\left\{\beta_\gamma|0, g\sigma^2(X_\gamma^\top X_\gamma)^{-1}\right\}, \\ \pi(\sigma^2|\gamma) &= \frac{1}{\sigma^2}, \end{aligned}$$

where $\phi(\cdot|\mu, \Sigma)$ denotes a multivariate normal density function with mean μ and variance Σ and g is a hyperparameter. We consider a flat prior for γ to show no preference over any candidate models,

$$\pi(\gamma) \propto \mathbb{I}(\gamma \in \mathcal{M}),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Given hierarchical model representation, we can derive the closed form expression of the marginal likelihood, which is

$$p(y|\gamma) = \frac{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}(1+g)^{\frac{p_\gamma}{2}}\left[y^\top y - \frac{1}{1+g^{-1}}y^\top H_\gamma y\right]^{\frac{n}{2}}}, \gamma \in \mathcal{M},$$

where $H_\gamma = X_\gamma(X_\gamma^\top X_\gamma)^{-1}X_\gamma^\top$.

Then, by the Bayes theorem, the posterior model probability of γ is proportional to

$$\mathrm{pr}(\gamma|y) \propto \frac{1}{(1+g)^{\frac{p_\gamma}{2}}\left[y^\top y - \frac{1}{1+g^{-1}} y^\top H_\gamma y\right]^{\frac{n}{2}}} \mathbb{I}(\gamma \in \mathcal{M}).$$

Posterior model probability measures the probability that the data is generated by the candidate model, thus it is a natural model selector. The model which maximizes the posterior model probability is selected as the true model. The model selection consistency means that the posterior model probability of true model becomes one as the sample size increases. The consistency of posterior model probability ensures that the chance of selecting the true model increases as we collect more data. In this paper, the Bayesian model selection consistency is defined as follows:

Definition 2.1. *Let γ_* be the true model. When the observations, y, is generated by $\mathcal{N}_n(X_{\gamma_*}\beta^0_{\gamma_*}, \sigma_0^2 I_n)$, where $\beta^0_{\gamma_*}$ is the $p_{\gamma_*} \times 1$ vector of the true non-zero coefficients and σ_0^2 is the true value of variance. The Bayesian model selection is consistent if*

$$pr\,(\gamma_*|y) \to 1,$$

in $\mathcal{Y}|\beta_{\gamma_*}$-probability as $n \to \infty$.

Our objective is to derive sufficient conditions to achieve posterior model probability consistency in high-dimensional regression while the model dimension grows with the sample size, including the cases in which the dimension grows faster than the sample size.

To achieve our goal, the following regularity conditions are needed.

Assumption 2.1. *$\gamma_* \in \mathcal{M}$, where $\mathcal{M}$ is the set of candidate models and γ_* is the true model.*

Assumption 2.1 ensures the true model belongs to the candidate model set so that the model selection consistency is achievable.

Assumption 2.2 (Asymptotic identifiability)**.** *Let $\mu_* = X_{\gamma_*}\beta_{\gamma_*}$ and $H_\gamma = X_\gamma(X_\gamma^\top X_\gamma)^{-1}X_\gamma^\top$.*

The true model γ_* is asymptotically identifiable if there exists $a_0 > 0$ such that

$$\lim_{n\to\infty} \min \left\{ n^{-1} \|(I_n - H_\gamma)\mu_*\|^2 : \gamma_* \not\subset \gamma, p_\gamma \leq K + p_{\gamma_*} \right\} > a_0,$$

where $\|\cdot\|$ is the Euclidean distance and K is an upper bound of p_γ.

The sparse Riesz condition is a common requirement in high-dimensional variable selection for model identifiability (Huang et al., 2012; Wei and Huang, 2010; Zhang et al., 2008, 2010). Assumption 2.2 is weaker than the sparse Riesz condition since it implies the sparse Riesz condition, see Appendix A.1 for the proof.

Assumption 2.3. *$K = o(n)$, where $K = \max\{p_\gamma : \gamma \in \mathcal{M}\}$.*

Assumption 2.3 restricts the size of candidate models to be smaller than sample size due to the belief that the true model is sparse.

2.3 Main results

We begin our discussion of posterior model selection consistency with two cases of Bayesian pairwise model comparison: (1) the true model is compared with a underfitting model, i.e. the candidate model is a sub-model of the true model; (2) the true model is compared with a overfitting model, i.e. the true model is a sub-model of a candidate model. The following lemmas ensures the the consistency in the Bayesian pairwise model comparisons in both cases.

Lemma 2.1. *Let $\mathcal{M}_1 = \{\gamma : \gamma \subsetneq \gamma_*\}$. Under Assumption 2.1-2.3, for any given $\gamma \in \mathcal{M}_1$, if g grows with n, then*

$$\log \frac{p(y|\gamma_*)}{p(y|\gamma)} > c_0 n - \frac{p_{\gamma_*} - p_\gamma}{2} \log(1+g)$$

as $n \to \infty$, where c_0 is a positive constant.

Since the posterior model probability is proportional to the marginal likelihood. The marginal likelihood ratio, namely the Bayes factor, is equivalent to the ratio of the posterior

model probability, that is,

$$\frac{p(y|\gamma_*)}{p(y|\gamma)} = \frac{\mathrm{pr}(\gamma_*|y)}{\mathrm{pr}(\gamma|y)}.$$

Lemma 2.1 implies that as $n \to \infty$, we have $\mathrm{pr}(\gamma_*|y) > \mathrm{pr}(\gamma|y)$ for any underfitting models γ if g is chosen to increase with n but $\log g = o(n)$. The proof of 2.1 is given in Appendix A.4

Lemma 2.2. *Let* $\mathcal{M}_2 = \{\gamma : \gamma_* \subsetneq \gamma, p_\gamma \leq K + p_{\gamma_*}\}$. *For any given* $\gamma \in \mathcal{M}_2$, *if* g *grows with* n, *we have*

$$\log \frac{p(y|\gamma_*)}{p(y|\gamma)} > \frac{p_\gamma - p_{\gamma_*}}{2} \log(1+g) - (p_\gamma - p_{\gamma_*}) \log p\{1 + o_p(1)\}$$

as $n \to \infty$.

Similarly, Lemma 2.2 implies that $\mathrm{pr}(\gamma_*|y) > \mathrm{pr}(\gamma|y)$ for any overfitting models γ under our consideration if g is chosen to grow faster than p along with n. The proof of Lemma 2.2 is given in Appendix A.5.

Remark 2.1. *Even though we already know* g *can not grow too fast, according to Lemma 2.2,* g *can not grow too slow either. The convergence rate of lower bound is dominated by* $\log g$ *instead of* n *in* $\mathcal{M}_2$. *A possible issue could be a significant slower convergence of the log ratio in* $\mathcal{M}_2$ *if* g *does not grow fast enough. Thus, in order to increase the convergence rate, we recommend to choose* g *to grow faster than* n *but* $\log g = o(n)$, *in other words,* g *should be large but not be too large.*

Case 1 and Case 2 are two basic types of candidate models. Besides these two cases, there is a third case which is the true model is compared with a misspecified candidate model. In this case, neither a candidate model is a sub-model of true model nor the true model is the sub-model of the candidate model. The following lemma shows that the Bayesian pairwise comparison still holds in this case.

Lemma 2.3. *Let* $\mathcal{M}_3 = \{\gamma : \gamma \not\subset \gamma_*, \gamma_* \not\subset \gamma, p_\gamma \leq K\}$. *Under Assumptions 2.1-2.3, for any*

given $\gamma \in \mathcal{M}_3$*, if* g *grows with* n*, then*

$$\log \frac{p(y|\gamma_*)}{p(y|\gamma)} > c_0 n + \frac{p_\gamma - p_{\gamma_*}}{2} \log(1+g) - p_\gamma \log p\{1 + o_p(1)\}$$

as $n \to \infty$*.*

Lemma 2.3 implies, we have $\text{pr}(\gamma_*) > \text{pr}(\gamma)$ for any $\gamma \in \mathcal{M}_3$ for sufficiently large n if g is chosen to increase with n but $\log g = o(n)$. The proof is shown in Appendix A.6.

The lemmas imply posterior probability of true model is consistent when comparing to a given candidate model as sample size grows. However, pairwise model comparison consistency does not always leads to consistency in posterior model probability. The following theorem proposes sufficient conditions to achieve posterior model probability consistency.

Theorem 2.1. *Under Assumptions 2.1-2.3, if* g *is chosen such that* $p^4 < g$ *and* $\log g = o(n)$*, then the Bayesian model selection with the g-prior is consistent, that is,*

$$pr\,(\gamma_*|y) \to 1,$$

in $\mathcal{Y}|\beta_{\gamma_*}$*-probability as* $n \to \infty$*.*

The proof of Theorem 2.1 is given in Appendix A.7. According to Theorem 2.1, g needs to meet the conditions $p^4 < g$ and $\log g = o(n)$ in order to achieve model selection consistency. The following corollary guarantees the existence of such g.

Corollary 2.1. *If* $\log p = o(n)$*, there exists* g *satisfying the sufficient conditions in Theorem 2.1, that is,* $p^4 < g$ *and* $\log g = o(n)$*.*

The proof of Corollary 2.1 is given in Appendix A.8. The sufficient condition $\log p = o(n)$ in the corollary indicates p can grow with the sample size. As long as p does not grow too fast, we can still find a g which satisfies the sufficient conditions in Theorem 2.1 and achieve model selection consistency, even under the "large p small n" cases. It's also not hard to see, since p can be smaller than the sample size, Theorem 2.1 is valid in the low dimensional

cases as well. To demonstrate our result, we perform a simulation study in the following section.

Remark 2.2. *The choice of g is crucial to achieve model selection consistency. Theorem 2.1 implies that g needs to grow with the sample size n, however, the growth rate needs to be controlled. In model selection cases in which n and p are fixed, letting g grow to infinity results in the Bayes factor favors the null model over any candidate models, which is known as "Lindley-Barlett" paradox (Foster and George, 1994; Jeffreys, 1961; Liang et al., 2008). In high dimensional analysis, n and p are most likely not fixed. They can be treated as fixed when g increases at a much higher rate than n and p. Even though g should not grow too fast, a g which does not grow fast enough leads to inconsistency in model selection, especially when the sample size is not sufficiently large. Our sufficient conditions restrict g to grow not too fast to trigger the "Lindley-Barlette" paradox while still large enough to establish the model selection consistency.*

The following are several examples of the choices of g value from existing literature:

Example 1. *Kass and Wasserman (1995) suggest to choose priors according to the information contained in a single observation which is referred as " Unit information prior". In the normal regression analysis, for the Zellner's g-priors, they recommended choosing $g = n$ which makes the BIC an asymptotically accurate approximation to Bayes factor. According to our sufficient conditions, choosing $g = n$ achieves posterior consistency in low dimensional cases such that $p < n^{-1/4}$. When $p > n^{-1/4}$, we have $g < p^4$ which violates the first condition which leads to the inconsistent model selection.*

Example 2. *Foster and George (1994) propose risk inflation criterion(RIC) for variable selection, they recommend to choose $g = p^2$, which minimizes the RIC. According to the sufficient conditions, choosing $g = p^2$ violates the first condition, in this case, g does not grow fast enough to allow the posterior model probability to achieve consistency asymptotically.*

Example 3. *Fernandez et al. (2001) suggest to use $g = \max(n, p^2)$ which is referred as "Benchmark prior". In this case, the posterior model consistency can be achieved in low*

dimensional cases where $n > p^4$. When $p^4 > n$, the first sufficient condition is violated, thus, the g is not large enough to yield posterior consistency.

2.4 Simulation study

In this section, we conduct simulation study to exam the performance of our theorem in model selection. We generate our data from the following model

$$\begin{aligned} y_i &= \sum_{j=1}^{p_n} \beta_j x_{ij} + \epsilon_i, \\ (x_{i1}, \ldots, x_{ij})^\top &\overset{iid}{\sim} N(0, \Sigma) \\ \epsilon_i &\overset{iid}{\sim} N(0, 1) \end{aligned}$$

with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma^{ij} = 2\rho^{|i-j|}$, for $i = 1, \ldots, n$. $\beta_j = 1$ for $j \leq 4$ and $\beta_j = 0$ for $j > 4$.

Table 2.1 shows the scenarios we consider. By the setting of κ and ρ, we can see that Scenario 1 is the low-dimensional case with uncorrelated predictor variables; Scenario 2 is the low-dimensional case with correlated predictor variables; Scenario 3 is the high-dimensional case with uncorrelated predictor variables; Scenario 4 is the high-dimensional case with correlated predictor variables. To satisfy Assumption 2.3, we set $K = n^{\frac{2}{3}}$. For the choices of the g value, we consider 5 cases. Table 2.2 contains our choices of g values for each scenario. We repeat the simulation 100 times with $n = \{100, 300, 500, 900\}$ for each case.

Table 2.1: *Simulation scenarios*

Scenario	κ	ρ
1	0.6	0
2	0.6	0.5
3	1.1	0
4	1.1	0.5

According to the sufficient conditions, the choices of g values in Case 1 and Case 2 violate the condition $p^4 < g$. The choice of g value in Case 5 violates the condition $\log g = o(n)$. Case 3 and Case 4 satisfy both sufficient conditions.

Table 2.2: *Choices of the g value*

Case	g
1	n
2	p^2
3	p^5
4	p^6
5	$\exp(0.7n)$

2.4.1 Gibbs sampler

The number of models in our candidate model set is $\sum_{k=1}^{K} \binom{p}{k} = O(p^K)$. When n is sufficiently large, it is computationally infeasible to compute posterior probabilities for all the candidate models. To address this computational challenge, we use Gibbs sampler to explore the candidate model space.

To evaluate the performance of Theorem 2.1 in the simulation, we use the Monte Carlo estimator to estimate the posterior probability of the true model

$$P(\gamma_*|y) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{\gamma^{(t)} = \gamma_*\},$$

where $\{\gamma^{(t)} : t = 1, \ldots, T\}$ is a Gibbs sequence generated from the posterior distribution of $\text{pr}(\gamma|y)$. We will run the Gibbs sample 6000 times with first 3000-iteration as the burning period. We compute the posterior probability of the true model using the a sample generated by resample very third iteration post the burn-in period. If Theorem 2.1 holds, we need to observe the estimated $\text{pr}(\gamma_*|y) \to 1$ as the sample size increases. The Gibbs sampler can be implemented by updating $\gamma^{(t)}$ as follows:

Algorithm 1 The Gibbs sampler.

Set $s = \gamma^{(t)}$

Repeat for $j = 1, \ldots, p$:

Set $s_1 = s \cup \{j\}$

Set $s_0 = s \setminus \{j\}$

Compute $w = \{p(y \mid s_1)\mathbb{I}(s_1 \in \mathcal{M})\}/\{p(y \mid s_1) + p(y \mid s_0)\}$

Sample $z \sim \text{Bernoulli}(w)$

If $z = 1$ then $s \leftarrow s_1$;

else $s \leftarrow s_0$

Output $\gamma^{(t+1)} = s$

2.4.2 Simulation results

Table 2.3 - Table 2.4 report the mean of the estimated $\text{pr}(\gamma_*|y)$ and the corresponding standard error over 100 Monte Carlo experiments of the four scenarios we consider.

Table 2.3 presents the results of model selection in the low-dimensional cases. For the cases which violate the sufficient conditions, namely Case 1, Case 2 and Case 5, the posterior probability of the true model fails to grow to one as sample size increases. For the cases which satisfy the sufficient conditions, namely Case 3 and Case 4, the posterior probability of the true model increases to one as sample size gets larger. The results are similar regardless of the correlation among the predictor variables.

Table 2.3: *Simulation results based on 100 Monte Carlo experiments.*

Scenario	Case	n=100	n=300	n=500	n=900
I	1	0.1430 (0.0076)	0.0544 (0.0038)	0.0393 (0.0028)	0.0211 (0.0016)
	2	0.2584 (0.0107)	0.1535 (0.0084)	0.1440 (0.0075)	0.1164 (0.0060)
	3	0.9587 (0.0066)	0.9794 (0.0035)	0.9891 (0.0007)	0.9936 (0.0004)
	4	0.9896 (0.0019)	0.9963 (0.0008)	0.9982 (0.0002)	0.9992 (0.0001)
	5	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
II	1	0.1860 (0.0072)	0.0597 (0.0040)	0.0384 (0.0026)	0.0290 (0.0020)
	2	0.2749 (0.0104)	0.1607 (0.0085)	0.1398 (0.0068)	0.1392 (0.0069)
	3	0.9618 (0.0043)	0.9764 (0.0056)	0.9882 (0.0024)	0.9948 (0.0004)
	4	0.9886 (0.0014)	0.9967 (0.0004)	0.9981 (0.0005)	0.9994 (0.0001)
	5	0.0089 (0.0056)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)

Notes: (Case 1) $g = n$; (Case 2) $g = p^2$; (Case 3) $g = p^5$; (Case 4) $g = p^6$; (Case 5) $g = \exp(0.7n)$. $p = 15, 30, 41, 59$ for $n = 100, 300, 500, 900$.

Table 2.4 presents the results of model selection in the high-dimensional cases. For the cases which violate the sufficient conditions, namely Case 1, Case 2 and Case 5, the posterior probability of the true model fails to grow to one as sample size increases. For the cases which satisfy the sufficient conditions, namely Case 3 and Case 4, the posterior probability of the true model increases to one as sample size gets larger. The results are similar regardless of the correlation among the predictor variables.

Figure 2.1 shows the trace of the $p(\gamma_*|y)$ as the sample size increases. In the trace plots, we only observe the trace increases along with the sample in Case 3 and 4, while it fails to grow in Case 1, Case 2 and Case 5 in each scenario.

Table 2.4: *Simulation results based on 100 Monte Carlo experiments.*

Scenario	Case	n=100	n=300	n=500	n=900
	1	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	2	0.0538 (0.0037)	0.0525 (0.0032)	0.0483 (0.0025)	0.0415 (0.0022)
III	3	0.9975 (0.0005)	0.9996 (0.0002)	0.9999 (0.0000)	1.0000 (0.0000)
	4	0.9998 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	5	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	1	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	2	0.0665 (0.0048)	0.0526 (0.0033)	0.0554 (0.0029)	0.0429 (0.0024)
IV	3	0.9974 (0.0005)	0.9997 (0.0001)	0.9998 (0.0000)	0.9999 (0.0000)
	4	0.9999 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	5	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)

Notes: (Case 1) $g = n$; (Case 2) $g = p^2$; (Case 3) $g = p^5$; (Case 4) $g = p^6$; (Case 5) $g = \exp(0.7n)$. $p = 158, 530, 930, 1776$ for $n = 100, 300, 500, 900$.

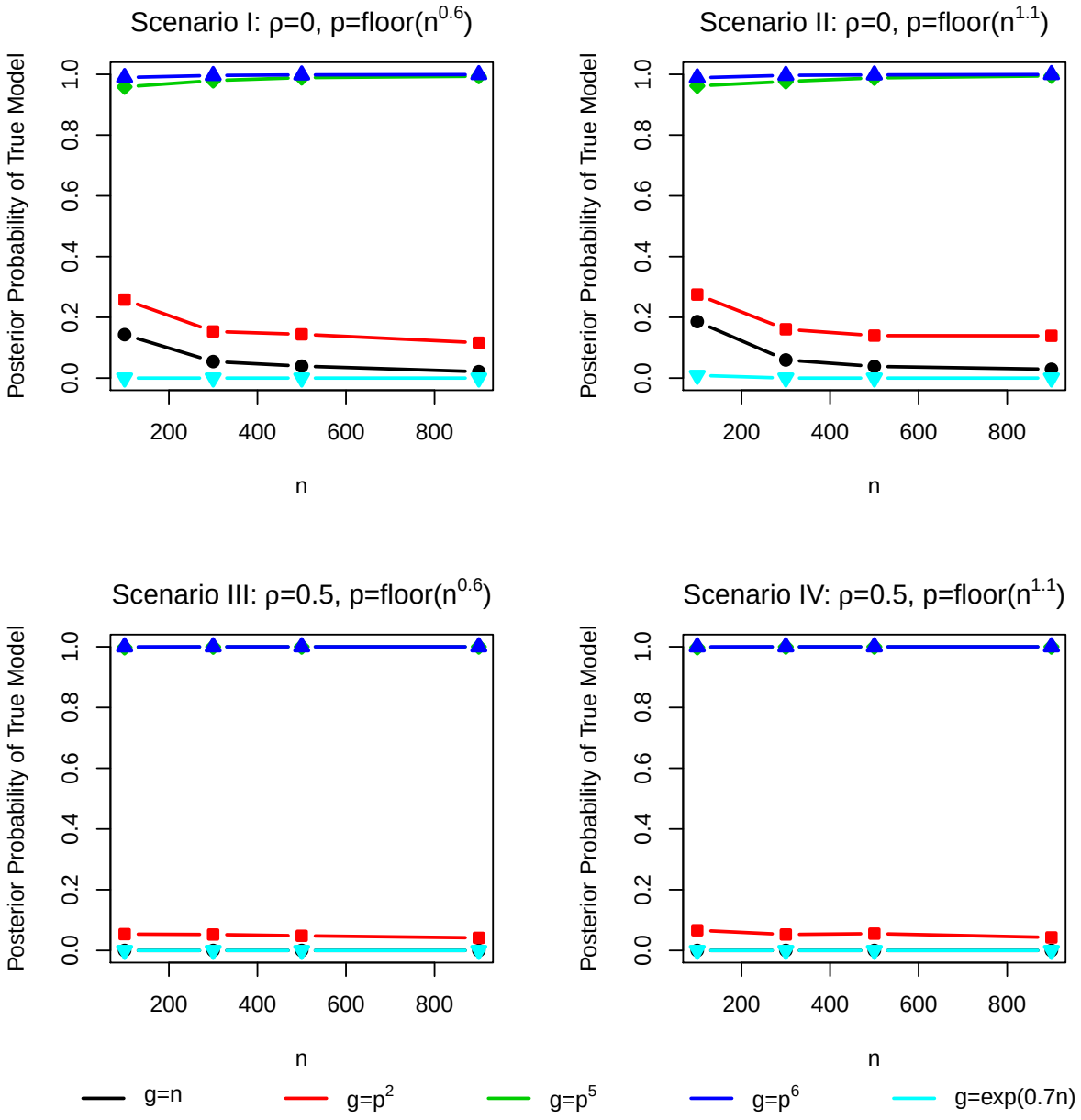

Figure 2.1: *The trace plots of the estimate of* $pr(\gamma_* \mid y)$ *as the sample size* n *increases.*

2.5 Discussion

In this chapter, we find out that, when $\log p = o(n)$, we can achieve posterior model probability consistency with the Zellner's g-priors by choosing the hyperparameter g such that $p < \sqrt{g}$ and $\log g = o(n)$. The two conditions propose useful guidelines for specifying the hy-

perparameter g. Our study also implies that the model selection consistency can be achieved in both low-dimensional and high-dimensional cases.

The Zellner's g-prior's ability to induce sparse solution makes it a popular prior choice in Bayesian high-dimensional model selection. Besides the Zellner's g-prior, there are other distributions which are popular prior choices in model selection study, such as the Laplace prior, Horseshoe prior, and scaled Student's t prior, to name a few. In the Bayesian regression analysis, prior specification is a challenging and important topic, thus an interesting direction for future research is to extend our framework to a class of general priors. By exploring variable selection consistency under the general prior settings, we expect to derive general conditions which can be applied to a variety of priors to induce Bayesian model selection consistency.

Chapter 3

The consistency of Bayesian high-dimensional model selection under arbitrary priors

3.1 Introduction

Bayesian model selection has become a popular approach to model selection in regression over the years. A few examples of Bayesian model selection are Zellner and Siow (1980), Stewart and Davis (1986), George and McCulloch (1993), George and McCulloch (1997) and Raftery et al. (1997) in linear regression; George et al. (1996) in general linear regression; Smith and Kohn (1996), Hansen and Yu (2001) and Kohn et al. (2000) in nonparametric regression. Under the Bayesian framework, we compute the posterior model probability for each candidate model. The model which maximizes the posterior model probability is selected as the true model. The model selection consistency means that the posterior probability of the true model approaches one as we accumulate more data.

The prior of the model parameters plays a important role in determining the posterior probability of the candidate model. The choice of the prior greatly affects the consistency of Bayesian model selection. There are many studies attempting to address the issue of

prior specification. For example, Moreno et al. (2015) prove that Bayesian model selection is consistent under the intrinsic priors, g-priors with $g = n$, and the mixture of g-priors when $C_1 \leq p_n < C_2 n^{1/2}$ for some positive constants C_1 and C_2, where p_n is the size of the full model. Shang and Clayton (2011) show the consistency of model selection under point mass spike and Gaussian flat priors. Liang et al. (2008) demonstrate model selection consistency under a mixture of g-priors. These studies provide valuable insights into the consistency of Bayesian model selection under a variety of priors. Nevertheless, prior specification still remains a challenging issue in the Bayesian model selection.

Our main focus and contribution in this chapter is that we derive sufficient conditions for the general prior of model parameters to achieve posterior model consistency in high-dimensional model selection. The sufficient conditions propose general guidelines for hyper-parameter specifications. The proposed guidelines are applicable to a variety of priors. Our simulation study and real data study demonstrate the validity of the sufficient conditions in high-dimensional model selection. It is worth noting that a unique feature of the study is that we assume the size of the full model grows with the sample size, possibly at a faster rate. Our main results hold in both low-dimensional and high-dimensional model selection.

3.2 Model set-up and assumptions

Consider a linear regression model,

$$y = X\beta + \epsilon, \tag{3.1}$$

where $y = (y_1, \ldots, y_n)^\top$ is the n-dimensional response vector, $X = (x_1, \ldots, x_{p_n})$ is the $n \times p_n$ design matrix, $\beta = (\beta_1, \ldots, \beta_{p_n})^\top$ is a p_n-dimensional regression coefficient vector, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a n-dimensional random error vector with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. For now, we assume σ^2 is known and will discuss the case when σ^2 is unknown later. Without loss of generality, we assume both y and X are centered so that the intercept can be omitted.

In linear regression, model selection is performed by identifying predictor variables with

non-zero coefficients. Let $\gamma \subset \{1, \ldots, p_n\}$ be an index set representing a submodel of size p_γ. Given the submodel γ, the full model reduces to

$$y = X_\gamma \beta_\gamma + \epsilon,$$

where X_γ is the $n \times p_\gamma$ submatrix of X and β_γ is the p_γ-dimensional vector of non-zero regression coefficients.

In this paper, we allow the size of the full model p_n to grow with the sample size n. Let $\mathcal{M}_n = \{\gamma : p_\gamma \leq K_n = o(n)\}$ be the set of candidate submodels under consideration, where K_n is the upper bound of the size of the candidate models (Abramovich et al., 2010; Martin et al., 2017; Wang, 2009). We believe it is necessary to impose such a restriction on K_n for the following reasons. First, models which are larger than the sample size are not identifiable. Second, the true model is usually a sparse model. Third, models with numerous predictors are hard to interpret. Lastly, for some priors such as the g-prior, it leads to improper posterior distributions for models that contain more predictors than the sample size.

Under the Bayesian framework, the model which maximizes the posterior model probability is selected as the true model, that is,

$$\hat{\gamma} = \arg \max_{\gamma \in \mathcal{M}_n} \text{pr}(\gamma | y).$$

The Bayesian model selection consistency is understood as the posterior probability of the true model approaches one as the sample size increases. Therefore, we formally define the Bayesian model selection consistency as follows (Liang et al., 2008; Moreno et al., 2015; Shang and Clayton, 2011):

Definition 3.1. *Let γ_* be the true model. When the observations, y, is generated by $\mathcal{N}_n(X_{\gamma_*}\beta^0_{\gamma_*}, \sigma_0^2 I_n)$, where $\beta^0_{\gamma_*}$ is the $p_{\gamma_*} \times 1$ vector of the true non-zero coefficients and σ_0^2*

is the true value of variance. The Bayesian model selection is consistent if

$$pr(\gamma_*|y) \to 1,$$

in $\mathcal{Y}|\beta_{\gamma_*}$*-probability as* $n \to \infty$.

To complete our Bayesian model set-up, we consider a uniform distribution for the candidate models which is given by

$$\text{pr}(\gamma) \propto \mathbb{I}(\gamma \in \mathcal{M}_n),$$

where $\mathbb{I}$ is an indicator function. For the prior of β_γ, we consider a general form

$$p(\beta_\gamma|\gamma) \sim \pi(\beta_\gamma),$$

which is the most distinctive difference of this chapter. Our objective is to derive general sufficient conditions to achieve posterior model probability consistency while p_n grows with the sample size, including cases where p_n grows at a faster rate than n. To accomplish our goal, the following regularity conditions are required:

Assumption 3.1. *The true data generating model is* γ_**, and* $\gamma_* \in \mathcal{M}_n$.

Assumption 3.2. $p_n = O(n^\alpha)$ *for* $\alpha \in [0, \infty)$.

Assumption 3.3. $K_n = O(n^b)$ *for* $b \in [0, 1)$.

Assumption 3.4. $p_{\gamma_*} = O(n^c)$ *for* $c \in [0, b)$.

Assumption 3.1 ensures the true model is contained in the set of the candidate models we consider. Assumption 3.2 allows the total number of predictors p_n to grow with the sample size. Assumption 3.3 restricts the size of candidate models to be smaller than the sample size. Assumption 3.4 controls the size of the true model.

Remark 3.1. *Under Assumption 3.2, the growth rate of p_n depends on α. It grows faster than n when $\alpha \in [1, \infty)$, and it grows slower than n when $\alpha \in [0, 1)$. In cases where $p_n < n$, p_n is the natural upper bound, thus Assumption 3.3 is no longer required in such circumstances.*

Assumption 3.5. *Let $\lambda_{\min,\gamma}$ and $\lambda_{\max,\gamma}$ be the smallest and largest eigenvalues of $\frac{1}{n}X_\gamma^T X_\gamma$ respectively. There exist $\lambda_{\min}$ and $\lambda_{\max}$ such that*

$$0 < \lambda_{\min} < \inf_{\gamma: p_\gamma \leq 2K_n} \lambda_{\min,\gamma} < \sup_{\gamma: p_\gamma \leq 2K_n} \lambda_{\max,\gamma} < \lambda_{\max} < \infty.$$

Assumption 3.6. *For any γ such that $p_\gamma \leq p_{\gamma_*}$, let $\tilde{\beta}_\gamma$ be the probability limit of $\hat{\beta}_\gamma$ where $\hat{\beta}_\gamma = (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top y$. The Euclidean norm of $\tilde{\beta}_\gamma$ is bounded in the sense that $\| \tilde{\beta}_\gamma \|_2 \leq \tilde{c}_\gamma p_\gamma$ for some constant $\tilde{c}_\gamma \in (0, \infty)$.*

Assumption 3.5 is referred as the sparse Riesz condition (Zhang et al., 2008) which is a common assumption in high-dimensional regression (Huang et al., 2012; Wei and Huang, 2010; Zhang et al., 2010). Assumption 3.6 allows the magnitude of $\hat{\beta}_\gamma$ to grow, however, it is bounded by the size of the model, i.e. $\|\hat{\beta}_\gamma\|_2 \leq c_\gamma p_\gamma$ for some constant $c_\gamma > 0$. Under Assumption 3.4 and Assumption 3.6, we have $\|\beta^0_{\gamma_*}\|_2 \leq c_{\gamma_*} p_{\gamma_*} = d_{\gamma_*} n^c$ for some constant $d_{\gamma_*} > 0$, where $\beta^0_{\gamma_*}$ is the true coefficients vector of the true model. Note that we use the same notation c_γ to denote the constant, however, c_γ is different for a different γ.

Assumption 3.7. *Let B_R be a ball such that $B_R(0) = \{\beta_\gamma \in \mathbb{R}^{p_\gamma} : \| \beta_\gamma \|_2 \leq d_\gamma n^c + 1\}$, for any β'_γ, $\beta_\gamma \in B_R(0)$, we have*

$$|\log p(\beta'_\gamma|\gamma) - \log p(\beta_\gamma|\gamma)| \leq F_1 \| \beta'_\gamma - \beta_\gamma \|_2,$$

for some constants $F_1 \in (0, \infty)$ and $d_\gamma \in (0, \infty)$.

Assumption 3.8. *For any $\beta_\gamma \in \mathbb{R}^{p_\gamma}$, we have*

$$\log p(\beta_\gamma|\gamma) - \log p(0|\gamma) \leq F_2,$$

for some constant $F_2 \in (0, \infty)$.

Assumption 3.7 and Assumption 3.8 require the prior of β_γ to be log-Lipschtz with respect to the radius $R > 0$ and the log-density ratios are bounded. Assumption 3.7 is a common assumption in high-dimensional regression (Barber et al., 2016; Ghosal et al., 1999).

3.3 Main results

Since we do not assume any specific distribution of $p(\beta_\gamma|\gamma)$, the marginal likelihood can not be expressed in a closed form. Hence, we employ the Laplace approximation to derive a general closed form expression of the marginal likelihood. We propose a general form of Laplace approximation for models with growing dimensions.

Lemma 3.1. *Suppose that Assumption 3.2 to 3.8 hold. Then,*

$$p(y|\gamma) = p(y|\hat{\beta}_\gamma)p(\hat{\beta}_\gamma|\gamma)|X_\gamma^\top X_\gamma|^{-1/2}(2\pi\sigma^2)^{p_\gamma/2}\left\{1 + O_p\left(\sqrt{\frac{p_\gamma \log n}{n}}\right)\right\},$$

for any $\gamma \in \mathcal{M}_n$, *where* $\hat{\beta}_\gamma = (X_\gamma^\top X_\gamma)^{-1}X_\gamma^\top y$.

The proof of Lemma 3.1 is given in Appendix B.3. Lemma 3.1 allows us to approximate the marginal likelihood when the dimensional of the model expands with the sample size. The result of Lemma 3.1 is similar to the result of Theorem 1 in Barber et al. (2016). However, the approximation error in Lemma 3.1 is relatively smaller due to the normal likelihood assumption.

Remark 3.2. *Lemma 3.1 indicates that the Laplace approximation error is* $O_p\left(\sqrt{\frac{p_\gamma \log n}{n}}\right)$. *For any* $\gamma \in \mathcal{M}_n$, *by Assumption 3.3, we have* $\sqrt{p_\gamma \log n/n} \leq \sqrt{c''_\gamma n^b \log n/n}$. *As* n *increases,* $\sqrt{c''_\gamma n^b \log n/n}$ *decreases to* 0 *where* c''_γ *is some positive constant.*

For the true model γ_*, *by Assumption 3.4,* $\sqrt{p_{\gamma_*} \log n/n}$ *decreases to 0 as* n *increases where* c''_{γ_*} *is some positive constant. When the sample size is sufficiently large, the approximation error is negligible.*

Remark 3.3. *We know that* $p(\gamma_*|y) = \frac{p(y|\gamma_*)p(\gamma_*)}{\sum p(y|\gamma)p(\gamma)} = \frac{1}{1+\sum_{\gamma\neq\gamma_*} p(y|\gamma)p(\gamma)/p(y|\gamma_*)p(\gamma_*)}$*, and we also have* $\sum_{\gamma\neq\gamma_*} \frac{p(y|\gamma)p(\gamma)}{p(y|\gamma_*)p(\gamma_*)} \propto \sum_{\gamma\neq\gamma_*} \frac{p(y|\gamma)}{p(y|\gamma_*)}$ *due to the uniform distribution assumption of model prior. We then have* $p(\gamma_*|y) \to 1$ *if* $\sum_{\gamma\neq\gamma_*} p(y|\gamma)/p(y|\gamma_*) \to 0$*. Following Lemma 3.1, we have* $\frac{p(y|\gamma)}{p(y|\gamma_*)} \propto \frac{p(y|\hat{\beta}_\gamma)p(\hat{\beta}_\gamma|\gamma)}{p(y|\hat{\beta}_{\gamma_*})p(\hat{\beta}_{\gamma_*}|\gamma_*)}$*. It is not hard to see that the posterior probability of the true model depends on the ratio of the likelihoods and the ratio of the priors of* β_γ*.*

We now focus our discussion on pairwise model comparisons. We first consider the case in which the true model is compared with the overfitting models, i.e. models which contain the true model.

Lemma 3.2. *Under Assumption 3.2, for any* γ *such that* $\gamma_* \subsetneq \gamma$ *and* $p_\gamma = o(n)$*, we have*

$$-2\log\left\{\frac{p(y \mid \hat{\beta}_{\gamma_*})}{p(y \mid \hat{\beta}_\gamma)}\right\} < r_\gamma \Lambda_n,$$

in probability as $n \to \infty$*, where* $r_\gamma = p_\gamma - p_{\gamma_*}$ *and* $\Lambda_n = 2(\log p_n + \delta_n) + \sqrt{2(\log p_n + \delta_n)} + 1$ *with* $\delta_n = o(\log p_n)$*.*

The proof of Lemma 3.2 is given in Appendix B.4. Lemma 3.2 implies that, when the true model is compared with the overfitting models, the likelihood fails to achieve consistency in the sense that $p(y|\hat{\beta}_{\gamma_*}) < p(y|\hat{\beta}_\gamma)$ in probability due to the negative lower bound of the log-likelihood ratio which is $-r_\gamma\Lambda_n/2$.

Next, we consider the case in which the true model is compared with the underfitting models, i.e. models which are nested in the true model.

Lemma 3.3. *Under Assumptions 3.2-3.3, for any* $\gamma \subsetneq \gamma_*$ *and* $p_{\gamma_*} = o(n)$*, there exists a positive constant* b_0 *such that*

$$\min_{\gamma\subsetneq\gamma_*} 2\log\left\{\frac{p(y \mid \hat{\beta}_{\gamma_*})}{p(y \mid \hat{\beta}_\gamma)}\right\} > b_0 n.$$

The proof of Lemma 3.3 is given in Appendix B.5. Lemma 3.3 implies that, when the true model is compared with any underfitting models, the likelihood achieves consistency in the sense that $p(y|\hat{\beta}_{\gamma_*}) > p(y|\hat{\beta}_\gamma)$ in probability as $n \to \infty$.

Remark 3.4. *In addition to the first two pairwise comparison cases, we have a third case in which the competing models are neither underfitting nor overfitting to the data. Suppose* γ *is a model such that* $\gamma \not\subset \gamma_*$ *and* $\gamma_* \not\subset \gamma$ *with* $p_\gamma = o(n)$ *and* $p_{\gamma_*} = o(n)$.

For any γ *in the third case, let* $\gamma_{**} = \gamma \cup \gamma_*$ *be the smallest overfitting model that contains both* γ *and* γ_*, *and* γ_{**} *can be regarded as the true model with some zero coefficients. By Lemma 2.2, we have*

$$\log\{p(y|\hat{\beta}_{\gamma_*})/p(y|\hat{\beta}_{\gamma_{**}})\} > -\frac{r_{**}}{2}\Lambda_n, \tag{3.2}$$

where $r_{**} = p_{\gamma_{**}} - p_{\gamma_*}$ *and* $\Lambda_n = 2(\log p_n + \delta_n) + \sqrt{2(\log p_n + \delta_n)} + 1$ *with* $\delta_n = o(\log p_n)$. *By Lemma 3.3, we have*

$$\log\{p(y|\hat{\beta}_{\gamma_{**}})/p(y|\hat{\beta}_{\gamma})\} > \frac{1}{2}b_0 n, \tag{3.3}$$

for some positive constant b_0. *By (3.2) and (3.3), it follows that*

$$\log\{p(y|\hat{\beta}_{\gamma_*})/p(y|\hat{\beta}_{\gamma})\} > \frac{1}{2}b_0 n\{1 + o(1)\}.$$

Hence, we have $p(y|\hat{\beta}_{\gamma_*}) > p(y|\hat{\beta}_{\gamma})$ *in probability as* $n \to \infty$ *for any* γ *that is neither underfitting nor overfitting to the sample data.*

Under our model set-up, selecting the model with the greatest posterior model probability is equivalent to selecting the model which maximizes the marginal likelihood, that is

$$\hat{\gamma} = arg\max_{\gamma \in \mathcal{M}_n} \mathrm{pr}(\gamma|y) = arg\max_{\gamma \in \mathcal{M}_n} p(y|\gamma).$$

To achieve the model selection consistency stated in Definition 3.1, it suffices that

$$\sum_{\gamma \in \mathcal{M}_n} \frac{p(y|\gamma)}{p(y|\gamma_*)} \to 0$$

in probability as $n \to \infty$. The inconsistency in the likelihoods suggests that we need certain conditions in order to achieve posterior model probability consistency. The following theorem proposes such conditions.

Theorem 3.1. *Under Assumptions 3.1 to 3.5, we have*

$$\sum_{\gamma \in \mathcal{M}_n} \frac{p(y|\gamma)}{p(y|\gamma_*)} \to 0,$$

in probability as $n \to \infty$ if the following conditions hold:

$$\textit{Condition 1: } \frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o(n) \textit{ for any } \gamma \in \mathcal{M}_n \setminus \{\gamma_*\},$$

$$\textit{Condition 2: } \frac{1}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} > \log \frac{p_n^2}{\sqrt{n}} \textit{ for any } \gamma \in \mathcal{M}_1 = \{\gamma \in \mathcal{M}_n : \gamma \supsetneq \gamma_*\}.$$

The proof of Theorem 3.1 is given in the Appendix B.6. Theorem 3.1 implies if the prior of β_γ satisfies the sufficient conditions, the posterior model probability of true model becomes one while the posterior model probabilities of other models become zero as the sample size increases. In other words, the posterior model probability distribution degenerates to a point mass distribution concentrates at the true model asymptotically.

Remark 3.5. *Theorem 3.1 proposes how non-informative the prior of β_γ should be. Condition 2 suggests that the prior should be non-informative to avoid the prior dominating the model selection. Condition 1 suggests that the prior should not be too non-informative to avoid the model selection favors the null model regardless of the information the data contains. Theorem 3.1 implies that, to achieve posterior model selection consistency, the prior of β_γ needs to be flat but not too flat.*

3.4 Examples of priors for model parameters

Theorem 3.1 proposes the general conditions for the prior of β_γ to achieve posterior model probability consistency. In this section, we explore model selection consistency under several priors. The priors we consider are frequently adopted in Bayesian model selection studies.

We establish model selection consistency under each prior by applying Theorem 3.1. The results imply that the sufficient conditions also provide useful guidelines for hyperparameter specification for a chosen prior.

3.4.1 Gaussian prior

As a conjugate prior, the Gaussian prior is the one of the most popular prior choices in the Bayesian model selection (George and McCulloch, 1993; Ishwaran and Rao, 2005; Narisetty et al., 2014) with Gaussian data. Suppose we assign the independent Gaussian prior to β_γ in the following way:

$$p(\beta_\gamma \mid \gamma, s) \quad = \quad \prod_{j \in \gamma} \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{1}{2s^2}\beta_j^2\right).$$

Under Assumption 3.2-3.4, the independent Gaussian prior yields consistent model selection if $s > p_n^2/\sqrt{n}$ and $p_{\gamma_*} \log s = o(n)$. The proof is given in Appendix B.8.

3.4.2 Laplace prior

As the prior of β_γ in Bayesian lasso (Park and Casella, 2008), the Laplace prior is another popular prior choice in Bayesian model selection (Casella et al., 2010; Hans, 2009, 2010). Suppose we assign the Laplace prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma, s) = \prod_{j=1}^{p_\gamma} \frac{1}{2s} \exp\left(-\frac{|\beta_j|}{s}\right).$$

Under Assumption 3.2-3.4, the Laplace prior yields consistent model selection if $s > p_n^2/\sqrt{n}$ and $p_{\gamma_*} \log s = o(n)$. The proof is given in Appendix B.8.

3.4.3 Scaled Student's t prior

The scaled Student's t prior (West, 1987) concentrates at zero with relatively thicker tails, which makes the scaled Student's t prior an appealing prior choice for model selection (Ar-

magan et al., 2011, 2013; Tipping, 2001). Suppose we assign scaled Student's t prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma, s, d) = \prod_{j=1}^{p_\gamma} [sd^{1/2}B(d/2, 1/2)]^{-1}(1 + \frac{\beta_j^2}{sd})^{-(d+1)/2}$$

where s is the scale parameter and d is the degrees of freedom. Under Assumption 3.2-3.4, the scaled Student's t prior yields consistent model selection if $s > p_n^2/\sqrt{n}$, $p_{\gamma_*} \log s = o(n)$, $\log d = o(n)$ and $p_{\gamma_*} \log d = o(n)$. The proof is given in Appendix B.8.

3.4.4 Generalized double Pareto prior

The generalized double Pareto prior proposed by Armagan et al. (2010) has some appealing properties such as spike at zero and Student's t-like tails which make it a frequent prior choice in Bayesian model selection(Armagan et al., 2013; Pal et al., 2017). Suppose we assign the generalized double Pareto prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma, \alpha, \eta) = \prod_{j=1}^{p_\gamma} \frac{\alpha}{2\eta} \left(1 + \frac{|\beta_j|}{\eta}\right)^{-(\alpha+1)}.$$

Under Assumption 3.2-3.4, the Generalized double Pareto prior yields consistent model selection if $\frac{\eta}{\alpha} > p_n^2/\sqrt{n}$ and $p_{\gamma_*} \log \frac{\eta}{\alpha} = o(n)$. The proof is given in Appendix B.8.

3.4.5 Horseshoe prior

The Horseshoe prior has some fairly desirable properties, such as heavy tail, an infinite spike at zero and so on. Due to such properties, it becomes a popular prior choice in Bayesian regression (Armagan et al., 2013; Carvalho et al., 2010). Suppose we assign the Horseshoe prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma, \tau) = \prod_{i=1}^{p_\gamma} K_0(\tau^2)^{-1/2} \exp(\frac{\beta_i^2}{2\tau^2}) E_1(\frac{\beta_i^2}{2\tau^2}),$$

where $K_0 = 1/(2\pi^3)^{1/2}$ and $E_1(\cdot)$ is the exponential integral function. Under Assumption 3.2-3.4, the Horseshoe prior yields consistent model selection if $\tau^2 > p_n^4/n$ and $p_{\gamma_*} \log \tau^2 = o(n)$. The proof is given in Appendix B.8.

3.5 Estimating unknown variance

It is worth noting that in deriving the main results, we assume the σ^2 to be known which is an assumption not always satisfied. In real practice, the σ^2 is usually unknown and the best solution is to replace it by its estimate. However, it is not clear how to estimate σ^2 in model selection study since we do not know the true model.

We use forward selection method to estimate σ^2. The forward selection method is able to estimate σ^2 consistently under two following conditions.

Assumption 3.9. *For any $\gamma \in \mathcal{M}_n$, we have $\|\beta^0_{\gamma_*}\|_2 \leq O(n^c)$ and $\beta^0_{\min} \geq \nu_{\beta^0_{\gamma_*}} n^{-c_{\min}}$, where $\beta^0_{\min} = \min_{j \in \gamma_*} |\beta^0_j|$ and $\nu_{\beta^0_{\gamma_*}}$ is a constant.*

Assumption 3.10. *There exists constants c_0 and ν such that $\log p_n \leq \nu n^{c_0}$, and $c_0 + 6c + 12c_{\min} < 1$.*

Let $\gamma^{(k)}$ denote the selected model in the kth step of the forward selection. Theorem 3.2 ensures that a model which contains the true model will be selected within a number of steps much smaller than the sample size.

Theorem 3.2. *Under Assumptions 3.9 and 3.10, we have*

$$p(\gamma_* \subset \gamma^{(K' n^{4c+4c_{\min}})}) \to 1,$$

as $n \to \infty$, where $K' = 2\lambda_{\max}\nu^2\lambda_{\min}^{-2}\nu_\beta^{-4}$.

The proof of Theorem 3.2 is given in Appendix B.10. By Theorem 3.2, if we set the number of steps in the forward selection to be K_n, then we are able to select an overfitting model consistently. Since an overfitting model can be consider as true model with some extra

predictor variables with zero regression coefficients, Theorem 3.2 enables us to estimate σ^2 consistently under an overfitting model.

3.6 Simulation study

In this section, we conduct the simulation study to exam the performance of Theorem 3.1. We use the priors discussed in the last section and construct the model selection procedures according to the sufficient conditions derived for each prior.

3.6.1 Simulation setting

First, we generate our data $\{(y_i, x_i) : i = 1, \ldots, n\}$ from the model $y_i = x_i^\top \beta + \epsilon_i$ for $i = 1, \ldots, n$, where

$$\begin{aligned} x_i &\overset{iid}{\sim} \mathcal{N}_{p_n}(0, \Sigma), \\ \beta_{\gamma_*} &= (1, 1, -1, -1)^\top, \\ \epsilon_i &\overset{iid}{\sim} \mathcal{N}(0, 1.3), \end{aligned}$$

with $\Sigma = (\sigma_{ij})_{p_n \times p_n}$ and $\sigma_{ij} = \rho^{|i-j|}$. The true model γ_* is a randomly generated index set with $p_{\gamma_*} = 4$. We consider two scenarios:(*Scenario* 1) $p_n = \lfloor n^{1.1} \rfloor$ and $\rho = 0.5$;(*Scenario* 2) $p_n = \lfloor n^{1.1} \rfloor$ and $\rho = 0$. The predictor variables are correlated in *Scenario* 1, while the predictor variables are uncorrelated in *Scenario* 2. By Assumption 3.3, we set $K_n = n^{2/3}$. To exam the limiting behavior of the Bayesian model selection as n increases, we run the simulation with $n = \{100, 300, 500, 900\}$ for each scenario.

For the specification of the hyperparameters, we consider 5 cases for each prior. Table 3.1 presents our choices for the values of the hyperparameters. By the sufficient conditions we derived for the priors, the settings in Case 1 and Case 5 violate the sufficient conditions, whereas the settings in Case 2, Case 3 and Case 4 satisfy the sufficient conditions. We expect to observe model selection consistency only in Case2, Case 3 and Case 4 under each prior.

Table 3.1: *Choices of hyperparameters*

Prior	**Case 1**	**Case 2**	**Case 3**	**Case4**	**Case5**
Gaussian	$s = p$	$s = p^2$	$s = p^3$	$s = p^4$	$s = \exp(0.6n)$
Laplace	$s = p$	$s = p^2$	$s = p^3$	$s = p^4$	$s = \exp(0.6n)$
Scaled student's t	$s = p$, $d = n$	$s = p^2$, $d = n$	$s = p^3$, $d = n$	$s = p^4$, $d = n$	$s = \exp(0.6n)$, $d = n$
Pareto	$\frac{\eta}{\alpha} = p$, $\alpha = n$	$\frac{\eta}{\alpha} = p^2$, $\alpha = n$	$\frac{\eta}{\alpha} = p^3$, $\alpha = n$	$\frac{\eta}{\alpha} = p^4$, $\alpha = n$	$\frac{\eta}{\alpha} = \exp(0.6n)$, $\alpha = n$
Horseshoe	$\tau = p$	$\tau = p^2$	$\tau = p^3$	$\tau = p^4$	$\tau = \exp(0.6n)$

To evaluate the performance of the model selection, we compute the posterior probability of the true model using a Monte Carlo estimator:

$$pr(\gamma_* \mid y) \approx T^{-1} \sum_{t=1}^{T} \mathbb{I}(\gamma^{(t)} = \gamma_*),$$

where $\{\gamma^{(t)} : t = 1, \ldots, T\}$ is a Gibbs sequence generated from $pr(\gamma \mid y)$. We totally generate 100 Monte Carlo experiments, and run the Gibbs sampler (see Section 3.6.2) $6,000$ times with first $3,000$ iterations as the burn-in period. We use $T = 1,000$ samples generated by resampling every third iteration after the burn-in period for each Monte Carlo experiment.

3.6.2 Gibbs sampler

In this section, we introduce our simulation algorithm. For each Monte Carlo experiment, we first run the forward selection to compute $\hat{\sigma}$. We then run the Gibbs sampler to generate the sample of selected models. The simulation algorithm can be implemented as follows:

Algorithm 2 The Gibbs sampler.

Set $\gamma_M = \gamma_0$, $A_1 = c(1, \ldots, p_n)$

Repeat for $h = 1, \ldots, K_n$:

Set $s_M = \gamma_M^{(h)}$, $C_1 = A_1^{(h)}$,$S_M = \{\Phi\}$

Repeat for $i \in C_1$:

Set $s_{M_i} = s_M \cup \{i\}$, $S_M = S_M \cup s_{M_i}$

Compute:

$$s_{M_i} = \underset{s_{M_i} \in S_M}{argmaxp}(y|\hat{\beta}_{s_{M_i}}, s_{M_i})$$

$s_M \leftarrow s_{M_i}$, $C_1 \leftarrow C_1 \setminus \{i\}$

Output $\gamma_M^{(h+1)} = s_M$, $A_1^{(h+1)} = C_1$

$\nu^2 = \frac{1}{n - p_{\gamma_M}} \|y - H_{\gamma_M} y\|_2^2$

Repeat for $t = 1, \ldots, T$:

Set $s = \gamma^{(t)}$

Repeat for $j = 1, \ldots, p_n$:

Set $s_1 = s \cup \{j\}$

Set $s_0 = s \setminus \{j\}$

Compute

$$p(y|s_1) = p(y \mid \hat{\beta}_{s_1}) p(\hat{\beta}_{s_1}|s_1) |X_{s_1}^\top X_{s_1}|^{-1/2} (2\pi\nu^2)^{p_{s_1}/2}$$

$$p(y|s_0) = p(y \mid \hat{\beta}_{s_0}) p(\hat{\beta}_{s_0}|s_0) |X_{s_0}^\top X_{s_0}|^{-1/2} (2\pi\nu^2)^{p_{s_0}/2}$$

$$w = p(y|s_1) \mathbb{I}(s_1 \in \mathcal{M}_n) / \{p(y \mid s_1) + p(y \mid s_0)\}$$

Sample $z \sim$ Bernoulli(w)

If $z = 1$ then $s \leftarrow s_1$;

else $s \leftarrow s_0$

Output $\gamma^{(t+1)} = s$

3.6.3 Simulation results

Table 3.2 and Table 3.3 report the mean and standard deviation of the estimate of $p(\gamma_*|y)$ over the 100 Monte Carlo experiments of *Scenario* 1. Clearly, we observe a strong trend of $p(\gamma_*|y)$ increasing to one in Case 2, Case 3 and Case 4 under each prior. On the contrary, we fail to observe $p(\gamma_*|y)$ increasing to one in Case 1 and Case 5. The results match our expectation.

Table 3.2: *Simulation results based on 100 Monte Carlo experiments: Scenario 1.*

Prior	**Case**	**n=100**	**n=200**	**n=500**	**n=900**
	1	0.1575(0.1951)	0.1190(0.1721)	0.1924(0.1728)	0.1637(0.1915)
	2	0.8307(0.2765)	0.9243(0.1595)	0.9305(0.2098)	0.9874(0.0650)
Gaussian	3	0.9656(0.1425)	0.9865(0.1017)	0.9958(0.0389)	0.9999(0.0001)
	4	0.9719(0.1354)	0.9932(0.0515)	1.0000(0.0000)	1.0000(0.0000)
	5	0.1497(0.3549)	0.0945(0.2859)	0.0076(0.0745)	0.0000(0.0000)
	1	0.1064(0.1595)	0.0987(0.1372)	0.1441(0.1616)	0.1349(0.1695)
	2	0.7923(0.3080)	0.8914(0.2178)	0.9594(0.1468)	0.9814(0.0917)
Laplace	3	0.9517(0.1746)	0.9959(0.0176)	0.9998(0.0022)	0.9999(0.0001)
	4	0.9563(0.1800)	0.9999(0.2792)	1.0000(0.0000)	1.0000(0.0000)
	5	0.1336(0.3346)	0.0918(0.2845)	0.0100(0.0986)	0.0000(0.0000)
	1	0.1550(0.1962)	0.1120(0.1525)	0.1924(0.1729)	0.1638(0.1915)
Scaled	2	0.7964(0.2906)	0.8571(0.2562)	0.9305(0.2098)	0.9924(0.02520)
Student's	3	0.9594(0.1535)	0.9865(0.1017)	0.9958(0.0389)	0.9999(0.0001)
t	4	0.9719(0.1355)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
	5	0.1497(0.3549)	0.1093(0.3070)	0.0076(0.07450)	0.0000(0.0000)

Table 3.3: *Simulation results based on 100 Monte Carlo experiments: Scenario 1.*

Prior	Case	n=100	n=200	n=500	n=900
Pareto	1	0.1200(0.1722)	0.0957(0.1402)	0.1685(0.1843)	0.1638(0.1916)
	2	0.8038(0.2947)	0.9038(0.2146)	0.9638(0.1324)	0.9870(0.0694)
	3	0.9617(0.1660)	0.9968(0.0156)	0.9999(0.0002)	0.9999(0.0001)
	4	0.9774(0.1210)	0.9999(0.0004)	1.0000(0.0000)	1.0000(0.0000)
	5	0.1637(0.3556)	0.1343(0.3282)	0.0411(0.1970)	0.0000(0.0000)
Horseshoe	1	0.2228(0.2644)	0.2487(0.25810)	0.3674(0.2709)	0.3518(0.2631)
	2	0.8110(0.3148)	0.9070(0.2168)	0.9886(0.0579)	0.9945(0.0379)
	3	0.9558(0.1848)	0.9956(0.0354)	0.9999(0.0002)	0.9999(0.0001)
	4	0.9580(0.1776)	0.9998(0.0018)	1.0000(0.0000)	1.0000(0.0000)
	5	0.1230(0.3086)	0.1859(0.37770)	0.0400(0.1969)	0.0010(0.0707)

Table 3.4 and Table 3.5 report the mean and standard deviation of the estimate of $p(\gamma_*|y)$ over the 100 Monte Carlo experiments of *Scenario* 2. We only observe posterior model probability of the true model tends to one in Case 2, Case 3 and Case 4 under each prior which is similar to *Scenario* 1. The results match our expectation as well.

Figure 3.1 - Figure 3.5 are the trace plots of the estimated $p(\gamma_*|y)$ as sample size increases under each prior. The plots demonstrate that when the sufficient conditions are met, the trace of $p(\gamma_*|y)$ shows a strong trend of increasing to one as the sample size increases. When the sufficient conditions are violated, the trace of $p(\gamma_*|y)$ fails to show such a trend as the sample size increases.

In general, the simulation results demonstrate that, when the priors satisfy the sufficient conditions in Theorem 3.1, the model selection achieves consistency in the sense that the posterior probability of the true model tends to one as the sample size increases. On the contrary, when the sufficient conditions are violated, the model selection fails to be consistent. The results imply that when the sufficient conditions are satisfied, the posterior model probability distribution degenerates to a point mass distribution concentrates at the true

model as the sample size increases.

Table 3.4: *Simulation results based on 100 Monte Carlo experiments: Scenario II.*

Prior	**Case**	**n=100**	**n=200**	**n=500**	**n=900**
	1	0.1267(0.1854)	0.0624(0.1227)	0.1452(0.1657)	0.1374(0.1695)
	2	0.7512(0.3358)	0.8570(0.2562)	0.9541(0.1507)	0.9871(0.0304)
Gaussian	3	0.9609(0.1417)	0.9873(0.0719)	0.9997(0.0018)	0.9999(0.0001)
	4	0.9719(0.1391)	0.9999(0.0001)	1.0000(0.0000)	1.0000(0.0000)
	5	0.1530(0.3376)	0.1285(0.3134)	0.0201(0.1407)	0.0099(0.0699)
	1	0.0719(0.1374)	0.0677(0.1104)	0.1105(0.1397)	0.0949(0.1397)
	2	0.7692(0.3168)	0.8265(0.2800)	0.9369(0.1734)	0.9842(0.0371)
Laplace	3	0.9480(0.1750)	0.9856(0.0783)	0.9998(0.0016)	0.9999(0.0001)
	4	0.9619(0.1575)	0.9999(0.0001)	1.0000(0.0000)	1.0000(0.0000)
	5	0.2475(0.4137)	0.0828(0.2644)	0.0310(0.1715)	0.0000(0.0000)
	1	0.0986(0.1470)	0.0625(0.1229)	0.1453(0.1658)	0.1375(0.1453)
Scaled	2	0.7933(0.3072)	0.8571(0.2562)	0.9436(0.1643)	0.9871(0.0304)
Student's	3	0.9615(0.1398)	0.9873 (0.0718)	0.9997(0.0018)	0.9999(0.0001)
t	4	0.9719(0.1391)	0.9999(0.0001)	1.0000(0.0000)	1.0000(0.0000)
	5	0.1609(0.3619)	0.1285(0.3135)	0.0201(0.1407)	0.0099(0.0699)

Table 3.5: *Simulation results based on 100 Monte Carlo experiments: Scenario II.*

Prior	**Case**	**n=100**	**n=200**	**n=500**	**n=900**
	1	0.0850(0.1435)	0.0677(0.1104)	0.1105(0.1397)	0.1053(0.1453)
	2	0.7102(0.3389)	0.8265(0.2800)	0.9491(0.1566)	0.9840(0.0370)
Pareto	3	0.9536(0.1589)	0.9873 (0.0713)	0.9995(0.0039)	0.9999(0.0001)
	4	0.9914(0.0337)	0.9999(0.0001)	1.0000(0.0000)	1.0000(0.0000)
	5	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
	1	0.0104(0.0254)	0.0011(0.0037)	0.0005(0.0015)	0.0003(0.0013)
	2	0.8314(0.2754)	0.8596(0.3031)	0.9672(0.1256)	0.9956(0.0103)
Horseshoe	3	0.9676(0.1478)	0.9954(0.0293)	0.9999(0.0005)	1.0000(0.0000)
	4	0.9849(0.0973)	0.9999(0.0001)	1.0000(0.00000)	1.0000(0.0000)
	5	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)

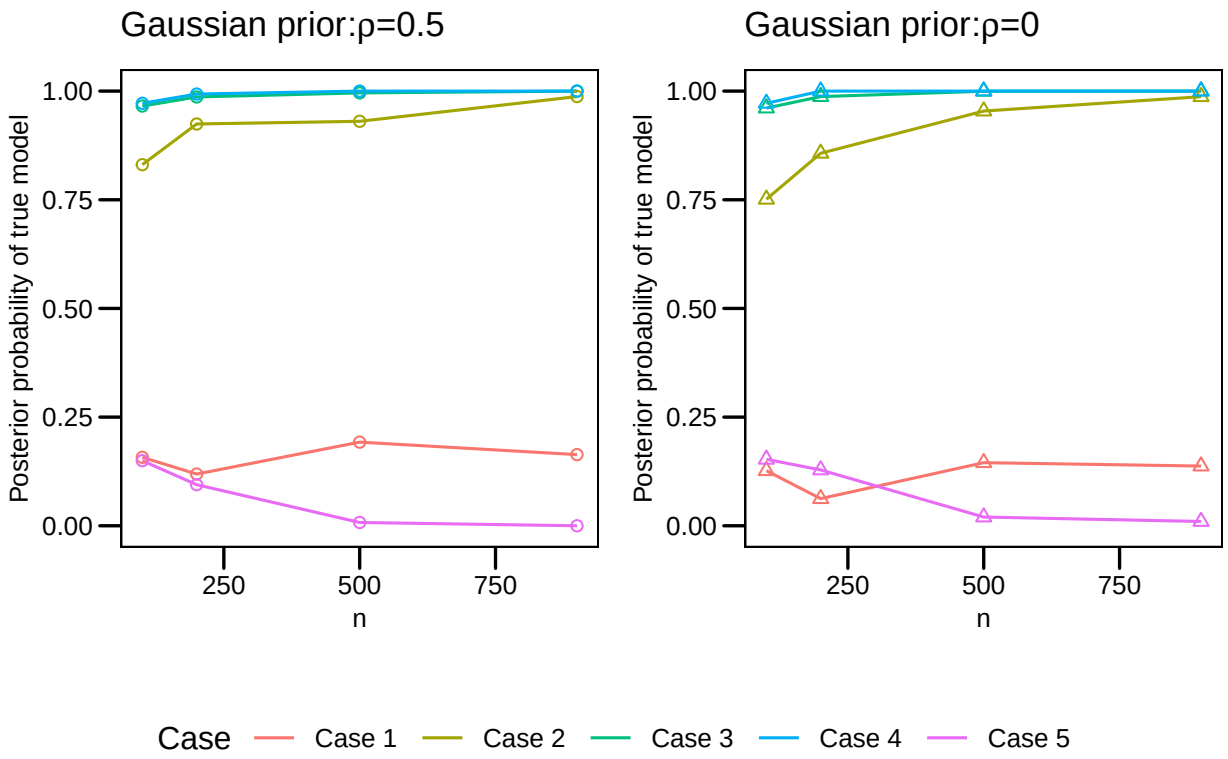

Figure 3.1: *The trace plots of the estimate of* $pr(\gamma_* \mid y)$ *as the sample size* n *increases.*

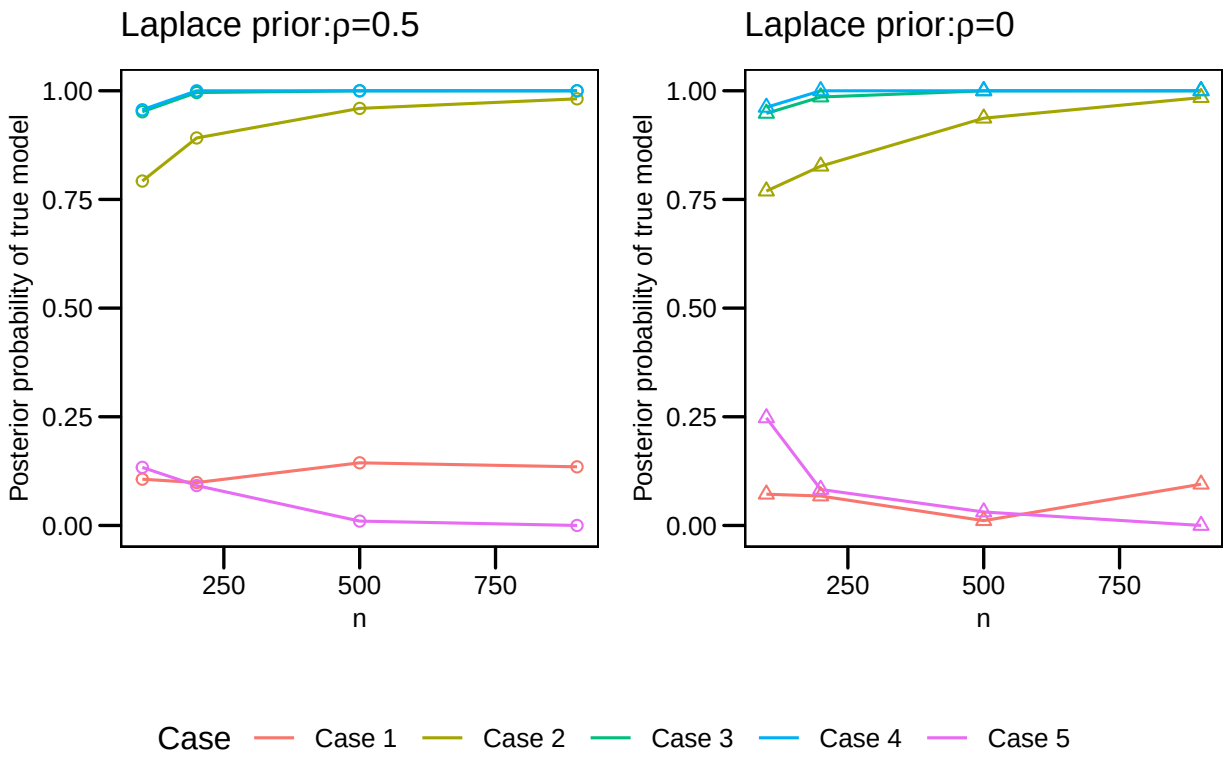

Figure 3.2: *The trace plots of the estimate of* $pr(\gamma_* \mid y)$ *as the sample size* n *increases.*

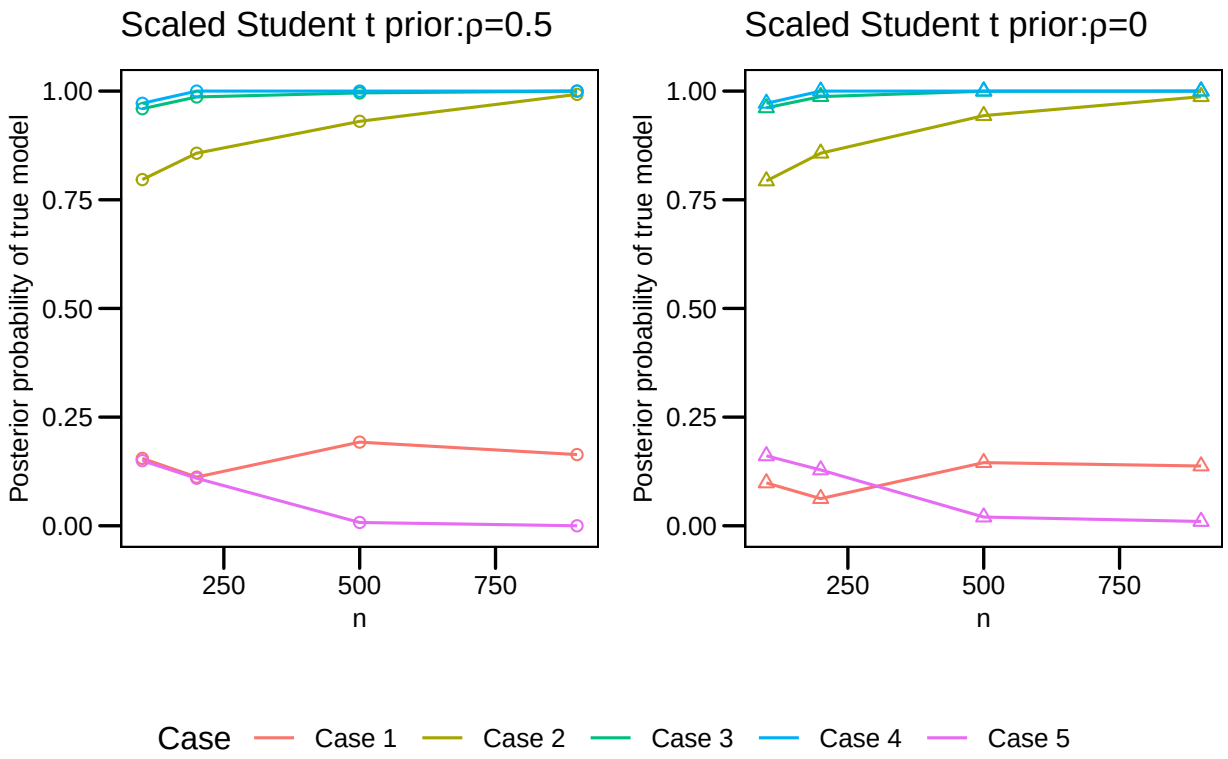

Figure 3.3: *The trace plots of the estimate of* $pr(\gamma_* \mid y)$ *as the sample size* n *increases.*

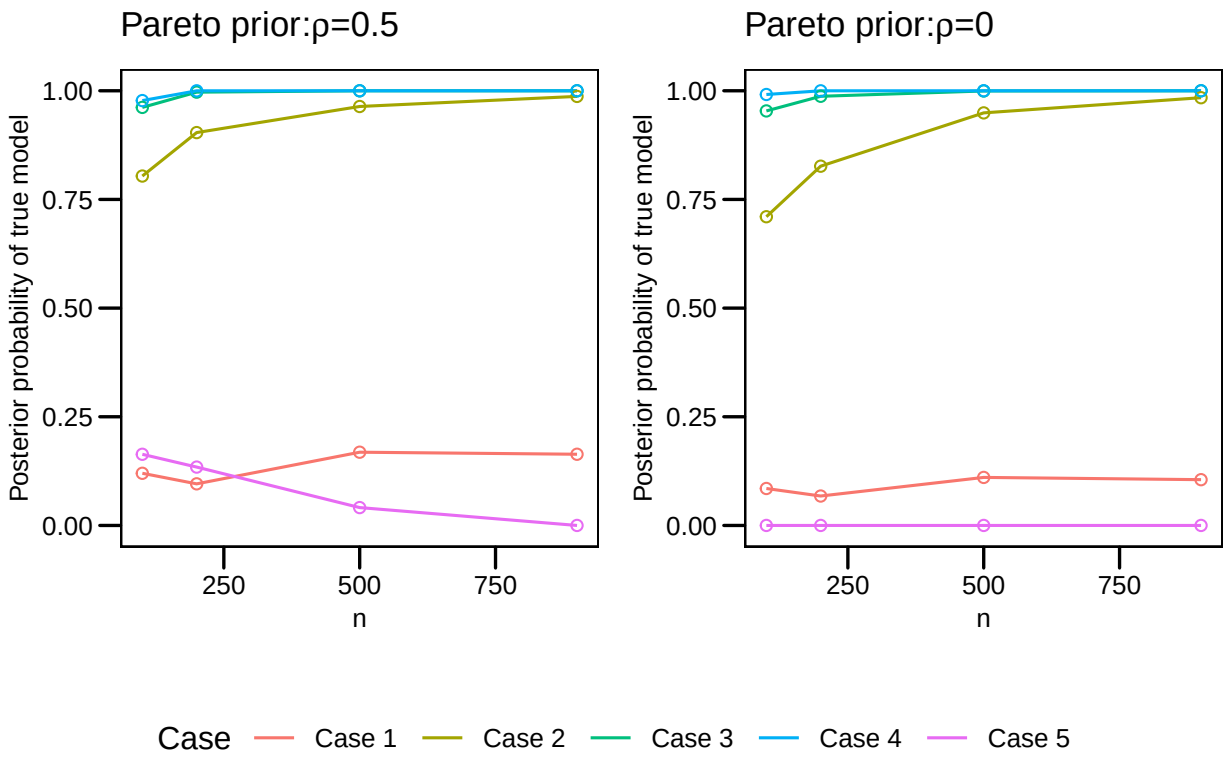

Figure 3.4: *The trace plots of the estimate of* $pr(\gamma_* \mid y)$ *as the sample size* n *increases.*

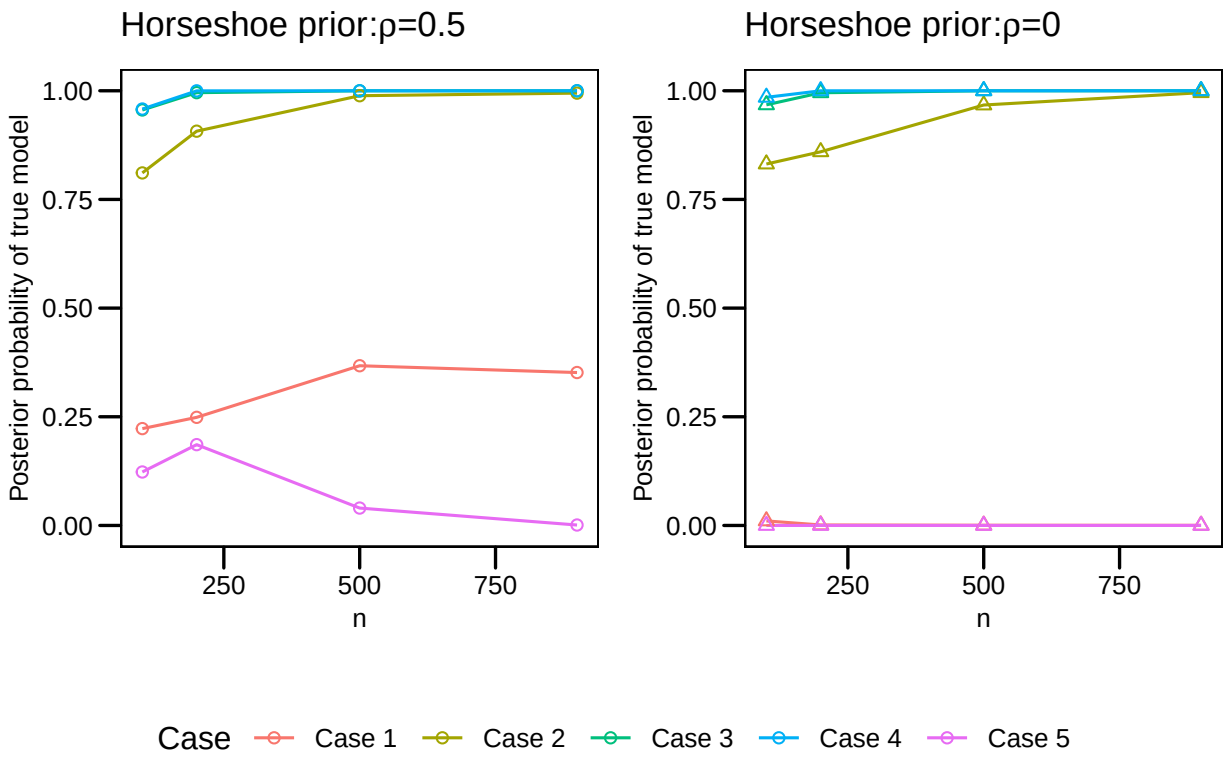

Figure 3.5: *The trace plots of the estimate of* $pr(\gamma_* \mid y)$ *as the sample size* n *increases.*

3.7 Real data study

3.7.1 Set-up

In this section, we examine the performance of Theorem 3.1 with real data. We use the $Bardet - Biedl$ syndrome gene expression data from Scheetz et al. (2006). The data set contains information of the expression level of the TRIM 32 gene and 200 gene probes of 120 rats. There are totally 120 observations with one response variable and 200 predictor variables in the data set. The goal of our study is to identify the genes that are highly associated with the expression level of the TRIM 32 gene.

We conduct the model selection using the same algorithm from the simulation study. We run the Gibbs sampler $6,000$ times with the first $3,000$ iterations as the burn-in period. We

then compute the relative frequency of each selected model, and select the model with the highest relative frequency as our estimate of the true model. We compute the Extended Bayesian Information Criteria(Chen and Chen, 2008) for model evaluation.

Table 3.6 presents the settings of the hyperparameter values, which are the same as the hyperparameter settings in the simulation study. We already know that only the settings in Case 2, Case 3 and Case 4 satisfy the sufficient conditions. We also run Lasso and SCAD regressions in comparison to our method.

Table 3.6: *Choices of hyperparameters*

Prior	**Case 1**	**Case 2**	**Case 3**	**Case4**	**Case5**
Gaussian	$s = p$	$s = p^2$	$s = p^3$	$s = p^4$	$s = \exp(0.6n)$
Laplace	$s = p$	$s = p^2$	$s = p^3$	$s = p^4$	$s = \exp(0.6n)$
Scaled student's t	$s = p$, $d = n$	$s = p^2$, $d = n$	$s = p^3$, $d = n$	$s = p^4$, $d = n$	$s = \exp(0.6n)$, $d = n$
Pareto	$\frac{\eta}{\alpha} = p$, $\alpha = n$	$\frac{\eta}{\alpha} = p^2$, $\alpha = n$	$\frac{\eta}{\alpha} = p^3$, $\alpha = n$	$\frac{\eta}{\alpha} = p^4$, $\alpha = n$	$\frac{\eta}{\alpha} = \exp(0.6n)$, $\alpha = n$
Horseshoe	$\tau = p$	$\tau = p^2$	$\tau = p^3$	$\tau = p^4$	$\tau^2 = \exp(0.6n)$

3.7.2 Results

Table 3.7 - Table 3.8 present the results of the real data study. There are several implications of the results which are worth noting. First, when the specified values of hyperparameters meet the sufficient conditions, the selected model is the best among all the selected models in terms of EBIC. Second, despite of the differences in the simulation settings, the model selection procedures unanimously select the same model when the priors satisfy the sufficient conditions. Thirdly, the selected models are inferior in terms of EBIC when the priors fail to meet the sufficient conditions. However, these models all contain the same genes selected by the best model. Lastly, the Lasso regression tends to select models that are relatively larger. The results in the SCAD regression resemble the results in Case 1 under each prior. The two models selected by Lasso and SCAD regression have larger EBIC, nevertheless, they both

select the 3 genes which are select by the best model.

Figure 3.6 is the heat map of the selected gene. The map shows that gene 153, 180 and 185 are selected by most of the models. The heat map indicates that gene 153, 180 and 185 have higher association with the expression level of the TRIM 32 gene.

In general, the real data study confirms our expectation. When the sufficient conditions are satisfied, the model selection tends to select the model which shows superiority.

Table 3.7: *Real data results.*

Prior	**Case**	**Selected Gene**	**EBIC**
	1	{87,153,180,185}	226.4538
	2	{153,180,185}	223.7867
Gaussian	3	{153,180,185}	223.7867
	4	{153,180,185}	223.7867
	5	{153}	251.5057
	1	{55,76,87,153,180,185}	234.2658
	2	{153,180,185}	223.7867
Laplace	3	{153,180,185}	223.7867
	4	{153,180,185}	223.7867
	5	{153}	251.5057
	1	{55,71,76,110,153,180,185}	237.8742
Scaled	2	{153,180,185}	223.7867
Student's	3	{153,180,185}	223.7867
t	4	{153,180,185}	223.7867
	5	{153}	251.5057

Note: The numbers represent the selected genes, e.g. 31 means the 31th gene is selected.

Table 3.8: *Real data results.*

Prior	Case	Selected Gene	EBIC
Pareto	1	{55,76,87,153,180,185}	234.2658
	2	{153,180,185}	223.7867
	3	{153,180,185}	223.7867
	4	{153,180,185}	223.7867
	5	{153}	251.5057
Horseshoe	1	{87,153,180,185}	226.4538
	2	{153,180,185}	223.7867
	3	{153,180,185}	223.7867
	4	{153,180,185}	223.7867
	5	{153}	251.5057
Lasso	$\lambda = 0.0056$	{11,50,54,62,76,87,90,96,...,200}	365.7635
SCAD	$a = 3.7,\ \lambda = 0.0120$	{87,153,180,181,185,200}	279.5519

Note: The numbers represent the selected genes, e.g. 31 means the 31th gene is selected.

Lasso:{11,50,54,62,76,87,90,102,110,127,134,136,140,146,153,155,161,164,180,184,185,187,188,200}

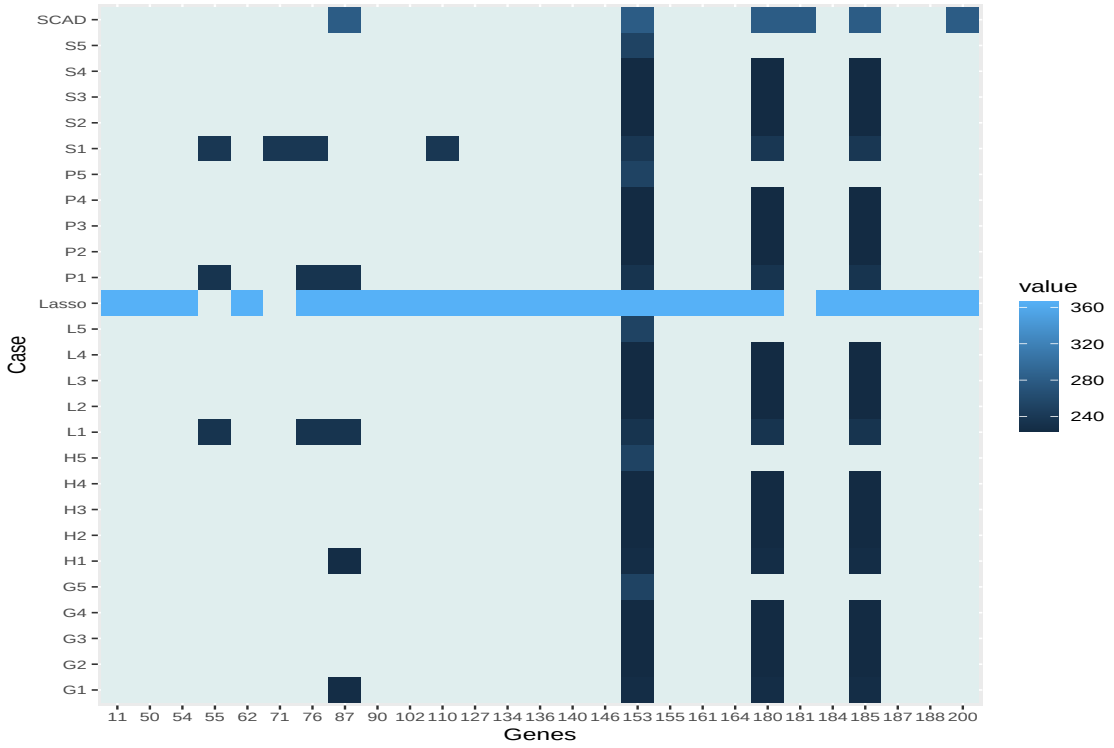

Figure 3.6: *Heat Map of Selected Genes*

3.8 Discussion

Theorem 3.1 proposes general sufficient conditions for the $p(\beta_\gamma|y)$ to induce posterior model probability consistency. We can establish model selection consistency under specific priors by applying the sufficient conditions. The sufficient conditions also provide useful guidelines for hyperparameter specifications. Even though we only demonstrate the consistency under the high-dimensional settings, Theorem 3.1 holds in low-dimensional model selection as well.

It is worth noting that, we assume σ^2 is known in our model set-up. We believe this assumption is not essential for Theorem 3.1 to hold. In cases where σ^2 is unknown, the forward selection method offers a way to estimate σ^2 consistently, which enables us to achieve model selection consistency without knowing σ^2.

While we consider the sufficient conditions in Theorem 3.1 as general conditions, the analysis still relies on the normal likelihood assumption. Thus, another interesting direction

for the future work is to extend the results to Bayesian generalized linear models by relaxing the normal likelihood assumption, which is discussed in the following chapter.

Chapter 4

The consistency of generalized Bayesian high-dimensional variable selection under arbitrary priors

4.1 Introduction

Bayesian model selection has enjoy increasingly popularity in the area of the generalized linear models. A few examples of Bayesian model selection in generalized linear regression are George and Foster (2000), Raftery (1996), Bernardo et al. (1999), Chen et al. (1999) and Ntzoufras et al. (2003). Under the Bayesian framework, we assign priors to model parameters and candidate models respectively, and compute the posterior probability for each candidate the model. By the Bayes theorem, the posterior probability of model γ is

$$\text{pr}(\gamma|y) = \frac{p(y|\gamma)p(\gamma)}{\sum_{\gamma\in\mathcal{M}_n} p(y|\gamma)p(\gamma)}$$

where

$$p(y|\gamma) = \int p(y|\beta_\gamma, \gamma)p(\beta_\gamma|\gamma)d\beta_\gamma,$$

and $\mathcal{M}_n$ is the set of candidate models under consideration. The true model is believed to be the model which maximizes the posterior model probability. Therefore, the model with the highest posterior model probability is usually selected as the true model.

In deriving the posterior model probability, $p(\beta_\gamma|y)$ plays a key role. For the Bayesian model selection using the posterior model probability, the specification of $p(\beta_\gamma|y)$ can be challenging, especially in the high-dimensional regression where we have numerous variables. For more discussion of prior specification, please refer to Moreno et al. (2015), Shang and Clayton (2011) and Liang et al. (2008). Despite the large number of studies using the Bayesian approach in high-dimensional model selection, the studies focus on the prior specification are very limited. We do not have a clear picture of how the choice of priors affects the model selection results in the generalized linear regression.

In this chapter, we extend our framework to Bayesian generalized linear regression models. In this generalization, we do not impose any specific distribution on the data which is a distinctive difference of the study in this chapter. Our approach to the model selection consistency focuses on the marginal likelihood. If we consider uniform distribution for $\text{pr}(\gamma)$, then the model with the greatest marginal likelihood is equivalent to model which maximizes the posterior model probability. Thus, if the model selection achieves marginal likelihood consistency, the posterior probability of the true model should become greater than any other candidate models asymptotically due to the equivalent.

In this chapter, we develop general conditions for $p(\beta_\gamma|\gamma)$ under which the true model tends to maximize the marginal likelihood even when the number of predictors increases with the sample size, in some cases, even faster than that of the sample size. The conditions provide useful guidelines for the specification of priors as well as the hyperparameter in the priors. Our simulation study demonstrates the validity of the our sufficient conditions for Bayesian model selection consistency with non-Gaussian data.

4.2 Model set-up

Consider a generalized linear regression model,

$$E(y|\beta) = g^{-1}(X\beta), \tag{4.1}$$

where $y = (y_1, \ldots, y_n)^\top$ is the n-dimensional response vector, $X = (x_1, \ldots, x_{p_n})$ is the $n \times p_n$ design matrix with $x_j = (x_{1j}, \ldots, x_{nj})^\top$, $\beta = (\beta_1, \ldots, \beta_{p_n})^\top$ is a p_n-dimensional regression coefficient vector, and $g^{-1}(\cdot)$ is the inverse link function.

In the generalized linear regression, the model selection is performed by identifying a subset of predictor variables that have non-zero regression coefficients. Let $\gamma \subset \{1, \ldots, p_n\}$ be an index set of size p_γ corresponding to a reduced model such that

$$E(y|\beta_\gamma, \gamma) = g^{-1}(X_\gamma \beta_\gamma),$$

where X_γ is the $n \times p_\gamma$ submatrix of X and β_γ is the p_γ-dimensional vector of non-zero regression coefficients. The goal of the model selection is to identify the subset γ_* which contain all the predictor variables with non-zero regression coefficients, i.e., the true model.

Let $\mathcal{M}_n = \{\gamma : p_\gamma \leq K_n\}$ be set of candidate models under our consideration, where p_γ is the size of the model γ and K_n is the upper bound of the candidate model size. Let $\mathrm{pr}(\gamma)$ be the prior probability of γ being the true model, we consider a uniform distribution for the candidate models which is given by

$$\mathrm{pr}(\gamma) \propto \mathbb{I}(\gamma \in \mathcal{M}_n),$$

where $\mathbb{I}$ is an indicator function. It then follows that

$$\mathrm{pr}(\gamma|y) \propto p(y|\gamma)\mathrm{pr}(\gamma) \propto p(y|\gamma)$$

. Under this model set-up, we select the model with the greatest posterior model probability

as the true model, which is equivalent to selecting the model which maximizes marginal likelihood, that is,

$$\hat{\gamma} = arg\max_{\gamma \in \mathcal{M}_n} \text{pr}(\gamma|y) = arg\max_{\gamma \in \mathcal{M}_n} p(y|\gamma).$$

Thus, the consistency in the posterior model probability is equivalent to the consistency in the marginal likelihood, which implies the marginal likelihood of true model becomes greater than any other candidate models as the sample increases. Therefore, in this chapter, we formally define the Bayesian model selection consistency as follows:

Definition 4.1. *Let γ_* be the true model. The Bayesian model selection is consistent if*

$$p(y|\gamma_*) > \sup_{\gamma \in \mathcal{M}_n \setminus \{\gamma_*\}} p(y|\gamma),$$

in $\mathcal{Y}|\beta_{\gamma_*}$-probability as $n \to \infty$.

Our objective in this chapter is to derive sufficient conditions under which the true model maximizes the marginal likelihood in the Bayesian generalized linear regression while p_n grows with the sample size, possibly at a faster rate than n. To achieve this goal, we require the following regularity conditions.

Assumption 4.1. *The true data generating model is γ_*, and $\gamma_* \in \mathcal{M}_n$.*

Assumption 4.2. *Suppose that γ_0 and γ_1 are two candidate models such that $\gamma_0 \subset \gamma_1$. Let γ_* be the true model. If $\gamma_* \subset \gamma_0$, then as $n \to \infty$,*

$$-2log\frac{p(y|X\hat{\beta}_{\gamma_0})}{p(y|X\hat{\beta}_{\gamma_1})} \to \chi^2_{p_{\gamma_1} - p_{\gamma_0}}$$

in distribution, where $\hat{\beta}_\gamma$ denotes the maximum likelihood estimator under model γ.

Assumption 4.1 ensures that the true model is contained in the set of the candidate models we consider so that the model selection consistency can be properly defined. Assumption 4.2 requires the log likelihood ratio of two overfitting models (models which contain the true model) converges to χ^2 distribution with the degrees of freedom $p_{\gamma_1} - p_{\gamma_0}$ asymptotically.

Assumption 4.3. *As $n \to \infty$, we have*

$$-2\log p(y|\gamma) = -2\log p(y|X\hat{\beta}_\gamma) - 2\log p(\hat{\beta}_\gamma|\gamma) + p_\gamma \log n + p_\gamma c_\gamma,$$

where c_γ is a constant, which depends on γ.

Since we do not impose any specific distribution on data and $p(\beta_\gamma|y)$, the marginal likelihood is not available in the closed form. Assumption 4.3 enables us to analyze the marginal likelihood without knowing the closed form expression of the marginal likelihood.

Assumption 4.4. *For any models γ_0 and γ_1 such that $\gamma_* \not\subset \gamma_0$ but $\gamma_* \subset \gamma_1$, then as $n \to \infty$,*

$$-2\log \frac{p(y|X\hat{\beta}_{\gamma_0})}{p(y|X\hat{\beta}_{\gamma_1})} > a_{\gamma_0,\gamma_1} n,$$

where a_{γ_0,γ_1} is a positive constant which depends on γ_0 and γ_1.

Assumption 4.4 requires the log likelihood ratio of overfitting model and misspecified model to be bounded. The positive lower bound implies the pairwise comparison between the overfitting model and misspecified model achieve consistency.

Assumption 4.5. *$p_n = O(n^\alpha)$ for $\alpha \in [0, \infty)$.*

Assumption 4.6. *$K_n = O(n^b)$ for $b \in [0, 1)$.*

Assumption 4.7. *For any γ such that $\gamma \in \mathcal{M}_n$, there exists a positive constant d such that $\|\hat{\beta}_\gamma\|_2 \leq dn^c$ in probability as $n \to \infty$, where constant $c \in [0, b)$.*

Assumption 4.5 allows the total number of predictors p_n to grow with the sample size. Assumption 4.6 restricts the size of the candidate models to be smaller than the sample size by controlling the upper bound K_n. Assumption 4.7 allows the magnitude of $\hat{\beta}_\gamma$ to grow, but the growth of the magnitude is bounded by K_n, i.e. $\|\hat{\beta}_\gamma\|_2 \leq c_\gamma K_n$ for some constant $c_\gamma > 0$.

4.3 Main results

We begin our discussion of the model selection consistency with two cases of pairwise model comparison. We first consider the pairwise comparisons between the true model and the overfitting models, i.e. models which contain the true model.

Lemma 4.1. *Define* $\mathcal{M}_1 = \{\gamma : \gamma_* \subsetneq \gamma \text{ and } \gamma \in \mathcal{M}_n\}$. *For any* $\gamma \in \mathcal{M}_1$, *by Assumption 4.2, 4.3 and 4.5, we have*

$$p(y|\gamma_*) > \sup_{\gamma \in \mathcal{M}_1} p(y|\gamma),$$

as $n \to \infty$, *if* $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} \geq (p_\gamma - p_{\gamma_*}) \log \frac{p_n}{\sqrt{n}}$.

The proof of Lemma 4.1 is given in Appendix C.1. We assume that $\text{pr}(\gamma) \propto \mathbb{I}(\gamma \in \mathcal{M}_n)$ in our model set-up. It follows that $p(y|\gamma_*)/p(y|\gamma) = \text{pr}(\gamma_*|y)/\text{pr}(\gamma|y)$. Lemma 4.1 implies that, if prior of the model parameters satisfies the condition $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} \geq (p_\gamma - p_{\gamma_*}) \log \frac{p_n}{\sqrt{n}}$, we achieve the consistency in pairwise model comparison between the true model and the overfitting models in the sense that $\text{pr}(y|\gamma_*) > \sup_{\gamma \in \mathcal{M}_1} \text{pr}(\gamma|y)$.

The next case we consider is the pairwise comparison between the true model and the non-overfitting models. Non-overfitting models include underfitting models and misspecified models. The underfitting models refer to models which are submodels of the true model, and the misspecified models refer to models which neither contain the true model nor are submodels of the true model.

Lemma 4.2. *Define* $\mathcal{M}_2 = \{\gamma : \gamma_* \not\subset \gamma \text{ and } \gamma \in \mathcal{M}_n\}$. *For any* $\gamma \in \mathcal{M}_2$, *by Assumption 4.1, 4.3 - 4.6, we have*

$$p(y|\gamma_*) > \sup_{\gamma \in \mathcal{M}_2} p(y|\gamma),$$

as $n \to \infty$, *if* $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o(n)$.

The proof of Lemma 4.2 is given in Appendix C.2. Lemma 4.2 implies the pairwise model comparison consistency holds in the comparisons between the true model and the non-overfitting models in the sense that $\text{pr}(\gamma_*|y) > \sup_{\gamma \in \mathcal{M}_2} \text{pr}(\gamma|y)$, under the condition that $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o(n)$.

Theorem 4.1. *Under Assumption 4.1 - 4.6, we have*

$$p(y|\gamma_*) > \sup_{\gamma \in \mathcal{M}_n \setminus \{\gamma_*\}} p(y|\gamma),$$

in probability as $n \to \infty$ *if*

Condition 1: $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} \geq (p_\gamma - p_{\gamma_*}) \log \frac{p_n}{\sqrt{n}}$ *for any* $\gamma \in \mathcal{M}_1$*;*

Condition 2: $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o(n)$ *for any* $\gamma \in \mathcal{M}_2$*.*

Proof. See the proof of Lemma 1 and Lemma 2. □

Theorem 4.1 ensures that if the prior of β_γ meets the two sufficient conditions, the model selection consistency can be achieve in the sense that $\text{pr}(\gamma_*|y) > \sup_{\gamma \in \mathcal{M}_n \setminus \{\gamma_*\}} \text{pr}(\gamma|y)$ as sample size increases.

4.4 Examples of priors for model parameters

Theorem 4.1 proposes two general conditions for the prior of β_γ to yield model selection consistency in the generalized linear regression. The sufficient conditions in Theorem 1 provide useful guidelines for the specification of hyperparameters. In this section, we select several priors from the shrinkage prior family. These priors are frequently adopted in Bayesian model selection studies. We derive conditions for these priors to achieve model selection consistency by applying Theorem 4.1.

4.4.1 Gaussian prior

The Gaussian prior is the one of the most popular prior choices in the Bayesian model selection (George and McCulloch, 1993; Ishwaran and Rao, 2005; Narisetty et al., 2014). Suppose we assign the independent Gaussian prior to β_γ in the following way:

$$p(\beta_\gamma \mid \gamma, s) \quad = \quad \prod_{j \in \gamma} \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{1}{2s^2}\beta_j^2\right).$$

Under Assumption 4.1, 4.5 - 4.7, the independent Gaussian prior yields consistent model selection if $s > p_n/\sqrt{n}$ and $(p_\gamma - p_{\gamma_*}) \log s = o(n)$. The proof is given as follows.

Proof of Gaussian Prior. For $\gamma \in \mathcal{M}_1$, by Assumption 4.7, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{1}{2s^2}\beta_i^2\right)}{\prod_{j=1}^{p_\gamma} \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{1}{2s^2}\beta_j^2\right)} \\
&= \frac{r_\gamma}{2}\log(2\pi s^2) + \frac{1}{2s^2}\left(\sum_{j=1}^{p_\gamma}\hat{\beta}_j^2 - \sum_{i=1}^{p_{\gamma_*}}\hat{\beta}_i^2\right) \\
&= r_\gamma \log s + \frac{r_\gamma}{2}\log(2\pi) + \frac{O_p(n^{2c})}{2s^2}.
\end{aligned}$$

where $r_\gamma = p_\gamma - p_{\gamma_*}$. Hence, to satisfy Condition 1 for any $\gamma \in \mathcal{M}_1$, we need to choose $s > \frac{p_n}{\sqrt{n}}$.

Similarly, for $\gamma \in \mathcal{M}_2$, by Assumption 4.7, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{1}{2s^2}\beta_i^2\right)}{\prod_{j=1}^{p_\gamma} \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{1}{2s^2}\beta_j^2\right)} \\
&= \frac{r_\gamma}{2}\log(2\pi s^2) + \frac{1}{2s^2}\left(\sum_{j=1}^{p_\gamma}\hat{\beta}_j^2 - \sum_{i=1}^{p_{\gamma_*}}\hat{\beta}_i^2\right) \\
&= r_\gamma \log s + \frac{r_\gamma}{2}\log(2\pi) + \frac{O_p(n^{2c})}{2s^2},
\end{aligned}$$

where $r_\gamma = p_\gamma - p_{\gamma_*}$. To satisfy Condition 2 that $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o_p(n)$ for any $\gamma \in \mathcal{M}_2$, we must choose $r_\gamma \log s = o(n)$. There are many ways to set up s to satisfy both conditions. For example, we can set $s = p_n^{1+\delta}$ for an arbitrary constant $\delta \geq 0$. Then, by Assumptions 4.1, 4.5 and 4.6, we have $s > p_n/\sqrt{n}$ and $(p_\gamma - p_{\gamma_*}) \log s = o(n)$. Thus, according to Theorem 4.1, the model selection consistency holds for the independent Gaussian prior. □

4.4.2 Laplace prior

Ever since Park and Casella (2008) introduced Bayesian lasso, the Laplace prior has become a popular choice in Bayesian regression analysis (Casella et al., 2010; Hans, 2009, 2010).

Suppose we assign the Laplace prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma, s) = \prod_{j=1}^{p_\gamma} \frac{1}{2s} \exp\left(-\frac{|\beta_j|}{s}\right).$$

Under Assumptions 4.1, 4.5 - 4.7, the Laplace prior yields consistent model selection if $s > p_n/\sqrt{n}$ and $(p_\gamma - p_{\gamma_*}) \log s = o(n)$. The proof is as follows:

Proof. First, by Assumption 4.7, for $\gamma \in \mathcal{M}_n$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{1}{2s} \exp\left(-\frac{|\hat{\beta}_i|}{s}\right)}{\prod_{j=1}^{p_\gamma} \frac{1}{2s} \exp\left(-\frac{|\hat{\beta}_j|}{s}\right)} \\
&= r_\gamma \log 2s + \frac{1}{s}\left\{\sum_{j=1}^{p_\gamma} |\hat{\beta}_j| - \sum_{i=1}^{p_{\gamma_*}} |\hat{\beta}_i|\right\} \\
&= r_\gamma \log 2s + \frac{O_p(n^c)}{s}
\end{aligned}$$

where $r_\gamma = p_\gamma - p_{\gamma_*}$.

When $s > p_n/\sqrt{n}$, for $\gamma \in \mathcal{M}_1$, as $n \to \infty$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= r_\gamma \log 2 + r_\gamma \log s + \frac{O_p(n^c)}{s} \\
&> r_\gamma \log s\{1 + o_p(1)\} \\
&> r_\gamma \log \frac{p_n}{\sqrt{n}}.
\end{aligned}$$

When $r_\gamma \log s = o(n)$, for any $\gamma \in \mathcal{M}_2$, as $n \to \infty$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= r_\gamma \log 2 + r_\gamma \log s + \frac{O_p(n^c)}{s} \\
&= o(n).
\end{aligned}$$

There are many ways to set up s to satisfy both conditions. For example, we can set $s = p_n^{1+\delta}$ for some $\delta \geq 0$. Then, by Assumptions 4.1, 4.5 and 4.6, we have $s > p_n/\sqrt{n}$ and $(p_\gamma - p_{\gamma_*}) \log s = o(n)$. Thus, according to Theorem 4.1, the model selection consistency

holds for the Laplace prior. □

4.4.3 Scaled Student's t prior

As one of the shrinkage priors, the scaled Student's t prior (West, 1987) concentrates at zero with relatively thicker tails, such properties make the scaled Student's t prior an appealing prior choice in high-dimensional model selection (Armagan et al., 2011, 2013; Tipping, 2001). Suppose we assign scaled Student's t prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma, s, d) = \prod_{j=1}^{p_\gamma}[sd^{1/2}B(d/2, 1/2)]^{-1}(1 + \frac{\beta_j^2}{sd})^{-(d+1)/2}$$

where s is the scale and d is the degrees of freedom. Under Assumptions 4.1, 4.5 - 4.7, the scaled Student's t prior yields consistent model selection if $s > p_n/\sqrt{n}$, $(p_\gamma - p_{\gamma_*}) \log s = o(n)$ and $(p_\gamma - p_{\gamma_*}) \log d = o(n)$. The proof is as follows:

Proof. For $\gamma \in \mathcal{M}$, if $\log d = o(n)$, by Assumption 4.7, we then have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s, d)}{p(\hat{\beta}_\gamma|\gamma, s, d)} &= \frac{\prod_{i=1}^{p_{\gamma_*}}[sd^{1/2}B(d/2, 1/2)]^{-1}(1 + \frac{\hat{\beta}_i^2}{sd})^{-(d+1)/2}}{\prod_{j=1}^{p_\gamma}[sd^{1/2}B(d/2, 1/2)]^{-1}(1 + \frac{\hat{\beta}_j^2}{sd})^{-(d+1)/2}} \\
&= r_\gamma \log s + \frac{r_\gamma}{2} \log d + r_\gamma \log B(d/2, 1/2) \\
&\quad + \frac{d+1}{2}\left[\sum_{j=1}^{p_\gamma} \log(1 + \frac{\hat{\beta}_j^2}{sd}) - \sum_{i=1}^{p_{\gamma_*}} \log(1 + \frac{\hat{\beta}_i^2}{sd})\right] \\
&= r_\gamma \log s + \frac{r_\gamma}{2} \log d + r_\gamma \log B(d/2, 1/2) + \left(\sum_{j=1}^{p_\gamma} \frac{\hat{\beta}_j^2}{2s} - \sum_{i=1}^{p_{\gamma_*}} \frac{\hat{\beta}_i^2}{2s}\right) \\
&= r_\gamma \log s + \frac{r_\gamma}{2} \log d + r_\gamma \log O(1) + \frac{O_p(n^c)}{2s}
\end{aligned}$$

where $r_\gamma = p_\gamma - p_{\gamma_*}$.

When $s > p_n/\sqrt{n}$, for $\gamma \in \mathcal{M}_1$, as $n \to \infty$, we have

$$\begin{aligned}\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s, d)}{p(\hat{\beta}_{\gamma}|\gamma, s, d)} &= r_\gamma \log s + \frac{r_\gamma \log d}{2} + r_\gamma \log O_p(1) + \frac{O(n^c)}{2s} \\ &> r_\gamma \log s\{1 + o(1)\} \\ &> r_\gamma \log \frac{p_n}{\sqrt{n}}.\end{aligned}$$

When $r_\gamma \log s = o(n)$ and $r_\gamma \log d = o(n)$, for $\gamma \in \mathcal{M}_2$, as $n \to \infty$, we have

$$\begin{aligned}\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s, d)}{p(\hat{\beta}_{\gamma}|\gamma, s, d)} &= r_\gamma \log s + \frac{r_\gamma}{2} \log d + r_\gamma \log O(1) + \frac{O_p(n^c)}{2s} \\ &= o(n).\end{aligned}$$

There are many ways to set up s to satisfy both conditions. For example, we set $s = p_n^{1+\delta}$ for some $\delta \geq 0$. Then, by Assumption 4.1, 4.5 and 4.6, we have $s > p_n/\sqrt{n}$ and $(p_\gamma - p_{\gamma_*}) \log s = o(n)$. Thus, according to Theorem 4.1, the model selection consistency holds for the Scaled Student's t prior. □

4.4.4 Generalized double Pareto prior

The generalized double Pareto prior proposed by Armagan et al. (2010) has some appealing properties such as spike at zero and Student's t-like tails which make it a popular choice of Bayesian prior (Armagan et al., 2013; Pal et al., 2017). Suppose we assign the generalized double Pareto prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma, \alpha, \eta) = \prod_{j=1}^{p_\gamma} \frac{\alpha}{2\eta} \left(1 + \frac{|\beta_j|}{\eta}\right)^{-(\alpha+1)}.$$

Under Assumptions 4.1, 4.5 - 4.7, the generalized double Pareto prior yields consistent model selection if $\frac{\eta}{\alpha} > p_n/\sqrt{n}$ and $(p_\gamma - p_{\gamma_*}) \log \frac{\eta}{\alpha} = o(n)$. The proof is as follows:

Proof. If α and η are chosen to grow with n, by Assumption 4.7, for $\gamma \in \mathcal{M}_n$, as $n \to \infty$,

we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\alpha,\eta)}{p(\hat{\beta}_{\gamma}|\gamma,\alpha,\eta)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{\alpha}{2\eta}\left(1+\frac{|\hat{\beta}_i|}{\eta}\right)^{-(\alpha+1)}}{\prod_{j=1}^{p_{\gamma}} \frac{\alpha}{2\eta}\left(1+\frac{|\hat{\beta}_j|}{\eta}\right)^{-(\alpha+1)}} \\
&= r_\gamma \log\frac{\alpha}{2\eta} + (\alpha+1)\left[\sum_{j=1}^{p_\gamma}\log(1+\frac{|\hat{\beta}_j|}{\eta}) - \sum_{i=1}^{p_{\gamma_*}}\log(1+\frac{|\hat{\beta}_i|}{\eta})\right] \\
&= r_\gamma \log\frac{2\eta}{\alpha} + \left[\sum_{j=1}^{p_\gamma}\frac{\alpha}{\eta}|\hat{\beta}_j| - \sum_{i=1}^{p_{\gamma_*}}\frac{\alpha}{\eta}|\hat{\beta}_i|\right] \\
&= r_\gamma \log 2 + r_\gamma \log\frac{\eta}{\alpha} + \frac{\alpha}{\eta}\left[\sum_{j=1}^{p_\gamma}|\hat{\beta}_j| - \sum_{i=1}^{p_{\gamma_*}}|\hat{\beta}_i|\right] \\
&= r_\gamma \log 2 + r_\gamma \log\frac{\eta}{\alpha} + \frac{O_p(n^c)}{\eta/\alpha}
\end{aligned}$$

where $r_\gamma = p_\gamma - p_{\gamma_*}$.

When $\frac{\eta}{\alpha} > p_n/\sqrt{n}$, for $\gamma \in \mathcal{M}_1$, as $n \to \infty$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\alpha,\eta)}{p(\hat{\beta}_{\gamma}|\gamma,\alpha,\eta)} &= r_\gamma \log 2 + r_\gamma \log\frac{\eta}{\alpha} + r_\gamma\frac{O_p(n^c)}{\eta/\alpha} \\
&> r_\gamma \log\frac{\eta}{\alpha}\{1+o_p(1)\} \\
&> r_\gamma \log\frac{p_n}{\sqrt{n}}.
\end{aligned}$$

When $r_\gamma \log\frac{\eta}{\alpha} = o(n)$, for $\gamma \in \mathcal{M}_2$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\alpha,\eta)}{p(\hat{\beta}_{\gamma}|\gamma,\alpha,\eta)} &= r_\gamma \log 2 + r_\gamma \log\frac{\eta}{\alpha} + \frac{O_p(n^c)}{\eta/\alpha} \\
&= o_p(n).
\end{aligned}$$

There are many ways to set up η/α to satisfy both conditions. For example, we can set $\eta/\alpha = p_n^{1+\delta}$ for some $\delta \geq 0$. Then, by Assumption 4.1, 4.5 and 4.6, we have $\eta/\alpha > p_n/\sqrt{n}$ and $(p_\gamma - p_{\gamma_*})\log\frac{\eta}{\alpha} = o(n)$. Thus, according to Theorem 4.1, the model selection consistency holds for the generalized double Pareto prior.

□

4.4.5 Horseshoe prior

The Horseshoe prior (Carvalho et al., 2010) has some fairly desirable properties, such as heavy tail, an infinite spike at zero and so on. Due to such properties, it becomes a popular prior choice in Bayesian regression (Armagan et al., 2013; Carvalho et al., 2010). Suppose we assign the Horseshoe prior to β_γ in the following way:

$$p(\beta_\gamma|\gamma,\tau) = \prod_{i=1}^{p_\gamma} K_0(\tau^2)^{-1/2} \exp(\frac{\beta_i^2}{2\tau^2}) E_1(\frac{\beta_i^2}{2\tau^2}),$$

where $K_0 = 1/(2\pi^3)^{1/2}$ and $E_1(\cdot)$ is the exponential integral function. Under Assumptions 4.1, 4.5 and 4.6, the Horseshoe prior yields consistent model selection if $\tau^2 > p_n^2/n$ and $(p_\gamma - p_{\gamma_*}) \log \tau^2 = o(n)$. The proof is as follows:

Proof. First, by Lemma C.1 (see Appendix C.3), for any $\gamma \in \mathcal{M}_n$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\tau)}{p(\hat{\beta}_\gamma|\gamma,\tau)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_i^2}{2\tau^2}) E_1(\frac{\hat{\beta}_i^2}{2\tau^2})}{\prod_{j=1}^{p_\gamma} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_j^2}{2\tau^2}) E_1(\frac{\hat{\beta}_j^2}{2\tau^2})} \\
&> \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{K_0}{2} (\tau^2)^{-1/2} \log\left(1 + \frac{4\tau^2}{\hat{\beta}_i^2}\right)}{\prod_{j=1}^{p_\gamma} K_0(\tau^2)^{-1/2} \log\left(1 + \frac{2\tau^2}{\hat{\beta}_j^2}\right)} \\
&= r_\gamma \log \frac{2}{K_0} - p_\gamma \log 2 + \frac{r_\gamma}{2} \log \tau^2 \\
&\quad + \sum_{i=1}^{p_{\gamma_*}} \log\left[\log\left(1 + \frac{4\tau^2}{\hat{\beta}_i^2}\right)\right] - \sum_{j=1}^{p_\gamma} \log\left[\log\left(1 + \frac{2\tau^2}{\hat{\beta}_j^2}\right)\right]
\end{aligned}$$

where $r_\gamma = p_\gamma - p_{\gamma_*}$.

We also have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\tau)}{p(\hat{\beta}_{\gamma}|\gamma,\tau)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_i^2}{2\tau^2}) E_1(\frac{\hat{\beta}_i^2}{2\tau^2})}{\prod_{j=1}^{p_{\gamma}} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_j^2}{2\tau^2}) E_1(\frac{\hat{\beta}_j^2}{2\tau^2})} \\
&< \log \frac{\prod_{i=1}^{p_{\gamma_*}} K_0(\tau^2)^{-1/2} \log\left(1+\frac{2\tau^2}{\hat{\beta}_i^2}\right)}{\prod_{j=1}^{p_{\gamma}} \frac{K_0}{2}(\tau^2)^{-1/2} \log\left(1+\frac{4\tau^2}{\hat{\beta}_j^2}\right)} \\
&= r_\gamma \log\frac{2}{K_0} + p_{\gamma_*}\log 2 + \frac{r_\gamma}{2}\log\tau^2 \\
&+ \sum_{i=1}^{p_{\gamma_*}} \log\left[\log\left(1+\frac{2\tau^2}{\hat{\beta}_i^2}\right)\right] - \sum_{j=1}^{p_\gamma} \log\left[\log\left(1+\frac{4\tau^2}{\hat{\beta}_j^2}\right)\right]
\end{aligned}$$

where $r_\gamma = p_\gamma - p_{\gamma_*}$.

For $\gamma \in \mathcal{M}_1$, when $\tau^2 \geq p_n^2/n$, as $n \to \infty$, we have

$$\begin{aligned}
&\frac{1}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\tau)}{p(\hat{\beta}_{\gamma}|\gamma,\tau)} \\
&> r_\gamma \log\frac{2}{K_0} - p_\gamma \log 2 + \frac{1}{2}\log\tau^2 \\
&+ \left\{ \sum_{i=1}^{p_{\gamma_*}} \log\left[\log\left(1+\frac{4\tau^2}{\hat{\beta}_i^2}\right)\right] - \sum_{j=1}^{p_\gamma} \log\left[\log\left(1+\frac{2\tau^2}{\hat{\beta}_j^2}\right)\right] \right\} \\
= \;& r_\gamma \log\frac{2}{K_0} - p_\gamma \log 2 + \frac{r_\gamma}{2}\log\tau^2 + \left\{ p_{\gamma_*} \log\left[o(\tau^2)\right] - p_\gamma \log\left[o(\tau^2)\right] \right\} \\
= \;& r_\gamma \log\frac{2}{K_0} - p_\gamma \log 2 + \frac{r_\gamma}{2}\log\tau^2 - r_\gamma \log\left[o(\tau^2)\right] \\
= \;& \frac{r_\gamma}{2}\log\tau^2\{1+o(n)\} \\
\geq \;& r_\gamma \log\frac{p_n}{\sqrt{n}},
\end{aligned}$$

For $\gamma \in \mathcal{M}_2$, when τ^2 grows with n and $r_\gamma \log\tau^2 = o(n)$, as $n \to \infty$ we have

$$
\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\tau)}{p(\hat{\beta}_{\gamma}|\gamma,\tau)} &> r_\gamma \log \frac{2}{K_0} - p_\gamma \log 2 + \frac{r_\gamma}{2}\log \tau^2 \\
&\quad + \left\{ \sum_{i=1}^{p_{\gamma_*}} \log \left[\log \left(1 + \frac{4\tau^2}{\hat{\beta}_i^2} \right) \right] - \sum_{j=1}^{p_\gamma} \log \left[\log \left(1 + \frac{2\tau^2}{\hat{\beta}_j^2} \right) \right] \right\} \\
&= r_\gamma \log \frac{2}{K_0} - p_\gamma \log 2 + \frac{r_\gamma}{2}\log \tau^2 + [p_{\gamma_*} \log o(\tau^2) - p_\gamma \log o(\tau^2)] \\
&= r_\gamma \log \frac{2}{K_0} - p_\gamma \log 2 + \frac{r_\gamma}{2}\log \tau^2 - r_\gamma \log o(\tau^2) \\
&= \frac{r_\gamma}{2}\log \tau^2 \{1 + o(n)\} \\
&= o(n).
\end{aligned}
$$

And, we also have

$$
\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\tau)}{p(\hat{\beta}_{\gamma}|\gamma,\tau)} &< r_\gamma \log \frac{2}{K_0} + p_{\gamma_*} \log 2 + \frac{r_\gamma}{2}\log \tau^2 \\
&\quad + \left\{ \sum_{i=1}^{p_{\gamma_*}} \log \left[\log \left(1 + \frac{2\tau^2}{\hat{\beta}_i^2} \right) \right] - \sum_{j=1}^{p_\gamma} \log \left[\log \left(1 + \frac{4\tau^2}{\hat{\beta}_j^2} \right) \right] \right\} \\
&= r_\gamma \log \frac{2}{K_0} + p_{\gamma_*} \log 2 + \frac{r_\gamma}{2}\log \tau^2 + [p_{\gamma_*} \log o(\tau^2) - p_\gamma \log o(\tau^2)] \\
&= r_\gamma \log \frac{2}{K_0} + p_{\gamma_*} \log 2 + \frac{r_\gamma}{2}\log \tau^2 - r_\gamma \log o(\tau^2) \\
&= \frac{r_\gamma}{2}\log \tau^2 \{1 + o(n)\} \\
&= o(n),
\end{aligned}
$$

thus, we have $\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*,\tau)}{p(\hat{\beta}_{\gamma}|\gamma,\tau)} = o(n)$ as $n \to \infty$. There are many ways to set up τ^2. For example, we set $\tau^2 = p_n^{2+\delta}$ for some $\delta \geq 0$. Then, by Assumptions 4.1, 4.5 and 4.6, we have $\tau^2 > p_n^2/n$ and $(p_\gamma - p_{\gamma_*}) \log \tau^2 = o(n)$. Thus, according to Theorem 4.1, the model selection consistency holds for the Horseshoe prior. □

4.5 Simulation study

In this section, we conduct a simulation study to examine the performance of Theorem 4.1. Since we already showed the model selection consistency with Gaussian data in Chapter 2 and Chapter 3, in this chapter we switch our focus to non-Gaussian data.

4.5.1 Simulation setting

We generate our data from the model $y_i \sim Ber(p_i)$, where

$$\begin{aligned} p_i &= \phi\left(\sum_{j=1}^{p_n} \beta_j x_{ij}\right), \\ (x_{i1}, \ldots, x_{ij})^\top &\overset{iid}{\sim} \mathcal{N}_{p_n}(0, \Sigma), \\ \beta_{\gamma_*} &= (1, 1, -1, -1)^\top, \end{aligned}$$

where $\Sigma = I_{p_n}$, γ_* is the true model which is a randomly generated index set with $p_{\gamma_*} = 4$, and $\phi(\cdot)$ is the cdf of the standard normal distribution. Following Assumption 4.5, we set $K_n = n^{2/3}$. To exam the limiting behavior of the marginal likelihood of true model as sample size increases, we run the simulation with $n = \{100, 200, 300\}$.

For the choice of priors, we assign all the five priors in Section 4.4 to β_γ respectively. Table 4.1 shows our choices of the hyperparameters for each prior. According to the sufficient conditions we derived for the prior, Case 1 and Case 3 violate the sufficient conditions, whereas Case 2 satisfies the conditions.

Table 4.1: *Hyperparameter settings*

Prior	**Case 1**	**Case 2**	**Case 3**
Gaussian	$s = p^{0.2}$	$s = p^{1.1}$	$s = \exp(0.1n)$
Laplace	$s = p^{0.2}$	$s = p^{1.1}$	$s = \exp(0.1n)$
Scaled	$s = p^{0.2}$	$s = p^{1.1}$	$s = \exp(0.1n)$
student's t	$d = n$	$d = n$	$d = n$
Pareto	$\frac{\eta}{\alpha} = p^{0.2}$, $\alpha = n^{0.3}$	$\frac{\eta}{\alpha} = p^{1.1}$, $\alpha = n$	$\frac{\eta}{\alpha} = \exp(0.1n)$, $\alpha = n$
Horseshoe	$\tau = p^{0.2}$	$\tau = p^{1.1}$	$\tau = \exp(0.1n)$

4.5.2 Shotgun stochastic search

Since the number of candidate models is large, it is computationally challenging to compute the marginal likelihood for all the candidate models. To address this challenge, we use the idea of SSS (Shotgun Stochastic Search) which explores the model space using MCMC (Markov Chain Monte Carlo) computation.

Let $nbd(\gamma) = \{\gamma^+, \gamma, \gamma^-\}$ be a neighborhood of model γ, where γ^+ is the set of models obtained by adding one predictor variable to model γ, γ^- is the set of models obtained by deleting one predictor variable from γ. The SSS updates the model by searching for the best candidate model in the $nbd(\gamma)$. Let $\gamma^{(t)}$ be the current model, $\gamma^{(t)}_{\max}$ be current model with the greatest marginal likelihood, γ_0 be the initial model and $\gamma_{\max_0}$ be the initial model with the greatest marginal likelihood. The SSS can be implemented as in Algorithm 3.

Algorithm 3 Shotgun Stochastic Search.

Set $\gamma^{(1)} = \gamma_0$, and $\gamma_{\max}^{(1)} = \gamma_{\max_0}$

Repeat for $t = 1, \ldots, T$

Set $s = \gamma^{(t)}$, $s^* = \gamma_{\max}^{(t)}$

Repeat for $j = 1, \ldots, p_n$:

If $s_j = 0$, set $s_j = s \cup \{j\}$

If $s_j = 1$, set $s_j = s \setminus \{j\}$

Compute $p(y|s_j) = p(y \mid \hat{\beta}_{s_j}) p(\hat{\beta}_{s_j}|s_j)(n)^{-p_{s_j}/2}$

Compute $\tilde{s} = \max_{1 \le j \le p_n} p(y|s_j) = p(y \mid \hat{\beta}_{s_j}) p(\hat{\beta}_{s_j}|s_j)(n)^{-p_{s_j}/2}$

If $p(y|\tilde{s}) > p(y|s^*)$, then $s^* \leftarrow \tilde{s}$;

else $s^* \leftarrow s^*$

Repeat for $j = 1, \ldots, p_n$:

Compute $w_j = p(y|s_j)\mathbb{I}(s_j \in \mathcal{M}_n) / \sum_{j=1}^{p_n+1} \{p(y \mid s_j)\mathbb{I}(s_j \in \mathcal{M}_n)\}$

Sample $z \sim \text{Categorical}(w_1, w_2, \ldots, w_{p_n})$, then $s \leftarrow s_z$

Update $\gamma^{(t+1)} = s$, $\gamma_{\max}^{(t+1)} = s^*$

Return $\gamma_{\max}$

We generate 100 Monte Carlo experiments, and run the simulation $T = 500$ times. To evaluate the performance of our marginal likelihood approach, we compute the relative frequency of the true model, which is

$$R.F.(\gamma_*) = N^{-1} \sum_{i=1}^{N} \mathbb{I}(\gamma^{(i)} = \gamma_*),$$

where N is the number of Monte Carlo experiments and $\gamma^{(i)}$ is the selected model in ith experiment. If Theorem 4.1 holds, we expect to observe the $R.F.(\gamma_*) \to 1$ only under the settings of Case 2 for each prior as sample size increases.

4.5.3 Simulation results

Table 4.2 reports the relative frequency of the true model out of the 100 Monte Carlo experiments.

We only observe a strong trend of the $R.F.(\gamma_*)$ increasing to one in Case 2 under each prior. In Case 1 and Case 3 which violate the sufficient conditions, the relative frequency of true model fails to grow with the sample size. Most of the results are similar across all the priors except for Case 1 under the Horseshoe prior. Compare to other priors, Horseshoe prior shows better model selection results when the hyperparameter settings violate the first sufficient condition. Nevertheless, we still do not observe a strong trend of $R.F.(\gamma_*)$ growing along with the sample size.

Figure 4.1 - Figure 4.5 are the trace plots of the $R.F.(\gamma_*)$ as sample size increases under each prior. The plots demonstrate that when the sufficient conditions are met, the trace of $R.F.(\gamma_*)$ shows a strong trend of increasing to one as the sample size increases. When the sufficient conditions are violated, we do not observe the trend of $R.F.(\gamma_*)$ increasing to one as the sample size increases.

In general, the simulation results match our expectation. When the prior meets the sufficient condition, the the true model becomes the model which maximizes the marginal likelihood as the sample size increases. Since the model with the maximum marginal likelihood is equivalent to the model with the greatest posterior model probability under our model set-up, therefore the marginal likelihood approach achieves model selection consistency in the sense that

$$\text{pr}(\gamma_*|y) > \max_{\gamma \in \mathcal{M}_n \setminus \{\gamma_*\}} \text{pr}(\gamma|y)$$

as $n \to \infty$.

Table 4.2: *Simulation results based on 100 Monte Carlo experiments.*

Prior	**Case**	**n=100**	**n=300**	**n=500**
	1	0.40	0.40	0.37
Gaussian	2	0.33	0.99	1.00
	3	0.00	0.02	0.01
	1	0.43	0.32	0.32
Laplace	2	0.36	1.00	1.00
	3	0.00	0.02	0.00
Scaled	1	0.52	0.48	0.43
Student's	2	0.33	0.99	1.00
t	3	0.00	0.02	0.01
	1	0.48	0.40	0.33
Pareto	2	0.36	1.00	1.00
	3	0.00	0.02	0.01
	1	0.66	0.87	0.69
Horseshoe	2	0.22	1.00	1.00
	3	0.00	0.01	0.00

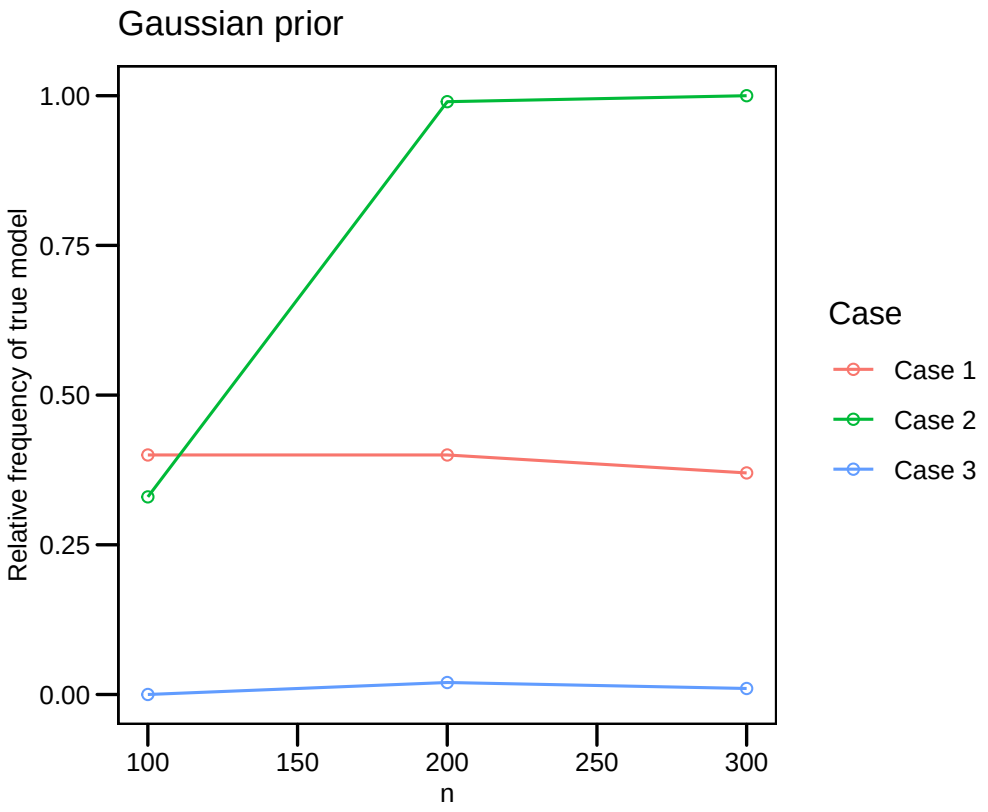

Figure 4.1: *The trace plots of the relative frequency of γ_* under the Gaussian prior as the sample size n increases.*

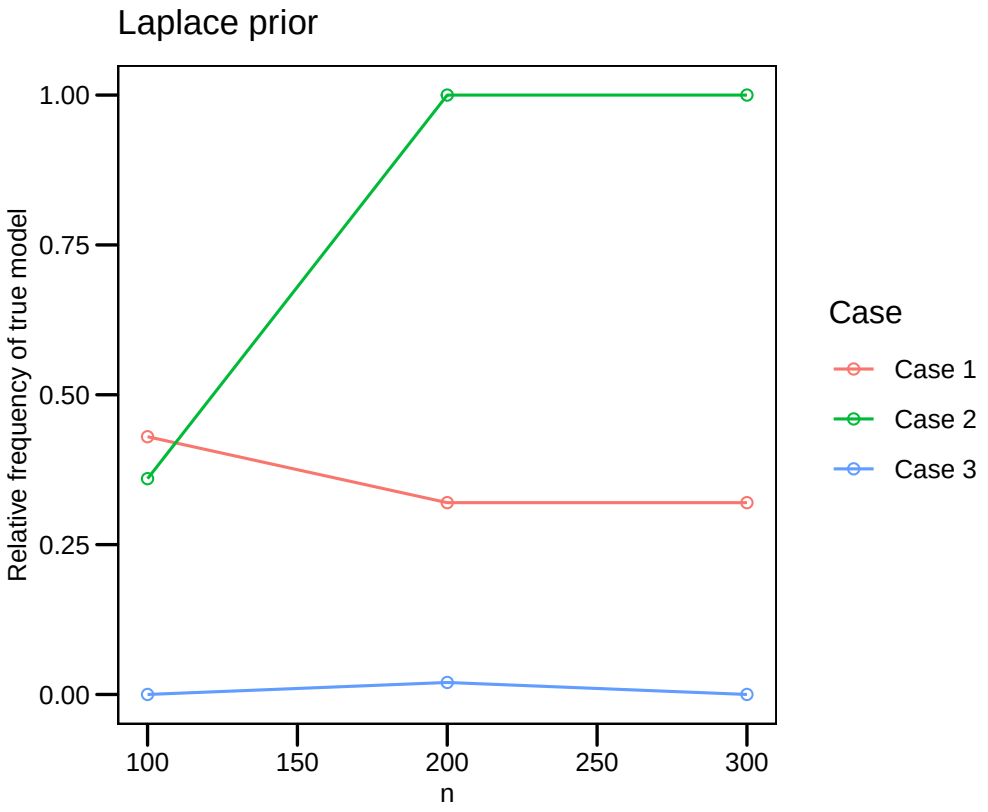

Figure 4.2: *The trace plots of the relative frequency of γ_* under the Laplace prior as the sample size n increases.*

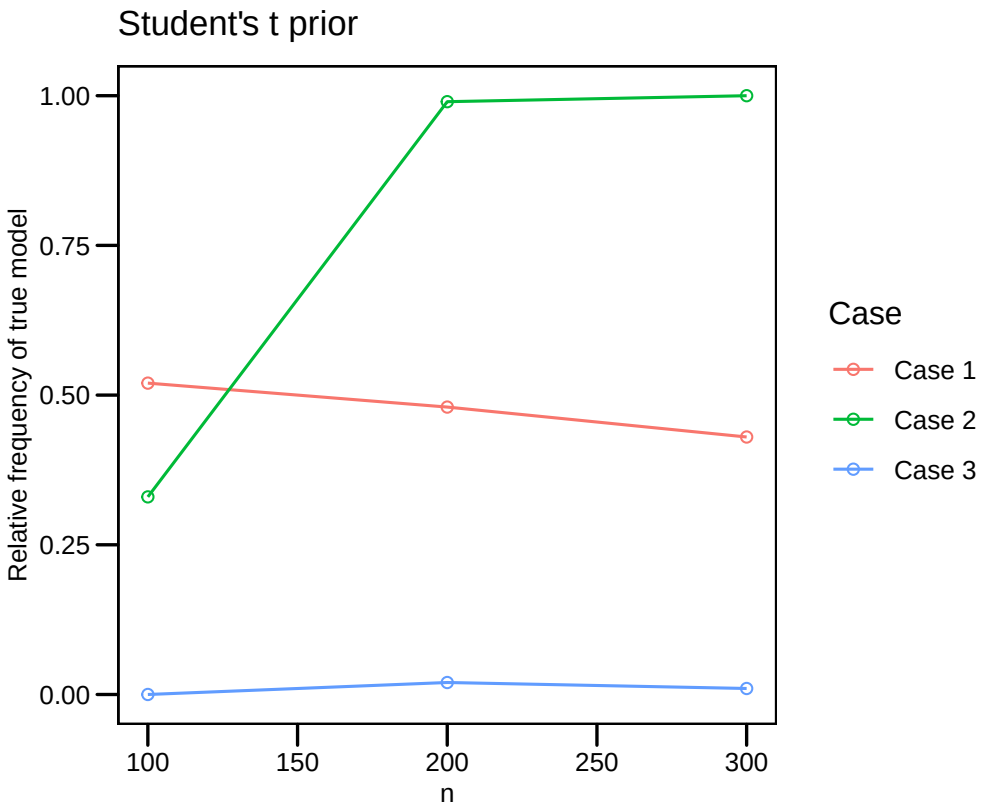

Figure 4.3: *The trace plots of the relative frequency of γ_* under the scaled Student's t prior as the sample size n increases.*

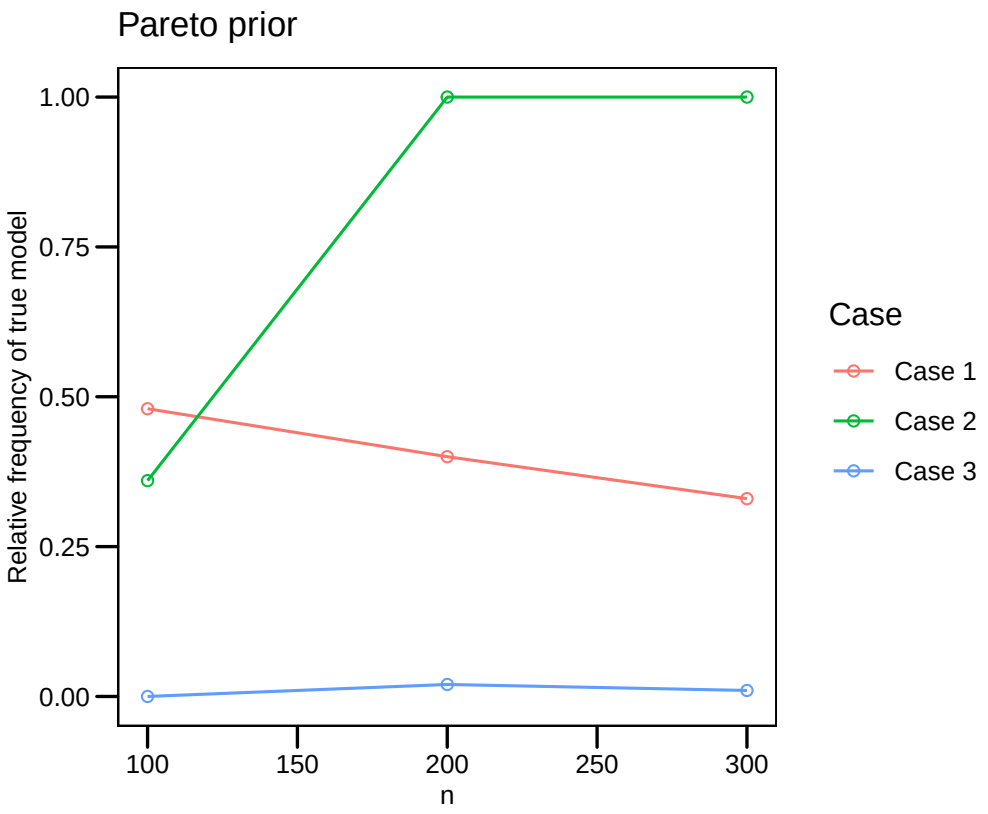

Figure 4.4: *The trace plots of the relative frequency of γ_* under the generalized double Pareto prior as the sample size n increases.*

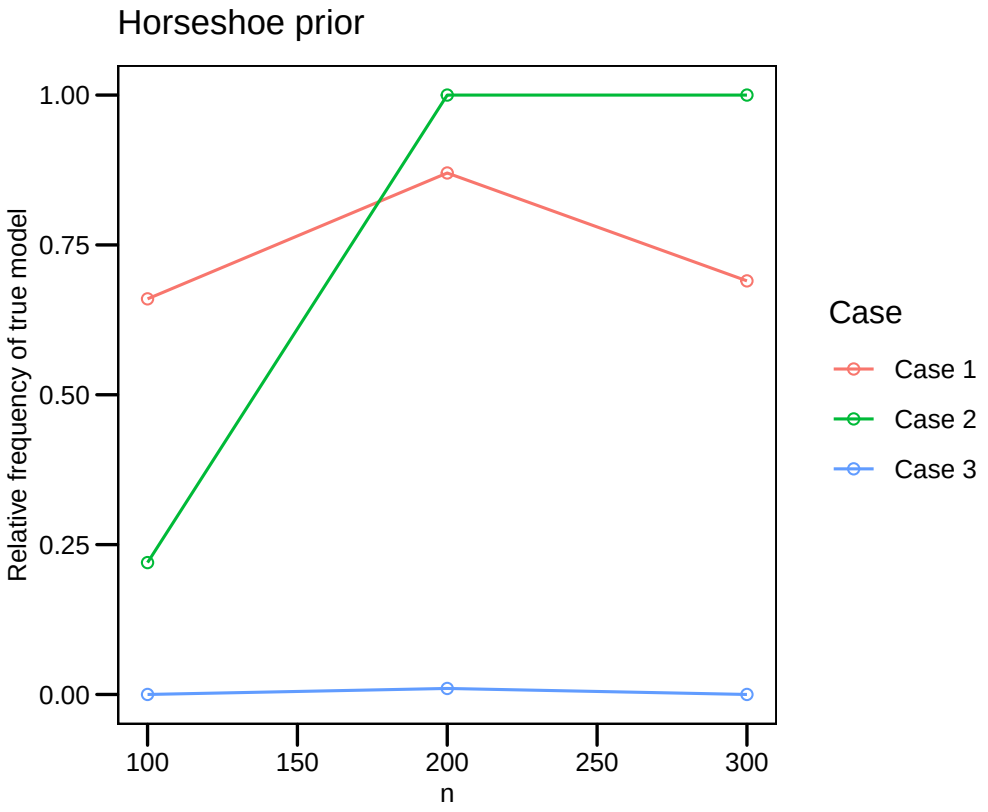

Figure 4.5: *The trace plots of the relative frequency of γ_* under the Horseshoe prior as the sample size n increases.*

4.6 Discussion

Theorem 4.1 proposes general sufficient conditions for $p(\beta_\gamma|y)$ under which the true model maximizes the marginal likelihood in the Bayesian generalized linear regression. Theorem 4.1 holds for both Gaussian and non-Gaussian data. The general sufficient conditions also provide useful guidelines for hyperparameter specifications under specific priors. Model selection consistency can be achieved by applying Theorem 4.1 to the specification of the prior for β_γ in high-dimensional and low-dimensional model selection.

In our analysis of model selection consistency, we assume some common properties the distribution of the likelihood should possess. Even though these properties are commonly shared by many likelihood distributions, such as normal distribution and binomial distribu-

tion, they do not cover all the likelihood distributions. An interesting direction for future work is to expand the general properties to include more likelihood distribution, so that we can derive more general sufficient conditions which will hold under a broader range of model selection.

Chapter 5

Summary and discussion

In this dissertation, we focus on obtaining posterior model probability consistency in the high-dimensional model selection. We approach the model selection from a Bayesian perspective. Our investigation mainly centers at deriving sufficient conditions for the $p(\beta_\gamma|y)$ to yield posterior model probability consistency. A distinctive feature in our setting is that we allow the size of the full model to increase with the sample size, even at a faster rate in some cases.

In Chapter 2, we begin our investigation of the asymptotic behavior of the posterior model probability with a specific prior, namely the Zellner's g-prior. It is well known that the value of hyperparameter g greatly affects posterior model probability. Moreno et al. (2015) point out:

> Large g values induce the Lindley-Bartlett paradox, and a fixed value for g induces inconsistency, which can be corrected if g were dependent on n.

The sufficient conditions we obtain require the g to grow with n. However, the growth rate of g can not be too fast in order to avoid the model selection favors the null model, namely the "Lindley-Bartlett" paradox. The sufficient conditions indicate that the value of g needs to be large but not too large. A large g value induces a flat prior, i.e. noninformative prior which is a desired property of the prior. However, a g value which is too large leads to the exceedingly noninformative prior. Therefore, in order to achieve model selection consistency, we need to control the informativeness of the prior. The sufficient conditions provide simple

and intuitive guidelines for the specification of the hyperparameter g in the Zellner's g-prior which restricts the informativeness of the prior within a desired range.

In Chapter 3, we explore the model selection consistency under the general prior. Under the general prior settings, we do not impose any specific distribution on the prior of β_γ. The general assumption for the prior allows us to derive general sufficient conditions for the prior to achieve model selection consistency. From Chapter 2, we know that the g value needs to be carefully chosen to control the informativeness of the prior. The sufficient conditions we obtain in Chapter 3 confine the informativeness within a certain range so that the prior of β_γ will achieve posterior model probability consistency. The results in Chapter 2 and Chapter 3 are consistent albeit the results in Chapter 2 are more specific which can be considered as a special case of the results in Chapter 3. The general sufficient conditions imply if we control the informativeness of the prior by carefully choosing the hyperparameters we will achieve posterior model probability consistency. Our study of a series of shrinkage priors demonstrate that the satisfaction of the sufficient conditions lead to model selection consistency.

Following the study of the general prior, we extend our framework to model selection in general linear regression in Chapter 4. Under the general framework, we relax our assumption of the model likelihood. We do not impose any specific distribution on the data. The sufficient conditions we derive under the general settings are consistent with the results from Chapter 2 and Chapter 3, albeit the results in Chapter 4 are more general which contain Chapter 2 and Chapter 3 as special cases. Our simulation study demonstrate that the sufficient conditions are valid with non-Gaussian data.

There is one point which is worth of noting here, even though we consider the results of Chapter 2 and Chapter 3 special cases of the Chapter 4, the three projects are not nested within each other due the special settings of each project. Therefore, the projects from Chapter 2 and Chapter 3 can not simply be treated as special cases of Chapter 4.

The results we obtain from the series of investigations are interesting and exciting, however, there are still several points which we see potential for improvements. The first point is the way we handle σ^2 in Chapter 2. We assume σ^2 is known. We use Laplace approximation to compute the marginal likelihood, the value of σ^2 needs to be known to compute

the marginal likelihood. This assumption is normally not satisfied in the real data analysis, even though there are ways to estimate σ^2 consistently, such as the forward selection method which we adopt in the simulation study. A better approach to address the unknown σ^2 challenge is to assume σ^2 is unknown and develop the theory under such assumption. Second, our assumptions of the likelihood in Chapter 4. We do not assign any specific distributions to the likelihood in our model set up, in order to have some control when deriving the conditions, we assume the properties the likelihood possesses. Even though these properties are commonly assumed for the likelihood, it would be better if we can prove the validity of these properties first. The last, but not the least, is the real data study. We need to include more real data, especially non-Gaussian data, to test the performance of our theories.

In the future, there are several directions we believe are worthy of further investigations. First, we can extend our results to multivariate regression. In our setting, there is only one response variable. In reality, a large portion of the data analysis cases contain more than one response variable. Therefore, the multivariate regression is a natural extension of our framework on model selection consistency. Second, our investigation so far is confined in the linear regression regime. Real data often times exhibits nonlinear traits. Thus, the second possible direction for future study is to expand our study to the nonlinear regression. Thirdly, we can include more priors in our investigation of model selection consistency. There is not a prior that fits all the model selection cases. The shrinkage priors are popular in model selection study, however, people are also interested in other priors. There are priors which possess properties that are preferred in high-dimensional variable selection, such as the spike-and-slab prior. The expansion of our investigation to more priors will enhance the validity of our theorem in a broader area. The next interesting direction is to investigate the relationship between Bayesian approach and frequentist approach. For example, we can investigate whether there is an connection between the sufficient conditions we derived for the Laplace prior and the tuning parameter λ in Lasso regression. The last, we will include more real data study to exam the performance of our theorems. Ultimately, we hope to develop R packages based on our discoveries to facilitates the model selection in various real cases of high-dimensional regression analysis.

Bibliography

Felix Abramovich, Vadim Grinshtein, et al. Map model selection in gaussian regression. *Electronic Journal of Statistics*, 4:932–949, 2010.

A Armagan, D Dunson, and J Lee. Bayesian generalized double pareto shrinkage. *Biometrika*, 2010.

Artin Armagan, David B Dunson, and Merlise Clyde. Generalized beta mixtures of gaussians. *Advances in neural information processing systems*, 24:523, 2011.

Artin Armagan, David B Dunson, Jaeyong Lee, Waheed U Bajwa, and Nate Strawn. Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018, 2013.

Rina Foygel Barber, Mathias Drton, and Kean Ming Tan. Laplace approximation in high-dimensional bayesian regression. In *Statistical Analysis for High-Dimensional Data*, pages 15–36. Springer, 2016.

JM Bernardo, JO Berger, AP Dawid, AFM Smith, et al. Bayesian model averaging and model search strategies. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, volume 6, page 157. Oxford University Press, 1999.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

Howard D Bondell and Brian J Reich. Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107 (500):1610–1624, 2012.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

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

George Casella, Malay Ghosh, Jeff Gill, and Minjung Kyung. Penalized regression, standard errors, and bayesian lassos. *Bayesian analysis*, 5(2):369–411, 2010.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

M-H Chen, Joseph G Ibrahim, and Constantin Yiannoutsos. Prior elicitation, variable selection and bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):223–242, 1999.

Shibasish Dasgupta. High-dimensional posterior consistency of the bayesian lasso. *Communications in Statistics-Theory and Methods*, 45(22):6700–6708, 2016.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Carmen Fernandez, Eduardo Ley, and Mark FJ Steel. Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, 1994.

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

Edward I George and Robert E McCulloch. Stochastic search variable selection. *Markov chain Monte Carlo in practice*, 68:203–214, 1995.

Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

Edward I George, Robert E McCulloch, and R Tsay. Two approaches to bayesian model selection with applications. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, 309:339, 1996.

EdwardI George and Dean P Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

Subhashis Ghosal et al. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331, 1999.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

Chris Hans. Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, 20(2):221–229, 2010.

Mark H Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008.

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

Gareth M James, Peter Radchenko, and Jinchi Lv. Dasso: connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):127–142, 2009.

Harold Jeffreys. Theory of probability, clarendon, 1961.

Wenxin Jiang. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *The Annals of Statistics*, 35(4):1487–1511, 2007.

Robert E Kass and Larry Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934, 1995.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.

Robert Kohn, James Stephen Marron, and Paul Yau. Wavelet estimation using bayesian basis selection and basis averaging. *Statistica Sinica*, pages 109–128, 2000.

Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

Ryan Martin, Raymond Mess, Stephen G Walker, et al. Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847, 2017.

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

Elías Moreno, Javier Girón, George Casella, et al. Posterior model consistency in variable selection as the model dimension grows. *Statistical Science*, 30(2):228–241, 2015.

Naveen Naidu Narisetty, Xuming He, et al. Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics*, 42(2):789–817, 2014.

Ioannis Ntzoufras, Petros Dellaportas, and Jonathan J Forster. Bayesian variable and link determination for generalised linear models. *Journal of statistical planning and inference*, 111(1-2):165–180, 2003.

Subahdip Pal, Kshitij Khare, and James P Hobert. Trace class markov chains for bayesian inference with generalized double pareto shrinkage priors. *Scandinavian Journal of Statistics*, 44(2):307–323, 2017.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Adrian E Raftery. Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266, 1996.

Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

Todd E Scheetz, Kwang-Youn A Kim, Ruth E Swiderski, Alisdair R Philp, Terry A Braun, Kevin L Knudtson, Anne M Dorrance, Gerald F DiBona, Jian Huang, Thomas L Casavant, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

Zuofeng Shang and Murray K Clayton. Consistency of bayesian linear model selection with a growing number of parameters. *Journal of Statistical Planning and Inference*, 141(11): 3463–3474, 2011.

Michael Smith and Robert Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

Douglas K Sparks, Kshitij Khare, and Malay Ghosh. Necessary and sufficient conditions for high-dimensional posterior consistency under g-priors. *Bayesian Analysis*, 10(3):627–664, 2015.

Leland Stewart and William W Davis. Bayesian posterior distributions over sets of possible models with inferences computed by monte carlo integration. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 35(2):175–182, 1986.

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

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

Mingqiu Wang, Lixin Song, and Guo-liang Tian. Scad-penalized least absolute deviation regression in high-dimensional models. *Communications in Statistics-Theory and Methods*, 44(12):2452–2472, 2015.

Fengrong Wei and Jian Huang. Consistent group selection in high-dimensional linear regression. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(4):1369, 2010.

Mike West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

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

Arnold Zellner and Aloysius Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 31(1):585–603, 1980.

Cun-Hui Zhang, Jian Huang, et al. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Appendix A

Chapter 2 Preliminaries

A.1 Sparse Riesz condition

Proof. The sparse Riesz condition requires that $\frac{1}{n}X_\gamma^\top X_\gamma$ be uniformly positive definite for any model γ of size $2K$. For any model γ such that $p_\gamma < K$, we have:

$$\begin{aligned}
\|(I_n - H_\gamma)\mu_*\|^2 &= \|\mu_* - H_\gamma\mu_*\|^2 \\
&= \|X_{\gamma_*-\gamma}\beta_{\gamma_*-\gamma} - X_\gamma a\|^2 \\
&= \left[\{\beta_{\gamma_*-\gamma}^\top, -a\}\{X_{\gamma_*\cup\gamma}^\top X_{\gamma_*\cup\gamma}\}\{\beta_{\gamma_*\cup\gamma}^\top, -a\}^\top\right] \\
&= n\left[\{\beta_{\gamma_*-\gamma}^\top, -a\}\frac{1}{n}\{X_{\gamma_*\cup\gamma}^\top X_{\gamma_*\cup\gamma}\}\{\beta_{\gamma_*-\gamma}^\top, -a\}^\top\right] \\
&\geq n\left[\{\beta_{\gamma_*-\gamma}^\top, -a\}\lambda_{\gamma_*\cup\gamma}\{\beta_{\gamma_*-\gamma}^\top, -a\}^\top\right] \\
&\geq n\lambda_{\gamma_*-\gamma}\|\beta_{\gamma_*-\gamma}\|^2
\end{aligned}$$

which goes to infinity by the sparse Riezs condition, where $\lambda_{\gamma_*\cup\gamma}$ is the smallest eigenvalue of $\frac{1}{n}X_{\gamma_*\cup\gamma}^\top X_{\gamma_*\cup\gamma}$.

□

Before proving Lemma 2.1, Lemma 2.2 and Lemma 2.3, we also need the following 2 lemmas.

A.2 Lemma A1

Lemma A1. *Under Assumption 2.1, for a given γ such that $\gamma_* \not\subset \gamma$ and $p_\gamma \leq K + p_{\gamma_*}$, we have*

$$\frac{\mu_*^\top (I_n - H_\gamma)\epsilon}{\mu_*^\top (I_n - H_\gamma)\mu_*} = o_p(1),$$

as $n \to \infty$.

Proof of Lemma A1. First, write

$$\mu_*^\top (I_n - H_\gamma)\epsilon/\sigma = \sqrt{\mu_*^\top (I_n - H_\gamma)\mu_*} Z_\gamma,$$

where

$$Z_\gamma = \frac{\mu_*^\top (I_n - H_\gamma)\epsilon}{\sigma\sqrt{\mu_*^\top (I_n - H_\gamma)\mu_*}} \sim N(0,1).$$

According to Assumption 2.2, we have

$$\begin{aligned}
\frac{\mu_*^\top (I_n - H_\gamma)\epsilon}{\mu_*^\top (I_n - H_\gamma)\mu_*} &= \frac{\sigma\sqrt{\mu_*^\top (I_n - H_\gamma)\mu_*} Z_\gamma}{\mu_*^\top (I_n - H_\gamma)\mu_*} \\
&\leq \sigma\sqrt{\frac{Z_\gamma^2/n}{\mu_*^\top \{I_n - H_\gamma\}\mu_*/n}} \\
&< \sigma\sqrt{\frac{Z_\gamma^2}{a_0 n}} = o_p(1),
\end{aligned}$$

as $n \to \infty$. This completes our proof. □

A.3 Lemma A2

Lemma A2. *Under Assumption 2.1-2.2, for a given γ such that $\gamma_* \not\subset \gamma$ and $p_\gamma < K + p_{\gamma_*}$, we have*

$$\frac{\epsilon^\top H_\gamma \epsilon}{\mu_*^\top (I_n - H_\gamma)\mu_*} = o_p(1),$$

as $n \to \infty$.

Proof of Lemma A2. Note that $\epsilon^\top H_\gamma \epsilon/\sigma^2 \sim \chi^2_{p_\gamma}$ and $\chi^2_{p_\gamma} = o_p(n)$. Under Assumption 2.1-2.2, this implies that

$$\frac{\epsilon^\top H_\gamma \epsilon}{\mu_*^\top (I_n - H_\gamma)\mu_*} = \frac{\sigma^2 \epsilon H_\gamma \epsilon/n\sigma^2}{\mu_*^\top (I_n - H_\gamma)\mu_*/n} < \frac{\sigma^2 \epsilon H_\gamma \epsilon/n\sigma^2}{a_0} = o_p(1),$$

as $n \to \infty$. This completes our proof. □

A.4 Proof of Lemma 2.1

Proof of Lemma 2.1. For a given $\gamma \in \mathcal{M}_1$, we have

$$\log \frac{p(y|\gamma_*)}{p(y|\gamma)} = \frac{p_\gamma - p_{\gamma_*}}{2}\log(1+g) + \frac{n}{2}\log\left\{\frac{y^\top(I_n - \frac{g}{1+g}H_\gamma)y}{y^\top(I_n - \frac{g}{1+g}H_{\gamma_*})y}\right\}.$$

From the condition that g grows with n, it follows that

$$\log\left\{\frac{y^\top(I_n - \frac{g}{1+g}H_\gamma)y}{y^\top(I_n - \frac{g}{1+g}H_{\gamma_*})y}\right\} - \log\left\{\frac{y^\top(I_n - H_\gamma)y}{y^\top(I_n - H_{\gamma_*})y}\right\} = o_p(1), \tag{A.1}$$

as $n \to \infty$. According to Lemma A1 and Lemma A2, we have

$$\begin{aligned}
&\log\left\{\frac{y^\top(I_n - H_\gamma)y}{y^\top(I_n - H_{\gamma_*})y}\right\} = \log\left\{1 + \frac{y^\top(I_n - H_\gamma)y - y^\top(I_n - H_{\gamma_*})y}{y^\top(I_n - H_{\gamma_*})y}\right\} \\
&= \log\left\{1 + \frac{\mu_*^\top(I_n - H_\gamma)\mu_* + 2\mu_*^\top(I_n - H_\gamma)\epsilon + \epsilon^\top(H_{\gamma_*} - H_\gamma)\epsilon}{\epsilon^\top(I_n - H_{\gamma_*})\epsilon}\right\} \\
&= \log\left\{1 + \frac{\mu_*^\top(I_n - H_\gamma)\mu_*}{\epsilon^\top(I_n - H_{\gamma_*})\epsilon}\{1 + o_p(1)\}\right\}
\end{aligned} \tag{A.2}$$

Note that if $X \sim \chi^2_{n-c}$ for a non-negative constant c, then $X/n \to 1 + o_p(1)$ as $n \to \infty$, because $E(X/n) = \frac{n-c}{n} \to 1$ and $Var(X/n) = \frac{2(n-c)}{n^2} \to 0$ as $n \to \infty$. Since $\epsilon^\top(I_n - H_{\gamma_*})\epsilon/\sigma^2 \sim \chi^2_{n-p_{\gamma^*}}$, under Assumption 2.2, Eq (A.2) leads to

$$\log\left\{\frac{y^\top(I_n - H_\gamma)y}{y^\top(I_n - H_{\gamma_*})y}\right\} > \log\left\{1 + \frac{a_0}{\sigma^2}\{1 + o_p(1)\}\right\}.$$

Hence, as $n \to \infty$, we have

$$\log \frac{p(y|\gamma_*)}{p(y|\gamma)} > c_0 n - \frac{p_{\gamma_*} - p_\gamma}{2} \log(1+g),$$

which completes the proof. □

A.5 Proof of Lemma 2.2

Proof of Lemma 2.2. Note that $y^\top (H_\gamma - H_{\gamma_*}) y = \epsilon^\top (H_\gamma - H_{\gamma_*}) \epsilon$, when $\gamma_* \subset \gamma$. Hence, for a given $\gamma \in \mathcal{M}_2$, we have

$$\begin{aligned} \frac{n}{2} \log \left\{ \frac{y^\top (I_n - H_\gamma) y}{y^\top (I_n - H_{\gamma_*}) y} \right\} &= \frac{n}{2} \log \left\{ 1 + \frac{y^\top (I_n - H_\gamma) y - y^\top (I_n - H_{\gamma_*}) y}{y^\top (I_n - H_{\gamma_*}) y} \right\} \\ &= \frac{n}{2} \log \left\{ 1 + \frac{y^\top (H_{\gamma_*} - H_\gamma) y}{y^\top (I_n - H_{\gamma_*}) y} \right\} \\ &= \log \left\{ 1 - \frac{\epsilon^\top (H_\gamma - H_{\gamma_*}) \epsilon}{\epsilon^\top (I_n - H_{\gamma_*}) \epsilon} \right\}^{\frac{n}{2}}. \end{aligned} \tag{A.3}$$

Since $\epsilon^\top (I_n - H_{\gamma_*}) \epsilon / n\sigma^2 \to 1 + o_p(1)$ as $n \to \infty$, by Slutsky theorem and continuous mapping theorem, Eq.(A.3) implies that

$$-\frac{n}{2} \log \left\{ \frac{y^\top (I_n - H_\gamma) y}{y^\top (I_n - H_{\gamma_*}) y} \right\} \to \frac{1}{2} \chi^2_{p_\gamma - p_{\gamma_*}}$$

in distribution. It follows from (A.1) that

$$\log \frac{p(y|\gamma_*)}{p(y|\gamma)} \to \frac{p_\gamma - p_{\gamma_*}}{2} \log(1+g) - \frac{1}{2} \chi^2_{p_\gamma - p_{\gamma_*}}$$

in distribution. Note that $Pr\{\chi^2_r \geq 2x + 2(rx)^{1/2} + r\} \leq \exp(-x)$ (Armagan et al., 2013; Laurent and Massart, 2000). Letting $r_\gamma = p_\gamma - p_{\gamma_*}$, we have

$$\begin{aligned}
&\text{pr}\left(\chi^2_{r_\gamma} \geq 2r_\gamma m_n + 2(m_n r_\gamma^2)^{1/2} + r_\gamma, \text{ some } 1 \leq r_\gamma \leq K\right) \\
&\leq \sum_{j=1}^{K} \binom{p}{j} \exp(-jm_n) \\
&\leq \sum_{j=1}^{K} p^j \exp(-jm_n) \\
&= \sum_{j=1}^{K} \exp(-j\{m_n - \log p\}) \\
&= \exp(-\{m_n - \log p\}) \frac{1 - \exp\left[-(K)\{m_n - \log p\}\right]}{1 - \exp(-\{m_n - \log p\})},
\end{aligned}$$

which goes to 0 if we set $m_n = \log p + \delta_n$ with $\delta_n \to \infty$. Since the above result holds for an arbitrary δ_n, we assume $\delta_n = o(\log n)$. Hence, we have $\chi^2_{p_\gamma - p_{\gamma_*}} < 2(p_\gamma - p_{\gamma_*}) \log p\{1 + o_p(1)\}$ in probability as $n \to \infty$. Hence, we have

$$\log \frac{p(y|\gamma_*)}{p(y|\gamma)} > \frac{p_\gamma - p_{\gamma_*}}{2} \log(1+g) - (p_\gamma - p_{\gamma_*}) \log p\{1 + o_p(1)\}.$$

□

A.6 Proof of Lemma 2.3

Proof of Lemma 2.3. Let $\gamma_{**} = \gamma_* \cup \gamma$. According to Lemma 2.1, for $\gamma \in \mathcal{M}_3$, we have

$$\log \frac{p(y|\gamma_{**})}{p(y|\gamma)} > c_0 n - \frac{p_{\gamma_{**}} - p_\gamma}{2} \log(1+g).$$

Similarly, by Lemma 2.2, for $\gamma \in M_3$, we have

$$\begin{aligned}\log \frac{p(y|\gamma_*)}{p(y|\gamma_{**})} &> \frac{p_{\gamma_{**}} - p_{\gamma_*}}{2} \log(1+g) - (p_{\gamma_{**}} - p_{\gamma_*}) \log p\{1 + o_p(1)\} \\ &> \frac{p_{\gamma_{**}} - p_{\gamma_*}}{2} \log(1+g) - p_{\gamma} \log p\{1 + o_p(1)\}.\end{aligned}$$

Hence, it concludes that

$$\begin{aligned}\log \frac{p(y|\gamma_*)}{p(y|\gamma)} &= \log \frac{p(y|\gamma_{**})}{p(y|\gamma)} + \log \frac{p(y|\gamma_*)}{p(y|\gamma_{**})} \\ &> c_0 n + \frac{p_{\gamma} - p_{\gamma_*}}{2} \log(1+g) - p_{\gamma} \log p\{1 + o_p(1)\}.\end{aligned}$$

□

A.7 Proof of Theorem 2.1

Proof of Theorem 2.1. Note that, by Bayes Theorem, we have

$$\mathrm{pr}(\gamma_*|y) = \frac{p(y|\gamma_*)\mathrm{pr}(\gamma_*)}{\sum_{\gamma} p(y|\gamma)\mathrm{pr}(\gamma)}.$$

Hence, to complete the proof, it suffices to show that

$$\sum_{\gamma \neq \gamma_*} \frac{p(y|\gamma)}{p(y|\gamma_*)} \to 0,$$

in probability as $n \to \infty$.

Define $\mathcal{M}' = \mathcal{M} \setminus \{\gamma_*\}$. Since $\mathcal{M}' = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$, we have

$$\sum_{\gamma \in \mathcal{M}'} \frac{p(y|\gamma)}{p(y|\gamma_*)} = \sum_{\gamma \in M_1} \frac{p(y|\gamma)}{p(y|\gamma_*)} + \sum_{\gamma \in M_2} \frac{p(y|\gamma)}{p(y|\gamma_*)} + \sum_{\gamma \in M_3} \frac{p(y|\gamma)}{p(y|\gamma_*)}.$$

When $\gamma \in \mathcal{M}_1$, according to Lemma 2.1, we have

$$\begin{aligned}
\sum_{\gamma\in\mathcal{M}_1} \frac{p(y|\gamma)}{p(y|\gamma_*)} &= \sum_{j=0}^{p_{\gamma_*}-1} \left\{ \sum_{\gamma\in\mathcal{M}_1, p_\gamma = j} \frac{p(y|\gamma)}{p(y|\gamma_*)} \right\} \\
&< \sum_{j=0}^{p_{\gamma_*}-1} \binom{p_{\gamma_*}}{j} \exp\left\{ -c_0 n + \frac{p_{\gamma_*}-j}{2} \log(1+g) \right\} \\
&\leq \exp\left\{ -c_0 n + \frac{p_{\gamma_*}}{2} \log(1+g) \right\} \sum_{j=0}^{p_{\gamma_*}} \binom{p_{\gamma_*}}{j} \exp\left\{ -\frac{j}{2} \log(1+g) \right\} \\
&= \exp\left\{ -c_0 n + \frac{p_{\gamma_*}}{2} \log(1+g) \right\} \left\{ 1 + (1+g)^{-\frac{1}{2}} \right\}^{p_{\gamma_*}} \\
&= \exp\left[-c_0 n + \frac{p_{\gamma_*}}{2} \log(1+g) + p_{\gamma_*} \log\left\{ 1 + (1+g)^{-\frac{1}{2}} \right\} \right],
\end{aligned}$$

which goes to 0 as $n \to \infty$ under the condition that $\log g = o(n)$.

When $\gamma \in \mathcal{M}_2$, according to Lemma 2.2, we have

$$\begin{aligned}
&\sum_{\gamma\in\mathcal{M}_2} \frac{p(y|\gamma)}{p(y|\gamma_*)} = \sum_{j=p_{\gamma_*}+1}^{K} \left\{ \sum_{\gamma\in\mathcal{M}_2, p_\gamma = j} \frac{p(y|\gamma)}{p(y|\gamma_*)} \right\} \\
&< \sum_{j=p_{\gamma_*}+1}^{K} \binom{p-p_{\gamma_*}}{j-p_{\gamma_*}} \exp\left[-\frac{j-p_{\gamma_*}}{2} \log(1+g) + (j-p_{\gamma_*}) \log p \{1+o_p(1)\} \right] \\
&\leq \sum_{j=p_{\gamma_*}+1}^{K} p^{j-p_{\gamma_*}} \exp\left\{ -\frac{j-p_{\gamma_*}}{2} \log(1+g) + (j-p_{\gamma_*}) \log p \right\} \\
&= \sum_{j=p_{\gamma_*}+1}^{K} \exp\left[-(j-p_{\gamma_*}) \left\{ \frac{\log(1+g)}{2} - 2\log p \right\} \right] \\
&= \exp\left[-\frac{1}{2} \left\{ \log(1+g) - \log p^4 \right\} \right] \frac{1 - \exp\left[-\frac{K-p_{\gamma_*}}{2} \left\{ \log(1+g) - \log p^4 \right\} \right]}{1 - \exp\left[-\frac{1}{2} \left\{ \log(1+g) - \log p^4 \right\} \right]},
\end{aligned}$$

which goes to 0 as $n \to \infty$, whenever $p^4 < g$.

When $\gamma \in \mathcal{M}_3$, according to Lemma 2.3, we have

$$\begin{aligned}
\sum_{\gamma\in\mathcal{M}_3} \frac{p(y|\gamma)}{p(y|\gamma_*)} &= \sum_{j=0}^{K} \left\{ \sum_{\gamma\in\mathcal{M}_3, p_\gamma = j} \frac{p(y|\gamma)}{p(y|\gamma_*)} \right\} \\
&< \sum_{j=0}^{K} \sum_{\gamma\in\mathcal{M}_3, p_\gamma = j} \exp\left\{ -c_0 n - \frac{p_\gamma - p_{\gamma_*}}{2} \log(1+g) + p_\gamma \log p \{1 + o_p(1)\} \right\} \\
&\leq \sum_{j=0}^{K} \binom{p}{j} \exp\left\{ -c_0 n - \frac{j - p_{\gamma_*}}{2} \log(1+g) + j \log p \{1 + o_p(1)\} \right\} \\
&\leq \sum_{j=0}^{K} p^j \exp\left[-c_0 n - \frac{j - p_{\gamma_*}}{2} \log(1+g) + j \log p \{1 + o_p(1)\} \right] \\
&\leq \exp\left[-c_0 n + \frac{p_{\gamma_*}}{2} \log(1+g) \right] \sum_{j=0}^{K} \exp\left[-j \left\{ \frac{\log(1+g)}{2} - 2 \log p \{1 + o_p(1)\} \right\} \right] \\
&= \exp\left[-c_0 n + \frac{p_{\gamma_*}}{2} \log(1+g) \right] \frac{1 - \exp\left[-\frac{K+1}{2} \left\{ \log(1+g) - \log p^4 \{1 + o_p(1)\} \right\} \right]}{1 - \exp\left[-\frac{1}{2} \left\{ \log(1+g) - \log p^4 \{1 + o_p(1)\} \right\} \right]},
\end{aligned}$$

which goes to 0 as $n \to \infty$ when $p^4 < g$ and $\log g = o(n)$. Hence, this completes our proof. □

A.8 Proof of Corollary 2.1

Proof of Corollary 2.1. First, consider the case when $p \to \infty$ as $n \to \infty$. Define

$$g = p^{4(1+\delta_0)},$$

where δ_0 is a positive constant. It follows that $g > p^4$. By the assumption that $\log p = o(n)$, we have

$$\log g = 4(1 + \delta_0) \log p = o(n).$$

Second, when p is fixed. Thus, define $g = n^{\delta_0'}$ where δ_0' is a positive constant. Then, as $n \to \infty$, we have $g > p^4$ and $\log g = o(n)$. This completes our proof. □

Appendix B

Chapter 3 Preliminaries

B.1 Lemma B.1

Lemma B.1. *Under Assumption 3.5, there exists a positive constant a_0 such that*

$$\lim_{n\to\infty} \frac{1}{n}\mu_*^\top (I_n - H_\gamma)\mu_* > a_0$$

for any $\gamma \in \{\gamma : \gamma \not\supset \gamma_*, p_\gamma < K_n\}$, where $\mu_* = X_{\gamma_*}\beta^0_{\gamma_*}$ and $H_\gamma = X_\gamma(X_\gamma^\top X_\gamma)^{-1}X_\gamma^\top$.

Proof of Lemma B.1. Under Assumption 3.5, for any model γ such that $p_\gamma < K_n$, we have

$$\begin{aligned}
\frac{1}{n}\mu_*^\top (I_n - H_\gamma)\mu_* &= \frac{1}{n}\|\mu_* - H_\gamma\mu_*\|^2 \\
&= \frac{1}{n}\|X_{\gamma_*\setminus\gamma}\beta^0_{\gamma_*\setminus\gamma} - X_\gamma b_\gamma\|^2 \\
&= \frac{1}{n}\left[(\beta^0_{\gamma_*\setminus\gamma}{}^\top, -b_\gamma{}^\top)\{X_{\gamma_*\cup\gamma}{}^\top X_{\gamma_*\cup\gamma}\}(\beta^0_{\gamma_*\setminus\gamma}{}^\top, -b_\gamma{}^\top)^\top\right] \\
&= \left[(\beta^0_{\gamma_*\setminus\gamma}{}^\top, -b_\gamma{}^\top)\{\frac{1}{n}X_{\gamma_*\cup\gamma}{}^\top X_{\gamma_*\cup\gamma}\}(\beta^0_{\gamma_*\setminus\gamma}{}^\top, -b_\gamma{}^\top)^\top\right] \\
&> \lambda_{\min}\|\beta_{\gamma_*\setminus\gamma}{}^0\|^2 \\
&> a_0,
\end{aligned}$$

where a_0 is a positive constant and $b_\gamma = (X_\gamma^\top X_\gamma)^{-1}X_\gamma^\top X_{\gamma_*\setminus\gamma}\beta^0_{\gamma_*\setminus\gamma}$. □

B.2 Lemma B.2

Lemma B.2. *Under Assumptions 3.2, 3.4, 3.5 and 3.6, for any $\gamma \in \mathcal{M}_n$, there exists a positive constant d such that $\|\hat{\beta}_\gamma\|_2 \leq dn^c$ in probability as $n \to \infty$.*

Proof of Lemma B.2. Under Assumption 3.5, we first have

$$\begin{aligned}
\max_{\gamma \in \mathcal{M}_n} \| \hat{\beta}_\gamma \|_2 &= \max_{\gamma \in \mathcal{M}_n} \left(\hat{\beta}_\gamma^\top \hat{\beta}_\gamma \right)^{1/2} \\
&= \max_{\gamma \in \mathcal{M}_n} \left\{ y^\top X_\gamma (X_\gamma^\top X_\gamma)^{-1} (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top y \right\}^{1/2} \\
&\leq \max_{\gamma \in \mathcal{M}_n} \left\{ \frac{y^\top X_\gamma (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top y}{n\lambda_{\min}} \right\}^{1/2} \\
&= \max_{\gamma \in \mathcal{M}_n} \left(\frac{y^\top H_\gamma y}{n\lambda_{\min}} \right)^{1/2} \\
&= \max_{\gamma \in \mathcal{M}_n} \left(\frac{\mu_*^\top H_\gamma \mu_* + 2\mu_*^\top H_\gamma \epsilon + \epsilon^\top H_\gamma \epsilon}{n\lambda_{\min}} \right)^{1/2}.
\end{aligned}$$

Note that $Pr\{\chi_r^2 \geq 2x + 2(rx)^{1/2} + r\} \leq \exp(-x)$ by Laurent and Massart (2000). We define

$$a_j = 2b'_j + 2(jb'_j)^{1/2} + j \quad \textit{and} \quad b'_j = j \log p_n + j \log\log n.$$

Then we have

$$\begin{aligned}
\mathrm{pr}(\chi^2_{p_\gamma} \geq a_{p_\gamma}, \textit{for some } \gamma \in \mathcal{M}_n) &\leq \sum_{\gamma \in \mathcal{M}_n} \mathrm{pr}\left\{\chi^2_{p_\gamma} \geq a_{p_\gamma}\right\} \\
&\leq \sum_{j=1}^{K_n} \binom{p_n}{j} \mathrm{pr}\left\{\chi^2_j \geq a_j\right\} \\
&\leq \sum_{j=1}^{K_n} \binom{p_n}{j} \exp(-b'_j) \\
&\leq \sum_{j=1}^{K_n} p_n^j \exp(-b'_j) \\
&= \sum_{j=1}^{K_n} \exp(j \log p_n - b'_j) \\
&= \sum_{j=1}^{K_n} \left(\frac{1}{\log n}\right)^j \\
&= \frac{1}{\log n}\left\{\frac{1 - \left(\frac{1}{\log n}\right)^{K_n}}{1 - \frac{1}{\log n}}\right\}, \qquad \text{(B.1)}
\end{aligned}$$

which goes to 0 as $n \to \infty$. This implies that $\chi^2_{p_\gamma} < 2p_\gamma \log p_n\{1 + o_p(1)\}$ for any $\gamma \in \mathcal{M}_n$. Since H_γ is idempotent with rank p_γ,

$$\frac{\epsilon^\top H_\gamma \epsilon}{\sigma^2} \sim \chi^2_{p_\gamma}$$

for each $\gamma \in \mathcal{M}_n$. By Assumption 3.2, we have $\log p_n = o(n)$. As $n \to \infty$, it follows from Eq. (B.1) that

$$\max_{\gamma \in \mathcal{M}_n} \frac{\epsilon^\top H_\gamma \epsilon}{n\sigma^2} = o_p(1).$$

Suppose $\gamma_* \subset \gamma$. By Assumption 3.5, we have

$$\frac{1}{n}\mu_*^\top H_\gamma \mu_* = \frac{1}{n}\mu_*^\top \mu_* < \lambda_{\max}\|\beta^0_{\gamma_*}\|^2.$$

Suppose $\gamma_* \not\subset \gamma$. From Lemma B.1, it follows that

$$\frac{1}{n}\mu_*^\top H_\gamma \mu_* < \frac{1}{n}\mu_*^\top \mu_* - a_0 < \lambda_{\max}\|\beta_{\gamma_*}^0\|^2.$$

By Cauchy-Schwarz inequality,

$$|n^{-1}\mu_*^\top H_\gamma \epsilon| \le \sqrt{\mu_*^\top \mu_*/n}\sqrt{\epsilon^\top H_\gamma \epsilon/n} < \sqrt{\lambda_{\max}\|\beta_{\gamma_*}^0\|^2}\sqrt{\epsilon^\top H_\gamma \epsilon/n} < \lambda_{\max}\|\beta_{\gamma_*}^0\|^2,$$

in probability as $n \to \infty$.

Thus, as $n \to \infty$, exists $d \in (0, \infty)$ such that

$$\begin{aligned}\max_{\gamma \in \mathcal{M}_n}\left[\frac{\mu_*^\top H_\gamma \mu_* + 2\mu_*^\top H_\gamma \epsilon + \epsilon^\top H_\gamma \epsilon}{n\lambda_{\min}}\right]^{1/2} &\le \left\{\frac{3\lambda_{\max}}{\lambda_{\min}}\|\beta_{\gamma_*}^0\|^2 + o_p(1)\right\}^{1/2} \\ &\le dn^c,\end{aligned}$$

where the last inequality is based on Assumption 3.4 and 3.6. This completes the proof. □

B.3 Proof of Lemma 3.1

Proof of Lemma 3.1. Let $N_\gamma = \{\beta_\gamma \in \mathbb{R}^{p_\gamma} : \|X_\gamma(\beta_\gamma - \hat{\beta}_\gamma)\|_2 \le \sqrt{3p_\gamma \log n}\}$. Then we divide the integral as follows

$$\begin{aligned}p(y|\gamma) &= \int_{N_\gamma} p(y|\beta_\gamma)p(\beta_\gamma|\gamma)d\beta_\gamma + \int_{\mathbb{R}^{p_\gamma}/N_\gamma} p(y|\beta_\gamma)p(\beta_\gamma|\gamma)d\beta_\gamma \\ &= I_1 + I_2\end{aligned}$$

First, we consider I_1. Note that

$$\log p(y|\beta_\gamma) = \log p(y|\hat{\beta}_\gamma) - \frac{1}{2\sigma^2}(\beta_\gamma - \hat{\beta}_\gamma)^\top X_\gamma^\top X_\gamma(\beta_\gamma - \hat{\beta}_\gamma). \tag{B.2}$$

Under Assumption 3.5, for $\beta_\gamma \in N_\gamma$, we have

$$(n\lambda_{\min})^{1/2} \parallel \beta_\gamma - \hat{\beta}_\gamma \parallel_2 \leq \parallel X_\gamma(\beta_\gamma - \hat{\beta}_\gamma) \parallel_2 \leq \sqrt{3p_\gamma \log n},$$

which implies that

$$\parallel \beta_\gamma - \hat{\beta}_\gamma \parallel_2 \ \leq \ \sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}. \tag{B.3}$$

Thus, when n is sufficiently large, we have

$$\|\beta_\gamma\|_2 - \|\hat{\beta}_\gamma\|_2 \leq \|\beta_\gamma - \hat{\beta}_\gamma\|_2 \leq 1$$

for any $\beta_\gamma \in N_\gamma$. By Lemma B.2, it follows that

$$\parallel \beta_\gamma \parallel_2 \leq dn^c + 1.$$

Therefore, by Assumption 3.7 and Eq. (B.3), we have

$$\log p(\hat{\beta}_\gamma|\gamma) - F_1\sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}} < \log p(\beta_\gamma|\gamma) < \log p(\hat{\beta}_\gamma|\gamma) + F_1\sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}. \tag{B.4}$$

For notation simplicity, we denote Eq. (B.4) by $\log p(\beta_\gamma|\gamma) = \log p(\hat{\beta}_\gamma|\gamma) \pm F_1\sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}$.

Using the singular value decomposition, we can write

$$X_\gamma = L_\gamma D_\gamma {U_\gamma}^\top,$$

where ${L_\gamma}^\top L_\gamma = I_{p_\gamma}$, ${U_\gamma}^\top U_\gamma = I_{p_\gamma}$ and D_γ is a $p_\gamma \times p_\gamma$ diagonal matrix.

From Eq. (B.2) and Eq. (B.4), it follows that

$$\begin{aligned} I_1 &= p(y|\hat{\beta}_\gamma)p(\hat{\beta}_\gamma|\gamma)\exp\left(\pm F_1\sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}\right) \\ &\quad \times \int_{N_\gamma} \exp\{-\frac{1}{2\sigma^2}(\beta_\gamma - \hat{\beta}_\gamma)^\top X_\gamma^\top X_\gamma(\beta_\gamma - \hat{\beta}_\gamma)\}d\beta_\gamma. \end{aligned}$$

Let $\xi = \frac{D_\gamma {U_\gamma}^\top}{\sigma}(\beta_\gamma - \hat{\beta}_\gamma)$. Note that $\|\xi\|_2 = \sqrt{\frac{1}{\sigma^2}(\beta_\gamma - \hat{\beta}_\gamma)^\top {X_\gamma}^\top X_\gamma(\beta_\gamma - \hat{\beta}_\gamma)}$, thus we have

$$\begin{aligned} &\int_{N_\gamma} \exp\{-\frac{1}{2\sigma^2}(\beta_\gamma - \hat{\beta}_\gamma)^\top X_\gamma^\top X_\gamma(\beta_\gamma - \hat{\beta}_\gamma)\}d\beta_\gamma \\ &= |\frac{X_\gamma^\top X_\gamma}{\sigma^2}|^{-\frac{1}{2}} \int_{\|\xi\|_2 \le \sqrt{3p_\gamma \log n}} \exp(-\frac{1}{2}\xi^\top \xi)d\xi \\ &= \frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \mathrm{pr}(\chi^2_{p_\gamma} \le 3p_\gamma \log n) \\ &\ge \frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \{1 - \exp(-p_\gamma \log n)\}. \end{aligned}$$

This implies that

$$I_l < I_1 < I_u, \tag{B.5}$$

where

$$I_l = p(y|\hat{\beta}_\gamma)p(\hat{\beta}_\gamma|\gamma)\frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \exp\left(-F_1\sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}\right)\{1 - \exp(-p_\gamma \log n)\},$$

and

$$I_u = p(y|\hat{\beta}_\gamma)p(\hat{\beta}_\gamma|\gamma)\frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \exp\left(F_1\sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}\right).$$

We now consider I_2. Under Assumptions 3.7 and 3.8, it follows that

$$\log p(\beta_\gamma|\gamma) \le \log p(0|\gamma) + F_2 \le \log p(\hat{\beta}_\gamma|\gamma) + F_1 \parallel \hat{\beta}_\gamma \parallel_2 + F_2.$$

Hence, by Assumption 3.6 and Lemma B.2, we have

$$\|\hat{\beta}_\gamma\|_2 \leq \min(c_\gamma p_\gamma, dn^c)$$

in probability as $n \to \infty$, form some positive constants c_γ. Hence, we have

$$\begin{aligned} I_2 &\leq p(y|\hat{\beta}_\gamma)p(\hat{\beta}_\gamma|\gamma)\exp(cp_\gamma) \\ &\quad \times \int_{\mathbb{R}^{p_\gamma}/N_\gamma} \exp\left\{-\frac{1}{2\sigma^2}(\beta_\gamma - \hat{\beta}_\gamma)^\top X_\gamma^\top X_\gamma(\beta_\gamma - \hat{\beta}_\gamma)\right\} d\beta_\gamma, \end{aligned}$$

in probability as $n \to \infty$.

Let $\xi = \frac{D_\gamma U_\gamma{}^\top}{\sigma}(\beta_\gamma - \hat{\beta}_\gamma)$, we have

$$\begin{aligned} &\int_{\mathbb{R}^{p_\gamma}/N_\gamma} \exp\left\{-\frac{1}{2\sigma^2}(\beta_\gamma - \hat{\beta}_\gamma)^\top X_\gamma^\top X_\gamma(\beta_\gamma - \hat{\beta}_\gamma)\right\} d\beta_\gamma \\ &= \left|\frac{X_\gamma^\top X_\gamma}{\sigma^2}\right|^{-1/2} \int_{\|\xi\|_2 > \sqrt{3p_\gamma \log n}} \exp\left(-\frac{1}{2}\xi^\top \xi\right) d\xi \\ &= \frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \mathrm{pr}(\chi^2_{p_\gamma} > 3p_\gamma \log n) \\ &\leq \frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \exp(-p_\gamma \log n). \end{aligned}$$

Thus, we have

$$I_2 < p(y|\hat{\beta}_\gamma)p(\hat{\beta}_\gamma|\gamma)|X_\gamma^\top X_\gamma|^{-1/2}(2\pi\sigma^2)^{p_\gamma/2}\exp\left(F_1 c_\gamma p_\gamma + F_2 - p_\gamma \log n\right) \tag{B.6}$$

Note that $e^x \leq 1 + 2x$ and $e^{-x} \geq 1 - 2x$ for $0 \leq x \leq 1$. From Eq. (B.5), Eq. (B.6) and

the fact that $0 \leq I_2$, it follows that

$$\begin{aligned} & p(y|\gamma) = I_1 + I_2 \\ & > p(y|\hat{\beta}_\gamma) p(\hat{\beta}_\gamma|\gamma) \frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \exp\left(-F_1 \sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}\right)\left(1 - \frac{1}{n^{p_\gamma}}\right) \\ & \geq p(y|\hat{\beta}_\gamma) p(\hat{\beta}_\gamma|\gamma) |X_\gamma^\top X_\gamma|^{-1/2} (2\pi\sigma^2)^{p_\gamma/2} \left(1 - 2F_1 \sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}\right)\left(1 - \frac{1}{n^{p_\gamma}}\right). \quad \text{(B.7)} \end{aligned}$$

We also have

$$\begin{aligned} & p(y|\gamma) = I_1 + I_2 \\ & \leq p(y|\hat{\beta}_\gamma) p(\hat{\beta}_\gamma|\gamma) \frac{(2\pi\sigma^2)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \left\{ \exp\left(F_1 \sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}}\right) + \exp\left(c_\gamma p_\gamma - p_\gamma \log n\right) \right\} \\ & \leq p(y|\hat{\beta}_\gamma) p(\hat{\beta}_\gamma|\gamma) \frac{(2\pi)^{p_\gamma/2}}{|X_\gamma^\top X_\gamma|^{1/2}} \left\{ 1 + 2F_1 \sqrt{\frac{3p_\gamma \log n}{n\lambda_{\min}}} + \exp\left(c_\gamma p_\gamma - p_\gamma \log n\right) \right\}. \quad \text{(B.8)} \end{aligned}$$

By Eq. (B.7) and Eq. (B.8), we have

$$p(y|\gamma) = p(y|\hat{\beta}_\gamma) p(\hat{\beta}_\gamma|\gamma) |X_\gamma^\top X_\gamma|^{-1/2} (2\pi\sigma^2)^{p_\gamma/2} \left\{ 1 + O_p\left(\sqrt{\frac{p_\gamma \log n}{n}}\right) \right\}.$$

This completes the proof. □

B.4 Proof of Lemma 3.2

Proof of Lemma 3.2. For any γ such that $\gamma_* \subsetneq \gamma$, we have

$$-2 \log \left\{ \frac{p(y \mid \hat{\beta}_{\gamma_*})}{p(y \mid \hat{\beta}_\gamma)} \right\} = \frac{\epsilon^\top (H_\gamma - H_{\gamma_*}) \epsilon}{\sigma^2}. \quad \text{(B.9)}$$

Since $\epsilon \sim N(0, \sigma^2)$, it follows that

$$\frac{\epsilon^\top (H_\gamma - H_{\gamma_*}) \epsilon}{\sigma^2} \sim \chi^2_{p_\gamma - p_{\gamma_*}}.$$

Recall that $\text{pr}\{\chi^2_r \geq 2x + 2(rx)^{1/2} + r\} \leq \exp(-x)$. Let $r_\gamma = p_\gamma - p_{\gamma_*}$, we have

$$\begin{aligned}
&\text{pr}(\chi^2_{r_\gamma} \geq 2r_\gamma m_n + 2(m_n r_\gamma^2)^{1/2} + r_\gamma, \textit{some } 1 \leq r_\gamma < n - p_{\gamma_*}) \\
&\leq \sum_{j=1}^{n-p_{\gamma_*}} \binom{p_n}{j} \exp(-jm_n) \\
&\leq \sum_{j=1}^{n} p_n^j \exp(-jm_n) \\
&= \sum_{j=1}^{n} \exp(-j\{m_n - \log p_n\}) \\
&= \exp\left[-\{m_n - \log p_n\}\right] \\
&\times \frac{1 - \exp\left[-n\{m_n - \log p_n\}\right]}{1 - \exp\left[-\{m_n - \log p_n\}\right]},
\end{aligned}$$

which goes to 0 as $n \to \infty$, if $m_n = \log p_n + \delta_n$ with $\delta_n \to \infty$ as n increases.

Thus, Eq. (B.9) implies that, for any γ such that $\gamma_* \subsetneq \gamma$,

$$\begin{aligned}
-2\log \frac{p(y|\hat{\beta}_{\gamma_*})}{p(y|\hat{\beta}_\gamma)} &< 2r_\gamma(\log p_n + \delta_n) + 2\sqrt{r_\gamma^2(\log p_n + \delta_n)} + r_\gamma \\
&\equiv r_\gamma \Lambda_n,
\end{aligned}$$

in probability as $n \to \infty$, where $\Lambda_n = 2(\log p_n + \delta_n) + 2\sqrt{\log p_n + \delta_n} + 1$. Since the above result holds for an arbitrary δ_n, we assume $\delta_n = o(\log n)$. The proof is thus completed. □

B.5 Proof of Lemma 2.3

Proof of Lemma 2.3. For any γ such that $\gamma \subsetneq \gamma_*$, we have

$$2\log\left\{\frac{p(y \mid \hat{\beta}_{\gamma_*})}{p(y \mid \hat{\beta}_\gamma)}\right\} = \frac{\mu_*^\top (I_n - H_\gamma)\mu_* + 2\mu_*^\top (I_n - H_\gamma)\epsilon + \epsilon^\top H_{\gamma_*}\epsilon - \epsilon^\top H_\gamma \epsilon}{\sigma^2}. \quad \text{(B.10)}$$

First, since H_γ is idempotent for any γ, thus

$$\frac{\epsilon^\top H_\gamma \epsilon}{\sigma^2} \sim \chi^2_{p_\gamma}.$$

Recall that $\text{pr}\{\chi^2_r \geq 2x + 2(rx)^{1/2} + r\} \leq \exp(-x)$. Define

$$a_j = 2b'_j + 2(jb'_j)^{1/2} + j \quad \textit{and} \quad b'_j = j \log p_n + j \log\log n.$$

It then follows that

$$\begin{aligned}
\text{pr}(\chi^2_{p_\gamma} \geq a_{p_\gamma}, \textit{for some } 0 \leq p_\gamma < p_{\gamma_*}) &\leq \sum_{j=1}^{p_{\gamma_*}} \binom{p_n}{j} \exp(-b'_j) \\
&\leq \sum_{j=1}^{p_{\gamma_*}} p_n^j \exp(-b'_j) \\
&= \sum_{j=1}^{p_{\gamma_*}} \exp(j \log p_n - b'_j) \\
&= \sum_{j=1}^{p_{\gamma_*}} \left(\frac{1}{\log n}\right)^j \\
&= \frac{1}{\log n} \left\{ \frac{1 - \left(\frac{1}{\log n}\right)^{p_{\gamma_*}}}{1 - \frac{1}{\log n}} \right\},
\end{aligned}$$

which goes to 0 as $n \to \infty$. Thus, we have $\chi^2_{p_\gamma} < 2p_\gamma \log p_n\{1 + o_p(1)\}$ in probability for any $\gamma \subsetneq \gamma_*$ as $n \to \infty$. $p_{\gamma_*} = o(n)$ implies that $\frac{\epsilon^\top H_\gamma \epsilon}{n\sigma^2}$ decreases to 0 in probability as $n \to 0$ for any $\gamma \subsetneq \gamma_*$. Hence, following Assumption 3.2 and Lemma B.1, we have

$$\begin{aligned}
\max_{\gamma \subsetneq \gamma_*} \frac{\epsilon^\top H_\gamma \epsilon}{\mu_*^\top (I_n - H_\gamma)\mu_*} &= \max_{\gamma \subsetneq \gamma_*} \frac{\epsilon^\top H_\gamma \epsilon / n\sigma^2}{\mu_*^\top (I_n - H_\gamma)\mu_* / n\sigma^2} \\
&< \frac{2p_\gamma \log p_n\{1 + o_p(1)\}/n}{b_0} \\
&= o_p(1). \qquad \text{(B.11)}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}\max_{\gamma\subsetneq\gamma_*}\frac{\epsilon^\top H_{\gamma_*}\epsilon}{\mu_*^\top(I_n-H_\gamma)\mu_*} &= \max_{\gamma\subsetneq\gamma_*}\frac{\epsilon^\top H_{\gamma_*}\epsilon/n\sigma^2}{\mu_*^\top(I_n-H_\gamma)\mu_*/n\sigma^2}\\ &< \frac{2p_{\gamma_*}\log p_n\{1+o_p(1)\}/n}{b_0}\\ &= o_p(1). \end{aligned} \tag{B.12}$$

Second, let

$$Z_\gamma = \frac{\mu_*^\top(I_n-H_\gamma)\epsilon}{\sigma\sqrt{\mu_*^\top(I_n-H_\gamma)\mu_*}},$$

which follows $\mathcal{N}(0,1)$ for any $\gamma\subsetneq\gamma_*$. We then have

$$\begin{aligned}\max_{\gamma\subsetneq\gamma_*}\left|\frac{\mu_*^\top(I_n-H_\gamma)\epsilon}{\mu_*^\top(I_n-H_\gamma)\mu_*}\right| &= \max_{\gamma\subsetneq\gamma_*}\left\{\frac{\sigma^2Z_\gamma^2}{\mu_*^\top(I_n-H_\gamma)\mu_*}\right\}^{1/2}\\ &< \left\{\frac{2\sigma^2\log p_n\{1+o_p(1)\}}{b_0 n}\right\}^{1/2}\\ &= o_p(1), \end{aligned} \tag{B.13}$$

as $n\to\infty$ by Assumption 3.2 and Lemma B.1.

Thus, by Eq. (B.11), Eq. (B.12), Eq. (B.13) and Lemma B.1, for any γ such that $\gamma\subsetneq\gamma_*$ and $p_{\gamma_*}=o(n)$, (B.10) implies that

$$\begin{aligned}\min_{\gamma\subsetneq\gamma_*}2\log\left\{\frac{p(y\mid\hat\beta_{\gamma_*})}{p(y\mid\hat\beta_\gamma)}\right\} &= \frac{\mu_*^\top(I_n-H_\gamma)\mu_*}{\sigma^2}\{1+o_p(1)\}\\ &> b_0 n,\end{aligned}$$

for some positive constant b_0. This completes our proof. □

B.6 Proof of Theorem 3.1

Proof of Theorem 3.1. To complete the proof of Theorem 3.1, we consider two cases. The high-dimensional case in which $p_n=O(n^\alpha)$ where $\alpha\in[1,\infty)$, and the low-dimensional case

in which $p_n = O(n^\alpha)$ where $\alpha \in (0,1)$. We only show the proof of the high-dimensional case. The low-dimensional case can be proved in a similar way with K_n replaced by p_n.

First, define $\mathcal{M}_1 = \{\gamma : \gamma_* \subsetneq \gamma, \gamma \in \mathcal{M}_n\}$. For any $\gamma \in \mathcal{M}_1$, under Assumption 3.5, Lemma 3.1 and Lemma 3.2, by ignoring smaller order term of $o_p(1)$, we have

$$\begin{aligned}
&-2\log\left\{\frac{p(y \mid \gamma_*)}{p(y \mid \gamma)}\right\} = -2\log\left\{\frac{p(y|\hat{\beta}_{\gamma_*}, \gamma_*)p(\hat{\beta}_{\gamma_*}|\gamma_*)|X_{\gamma_*}^\top X_{\gamma_*}|^{-1/2}(2\pi\sigma^2)^{p_{\gamma_*}/2}}{p(y|\hat{\beta}_\gamma, \gamma)p(\hat{\beta}_\gamma|\gamma)|X_\gamma^\top X_\gamma|^{-1/2}(2\pi\sigma^2)^{p_\gamma/2}}\right\} \\
&< r_\gamma \Lambda_n - 2\log\left\{\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)}\right\} - \log\left(\frac{|X_\gamma^\top X_\gamma|}{|X_{\gamma_*}^\top X_{\gamma_*}|}\right) - (p_{\gamma_*} - p_\gamma)\log 2\pi\sigma^2 \\
&< r_\gamma \Lambda_n - r_\gamma \log n - r_\gamma \pi_n, \qquad \text{(B.14)}
\end{aligned}$$

in probability as $n \to \infty$, where $r_\gamma = p_\gamma - p_{\gamma_*}$, $\pi_n = \frac{2}{r_\gamma}\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)}$, and $\Lambda_n = 2(\log p_n + \delta_n) + \sqrt{2(\log p_n + \delta_n)} + 1$ with $\delta_n = o(\log n)$. By Assumption 3.2, we have

$$\Lambda_n = 2\log p_n + 2\delta_n + \sqrt{2(\log p_n + \delta_n)} + 1 = 2\log p_n + o(\log n).$$

By Eq. (B.14), it follows that

$$\begin{aligned}
&\sum_{\gamma \in \mathcal{M}_1} \frac{p(y|\gamma)}{p(y|\gamma_*)} = \sum_{j=1}^{K_n - p_{\gamma_*}} \left\{ \sum_{\gamma \in \mathcal{M}_1, p_\gamma - p_{\gamma_*} = j} \frac{p(y|\gamma)}{p(y|\gamma_*)} \right\} \\
&< \sum_{j=1}^{K_n - p_{\gamma_*}} \left[\binom{p_n - p_{\gamma_*}}{j} \exp\left\{ -\frac{j}{2}(\log n + \pi_n - \Lambda_n) \right\} \right] \\
&< \sum_{j=1}^{K_n} p_n^j \exp\left\{ -\frac{j}{2}(\log n + \pi_n - \Lambda_n) \right\} \\
&= \sum_{j=1}^{K_n} \exp\left\{ -\frac{j}{2}(\log n + \pi_n - \Lambda_n - 2\log p_n) \right\} \\
&= \exp\left\{ -\frac{1}{2}(\log n + \pi_n - \Lambda_n - 2\log p_n) \right\} \frac{1 - \exp\left\{ -\frac{K_n}{2}(\log n + \pi_n - \Lambda_n - 2\log p_n) \right\}}{1 - \exp\left\{ -\frac{1}{2}(\log n + \pi_n - \Lambda_n - 2\log p_n) \right\}} \\
&= \exp\left[-\frac{1}{2}\{\log n + \pi_n - 4\log p_n - o(\log n)\} \right] \\
&\quad \times \frac{1 - \exp\left[-\frac{K_n}{2}\{\log n + \pi_n - 4\log p_n - o(\log n)\} \right]}{1 - \exp\left[-\frac{1}{2}\{\log n + \pi_n - 4\log p_n - o(\log n)\} \right]},
\end{aligned}$$

which goes to 0 as $n \to \infty$ under the condition that $\pi_n > 4 \log p_n$, i.e. $\frac{1}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} > \log \frac{p_n^2}{\sqrt{n}}$.

Second, let $\mathcal{M}_2 = \{\gamma : \gamma \subsetneq \gamma_*, \gamma \in \mathcal{M}_n\}$. For any $\gamma \in \mathcal{M}_2$, according to Assumption 3.5, Lemma 3.1 and Lemma 3.3, by ignoring the smaller order term of $o_p(1)$, we obtain that

$$\begin{aligned}
-2 \log \left\{ \frac{p(y|\gamma_*)}{p(y|\gamma)} \right\} &= -2 \log \left\{ \frac{p(y|\hat{\beta}_{\gamma_*}, \gamma_*) p(\hat{\beta}_{\gamma_*}|\gamma_*) |X_{\gamma_*}^\top X_{\gamma_*}|^{-1/2} (2\pi\sigma^2)^{p_{\gamma_*}/2}}{p(y|\hat{\beta}_\gamma, \gamma) p(\hat{\beta}_\gamma|\gamma) |X_\gamma^\top X_\gamma|^{-1/2} (2\pi\sigma^2)^{p_\gamma/2}} \right\} \\
&< -b_0 n - 2 \log \left\{ \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} \right\} - \log \left(\frac{|X_\gamma^\top X_\gamma|}{|X_{\gamma_*}^\top X_{\gamma_*}|} \right) - (p_{\gamma_*} - p_\gamma) \log 2\pi\sigma^2 \\
&< -b_0 n + (p_{\gamma_*} - p_\gamma) \log n + (p_{\gamma_*} - p_\gamma) \pi_n, \qquad \text{(B.15)}
\end{aligned}$$

in probability as $n \to \infty$, where $\pi_n = \frac{2}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)}$. From Assumption 3.4 and Eq. (B.15), it follows that

$$\begin{aligned}
\sum_{\gamma \in \mathcal{M}_2} \frac{p(y|\gamma)}{p(y|\gamma_*)} &= \sum_{j=0}^{p_{\gamma_*}-1} \left\{ \sum_{\gamma \in \mathcal{M}_2 : p_\gamma = j} \frac{p(y|\gamma)}{p(y|\gamma_*)} \right\} \\
&< \sum_{j=0}^{p_{\gamma_*}-1} \left[\sum_{\gamma \in \mathcal{M}_2 : p_\gamma = j} \exp \left\{ -\frac{1}{2} \left[b_0 n - (p_{\gamma_*} - p_\gamma) \log n - (p_{\gamma_*} - p_\gamma) \pi_n \right] \right\} \right] \\
&\leq \exp \left\{ -\frac{1}{2} (b_0 n - p_{\gamma_*} \log n - p_{\gamma_*} \pi_n) \right\} \sum_{j=0}^{p_{\gamma_*}} \binom{p_{\gamma_*}}{j} \exp \left\{ -\frac{j}{2} (\log n + \pi_n) \right\} \\
&= \exp \left\{ -\frac{1}{2} (b_0 n - p_{\gamma_*} \log n - p_{\gamma_*} \pi_n) \right\} \left[1 + \exp \left\{ -\frac{\log n + \pi_n}{2} \right\} \right]^{p_{\gamma_*}},
\end{aligned}$$

which goes to 0 as $n \to \infty$ under the condition that $\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o(n)$.

The last, let $\mathcal{M}_3 = \{\gamma : \gamma \neq \gamma_*, \gamma \notin \mathcal{M}_1 \cup \mathcal{M}_2, \gamma \in \mathcal{M}_n\}$. Suppose $\gamma_{**} = \gamma_* \cup \gamma$ for $\gamma \in \mathcal{M}_3$. By Lemma 3.2 and Lemma 3.3, we obtain that

$$\begin{aligned}
-2 \log \left\{ \frac{p(y|\hat{\beta}_{\gamma_*})}{p(y|\hat{\beta}_\gamma)} \right\} &= -2 \log \left\{ \frac{p(y|\hat{\beta}_{\gamma_*})}{p(y|\hat{\beta}_{\gamma_{**}})} \right\} - 2 \log \left\{ \frac{p(y|\hat{\beta}_{\gamma_{**}})}{p(y|\hat{\beta}_\gamma)} \right\} \\
&< (p_{\gamma_{**}} - p_{\gamma_*}) \Lambda_n - c_0 n \\
&< -\frac{c_0}{2} n \qquad \text{(B.16)}
\end{aligned}$$

in probability for some positive constant c_0 as $n \to \infty$, where the last inequality is obtained by following the fact that $(p_{\gamma_{**}} - p_{\gamma_*})\Lambda_n \leq 2p_\gamma \log p_n + p_\gamma o(\log n) < c_0 n/2$ for sufficiently large n. By Assumption 3.5 and Eq. (B.16), it follows that

$$\begin{aligned}
&-2\log\left\{\frac{p(y|\gamma_*)}{p(y|\gamma)}\right\} = -2\log\left\{\frac{p(y|\hat{\beta}_{\gamma_*},\gamma_*)p(\hat{\beta}_{\gamma_*}|\gamma_*)|X_{\gamma_*}^\top X_{\gamma_*}|^{-1/2}(2\pi\sigma^2)^{p_{\gamma_*}/2}}{p(y|\hat{\beta}_\gamma,\gamma)p(\hat{\beta}_\gamma|\gamma)|X_\gamma^\top X_\gamma|^{-1/2}(2\pi\sigma^2)^{p_\gamma/2}}\right\} \\
&< -\frac{c_0}{2}n + (p_{\gamma_*} - p_\gamma)\log n + (p_{\gamma_*} - p_\gamma)\pi_n - (p_{\gamma_*} - p_\gamma)\log 2\pi\sigma^2 \\
&< -d_0 n, \qquad \text{(B.17)}
\end{aligned}$$

in probability as $n \to \infty$, where $d_0 = c_0/4$ and the last inequality holds when $p_{\gamma_*}\pi_n = o(n)$ where $\pi_n = \frac{2}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)}$. It then follows from Eq. (B.17) that

$$\begin{aligned}
\sum_{\gamma\in\mathcal{M}_3}\frac{p(y|\gamma)}{p(y|\gamma_*)} &= \sum_{j=0}^{K_n}\left\{\sum_{\gamma\in\mathcal{M}_3:p_\gamma=j}\frac{p(y|\gamma)}{p(y|\gamma_*)}\right\} \\
&< \sum_{j=0}^{K_n}\left\{\binom{p_n}{j}\exp\left(-\frac{1}{2}d_0 n\right)\right\} \\
&< \exp\left(-\frac{d_0}{2}n\right)\sum_{j=0}^{K_n} p_n^j \\
&= \exp\left(-\frac{d_0}{2}n\right)p_n^{K_n}\{1+o(1)\} \\
&= \exp\left(-\frac{d_0}{2}n + K_n\log p_n\right)\{1+o(1)\},
\end{aligned}$$

which goes to 0 as $n \to \infty$, since $K_n \log p_n = o(n)$ under Assumptions 3.2 and 3.3.

Since, $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are disjoint, it follows that

$$\sum_{\gamma\in\mathcal{M}_n}\frac{p(y|\gamma)}{p(y|\gamma_*)} = \sum_{\gamma\in\mathcal{M}_1}\frac{p(y|\gamma)}{p(y|\gamma_*)} + \sum_{\gamma\in\mathcal{M}_2}\frac{p(y|\gamma)}{p(y|\gamma_*)} + \sum_{\gamma\in\mathcal{M}_3}\frac{p(y|\gamma)}{p(y|\gamma_*)},$$

which goes to 0 in probability as $n \to \infty$. This completes the proof. □

B.7 Lemma B.3

Lemma B.3. *For model $\gamma \in \mathcal{M}_n$, the univariate horseshoe density $p(\beta_i|\tau)$ satisfies the following: (a) $\lim_{\beta_i \to 0} p(\beta_i|\tau) = \infty$. (b) For $\beta_i \neq 0$, we have*

$$\frac{K_0}{2}(\tau^2)^{-1/2} \log\left(1 + \frac{4\tau^2}{\beta_i^2}\right) < p(\beta_i|\tau) < K_0(\tau^2)^{-1/2} \log\left(1 + \frac{2\tau^2}{\beta_i^2}\right)$$

where $K_0 = 1/(2\pi^3)^{1/2}$ and $i = 1, \ldots, p_\gamma$.

Proof. First, for model γ, we have

$$p(\beta_i|\tau) = \int_0^\infty (2\pi\lambda_i^2)^{-1/2} \exp\left(-\beta_i^2/2\lambda_i^2\right) \frac{2}{\pi\tau[1 + (\lambda_i/\tau)^2]} d\lambda_i.$$

Let $u = (\tau/\lambda_i)^2$, then we have

$$\begin{aligned} p(\beta_i|\tau) &= \int_0^\infty (2\pi\tau^2)^{-1/2} u^{1/2} \exp\left(-\frac{\beta_i^2}{2\tau^2}u\right) \frac{u^{-3/2}\tau}{\pi\tau(1 + 1/u)} du \\ &= \frac{K_0}{(\tau^2)^{1/2}} \int_0^\infty \frac{\exp\left(-\frac{\beta_i^2}{2\tau^2}u\right)}{1 + u} du, \end{aligned}$$

where $K_0 = 1/(2\pi^3)^{1/2}$. Let $z = u + 1$, it follows that

$$\begin{aligned} p(\beta_i|\tau) &= \frac{K_0}{(\tau^2)^{1/2}} \int_1^\infty \exp\left[-\frac{\beta_i^2}{2\tau^2}(z - 1)\right] \frac{1}{z} dz \\ &= \frac{K_0}{(\tau^2)^{1/2}} \exp(\frac{\beta_i^2}{2\tau^2}) \int_1^\infty \frac{1}{z} \exp\left[-\frac{\beta_i^2}{2\tau^2}z\right] dz \\ &= \frac{K_0}{(\tau^2)^{1/2}} \exp(\frac{\beta_i^2}{2\tau^2}) E_1(\frac{\beta_i^2}{2\tau^2}), \end{aligned}$$

where $E_1(\cdot)$ is the exponential integral function. This function satisfies tight upper and lower bounds:

$$\frac{\exp(-t)}{2} \log\left(1 + \frac{2}{t}\right) < E_1(t) < \exp(-t) \log\left(1 + \frac{1}{t}\right)$$

for all $t > 0$. Thus we have

$$\frac{K_0}{2}(\tau^2)^{-1/2}\log\left(1+\frac{4\tau^2}{\beta_i^2}\right) < p(\beta_i|\tau) < K_0(\tau^2)^{-1/2}\log\left(1+\frac{2\tau^2}{\beta_i^2}\right)$$

which proves Part (b) and Part (a) then follows from the lower bound approaches ∞ as $\beta_i \to 0$.

□

B.8 Examples of priors

Proof of Gaussian prior. First, by Lemma B.2, for $\gamma \in \mathcal{M}_1$, we have

$$\begin{aligned}
\frac{1}{p_\gamma - p_{\gamma_*}}\log\frac{p(\hat\beta_{\gamma_*}|\gamma_*,s)}{p(\hat\beta_\gamma|\gamma,s)} &= \frac{1}{p_\gamma - p_{\gamma_*}}\log\frac{\prod_{k\in\gamma_*}\frac{1}{(2\pi s^2)^{1/2}}\exp\left(-\frac{1}{2s^2}\beta_k^2\right)}{\prod_{j\in\gamma}\frac{1}{(2\pi s^2)^{1/2}}\exp\left(-\frac{1}{2s^2}\beta_j^2\right)} \\
&= \frac{1}{2}\log(2\pi s^2) + \frac{1}{2(p_\gamma - p_{\gamma_*})s^2}\left(\sum_{j\in\gamma}\hat\beta_j^2 - \sum_{k\in\gamma_*}\hat\beta_k^2\right) \\
&= \log s + \frac{1}{2}\log(2\pi) + \frac{O_p(n^c)}{2(p_\gamma - p_{\gamma_*})s^2} \\
&> \log s.
\end{aligned}$$

When $s > p_n^2/\sqrt{n}$, the prior satisfies Condition 1 that $\frac{1}{p_\gamma - p_{\gamma_*}}\log\frac{p(\hat\beta_{\gamma_*}|\gamma_*)}{p(\hat\beta_\gamma|\gamma)} > \log p_n^2/\sqrt{n}$ for $\gamma \in \mathcal{M}_1$.

Similarly, for $\gamma \in \mathcal{M}_n \setminus \gamma_*$, we have

$$\begin{aligned}
\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}}\log\frac{p(\hat\beta_{\gamma_*}|\gamma_*,s)}{p(\hat\beta_\gamma|\gamma,s)} &= \frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}}\log\frac{\prod_{i=1}^{p_{\gamma_*}}\frac{1}{(2\pi s^2)^{1/2}}\exp\left(-\frac{1}{2s^2}\beta_k^2\right)}{\prod_{j=1}^{p_\gamma}\frac{1}{(2\pi s^2)^{1/2}}\exp\left(-\frac{1}{2s^2}\beta_j^2\right)} \\
&= \frac{p_{\gamma_*}}{2}\log(2\pi s^2) + \frac{p_{\gamma_*}}{2s^2(p_\gamma - p_{\gamma_*})}\left(\sum_{j\in\gamma}\hat\beta_j^2 - \sum_{k\in\gamma_*}\hat\beta_k^2\right) \\
&= \frac{p_{\gamma_*}}{2}\log(s^2) + \frac{p_{\gamma_*}}{2}\log(2\pi) + \frac{p_{\gamma_*}O_p(n^c)}{2s^2(p_\gamma - p_{\gamma_*})}
\end{aligned}$$

Hence, to satisfy $\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}}\log\frac{p(\hat\beta_{\gamma_*}|\gamma_*)}{p(\hat\beta_\gamma|\gamma)} = o(n)$, we must set $p_{\gamma_*}\log(s) = o(n)$. For example, we

set $s = p_n^{2+\delta}$ for an arbitrary positive constant $\delta > 0$. We then have $p_{\gamma_*} \log(s) = o(n)$ and $s > p_n^2/\sqrt{n}$. Thus, according to Theorem 3.1, the model selection consistency holds for the independent Gaussian prior. □

Proof of Laplace prior. First, by Lemma B.2, for $\gamma \in \mathcal{M}_n$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{1}{2s} \exp\left(-\frac{|\hat{\beta}_i|}{s}\right)}{\prod_{j=1}^{p_\gamma} \frac{1}{2s} \exp\left(-\frac{|\hat{\beta}_j|}{s}\right)} \\
&= (p_\gamma - p_{\gamma_*}) \log 2s + \frac{1}{s}\left(\sum_{j=1}^{p_\gamma} |\hat{\beta}_j| - \sum_{i=1}^{p_{\gamma_*}} |\hat{\beta}_i|\right) \\
&= (p_\gamma - p_{\gamma_*}) \log 2s + \frac{O_p(n^c)}{s}
\end{aligned}$$

When $s > p_n^2/\sqrt{n}$, for $\gamma \in \mathcal{M}_1$, as $n \to \infty$, we have

$$\begin{aligned}
\frac{1}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= \log 2 + \log s + \frac{O_p(n^c)}{(p_\gamma - p_{\gamma_*})s} \\
&> \log s\{1 + o_p(1)\} \\
&> \log p_n^2/\sqrt{n},
\end{aligned}$$

which satisfies Condition 1 that $\frac{1}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} > \log p_n^2/\sqrt{n}$ for $\gamma \in \mathcal{M}_1$.

When $p_{\gamma_*} \log s = o(n)$, for any $\gamma \in \mathcal{M}_n \setminus \gamma_*$, as $n \to \infty$, we have

$$\begin{aligned}
\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s)}{p(\hat{\beta}_\gamma|\gamma, s)} &= p_{\gamma_*} \log 2 + p_{\gamma_*} \log s + \frac{p_{\gamma_*} O_p(n^c)}{(p_\gamma - p_{\gamma_*})s} \\
&= o(n),
\end{aligned}$$

which satisfies Condition 2 that $\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o(n)$.

There are many ways to set up s to satisfy both conditions. For example, we can set $s = p_n^{2+\delta}$ for some $\delta > 0$, we then have $s > p_n^2/\sqrt{n}$ and $p_{\gamma_*} \log s = o(n)$. Thus, by Theorem 3.1, the model selection consistency holds for the Laplace prior. □

Proof of scaled Student's t prior. For $\gamma \in \mathcal{M}$, if $\log d = o_p(n)$, by Lemma B.2, we then have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s, d)}{p(\hat{\beta}_{\gamma}|\gamma, s, d)} &= \frac{\prod_{i=1}^{p_{\gamma_*}}[sd^{1/2}B(d/2,1/2)]^{-1}(1+\frac{\hat{\beta}_i^2}{sd})^{-(d+1)/2}}{\prod_{j=1}^{p_{\gamma}}[sd^{1/2}B(d/2,1/2)]^{-1}(1+\frac{\hat{\beta}_j^2}{sd})^{-(d+1)/2}} \\
&= (p_\gamma - p_{\gamma_*})\log s + \frac{p_\gamma - p_{\gamma_*}}{2}\log d + (p_\gamma - p_{\gamma_*})\log B(d/2,1/2) \\
&\quad + \frac{d+1}{2}\left\{\sum_{j=1}^{p_\gamma}\log(1+\frac{\hat{\beta}_j^2}{sd}) - \sum_{i=1}^{p_{\gamma_*}}\log(1+\frac{\hat{\beta}_i^2}{sd})\right\} \\
&= (p_\gamma - p_{\gamma_*})\log s + \frac{p_\gamma - p_{\gamma_*}}{2}\log d + (p_\gamma - p_{\gamma_*})\log B(d/2,1/2) \\
&\quad + \left(\sum_{j=1}^{p_\gamma}\frac{\hat{\beta}_j^2}{2s} - \sum_{i=1}^{p_{\gamma_*}}\frac{\hat{\beta}_i^2}{2s}\right) \\
&= (p_\gamma - p_{\gamma_*})\log s + \frac{p_\gamma - p_{\gamma_*}}{2}\log d + (p_\gamma - p_{\gamma_*})\log O_p(1) + \frac{O_p(n^c)}{2s}.
\end{aligned}$$

When $s > p_n^2/\sqrt{n}$, for $\gamma \in \mathcal{M}_1$, as $n \to \infty$, we have

$$\begin{aligned}
\frac{1}{p_\gamma - p_{\gamma_*}}\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s, d)}{p(\hat{\beta}_{\gamma}|\gamma, s, d)} &= \log s + \frac{\log d}{2} + \log O_p(1) + \frac{O_p(n^c)}{2s(p_\gamma - p_{\gamma_*})} \\
&> \log s\{1 + o_p(1)\} \\
&> \log p_n^2/\sqrt{n},
\end{aligned}$$

which satisfies Condition 1 that $\frac{1}{p_\gamma - p_{\gamma_*}}\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} > \log p_n^2$ for $\gamma \in \mathcal{M}_1$.

When $p_{\gamma_*}\log s = o(n)$ and $p_{\gamma_*}\log d = o(n)$, for $\gamma_* \in \mathcal{M}_n \setminus \gamma_*$, as $n \to \infty$, we have

$$\begin{aligned}
\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}}\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, s, d)}{p(\hat{\beta}_{\gamma}|\gamma, s, d)} &= p_{\gamma_*}\log s + \frac{p_{\gamma_*}}{2}\log d + p_{\gamma_*}\log O_p(1) + \frac{p_{\gamma_*}O_p(n^c)}{2s(p_\gamma - p_{\gamma_*})} \\
&= o(n),
\end{aligned}$$

which satisfies Condition 2 that $\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}}\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} = o(n)$ for $\gamma \in \mathcal{M}_n \setminus \gamma_*$.

There are many ways to set up s to satisfy both conditions. For example, we set $s = p_n^{2+\delta}$ for some $\delta > 0$, we then have $s > p_n^2/\sqrt{n}$ and $p_{\gamma_*}\log s = o(n)$. Thus, according to Theorem 3.1, the model selection consistency holds for the Scaled Student's t prior. □

Proof of generalized double Pareto prior. If α and η are chosen to grow with n, by Lemma

B.2, for $\gamma \in \mathcal{M}_n$, as $n \to \infty$, we have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \alpha, \eta)}{p(\hat{\beta}_{\gamma}|\gamma, \alpha, \eta)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{\alpha}{2\eta} \left(1 + \frac{|\hat{\beta}_i|}{\eta}\right)^{-(\alpha+1)}}{\prod_{j=1}^{p_{\gamma}} \frac{\alpha}{2\eta} \left(1 + \frac{|\hat{\beta}_j|}{\eta}\right)^{-(\alpha+1)}} \\
&= (p_{\gamma_*} - p_{\gamma}) \log \frac{\alpha}{2\eta} + (\alpha+1) \left[\sum_{j=1}^{p_{\gamma}} \log(1 + \frac{|\hat{\beta}_j|}{\eta}) - \sum_{i=1}^{p_{\gamma_*}} \log(1 + \frac{|\hat{\beta}_i|}{\eta})\right] \\
&= (p_{\gamma} - p_{\gamma_*}) \log \frac{2\eta}{\alpha} + \left[\sum_{j=1}^{p_{\gamma}} \frac{\alpha}{\eta} |\hat{\beta}_j| - \sum_{i=1}^{p_{\gamma_*}} \frac{\alpha}{\eta} |\hat{\beta}_i|\right] \\
&= (p_{\gamma} - p_{\gamma_*}) \log 2 + (p_{\gamma} - p_{\gamma_*}) \log \frac{\eta}{\alpha} + \frac{\alpha}{\eta} \left[\sum_{j=1}^{p_{\gamma}} |\hat{\beta}_j| - \sum_{i=1}^{p_{\gamma_*}} |\hat{\beta}_i|\right] \\
&= (p_{\gamma} - p_{\gamma_*}) \log 2 + (p_{\gamma} - p_{\gamma_*}) \log \frac{\eta}{\alpha} + \frac{O_p(n^c)}{\eta/\alpha}.
\end{aligned}$$

When $\frac{\eta}{\alpha} > p_n^2/\sqrt{n}$, for $\gamma \in \mathcal{M}_1$, as $n \to \infty$, we have

$$\begin{aligned}
\frac{1}{p_{\gamma} - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \alpha, \eta)}{p(\hat{\beta}_{\gamma}|\gamma, \alpha, \eta)} &= \log 2 + \log \frac{\eta}{\alpha} + \frac{O_p(n^c)}{\eta/\alpha} \\
&> \log \frac{\eta}{\alpha} \{1 + o_p(1)\} \\
&> \log p_n^2/\sqrt{n},
\end{aligned}$$

which satisfies Condition 1 that $\frac{1}{p_{\gamma} - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} > \log p_n^2$ for $\gamma \in \mathcal{M}_1$.

When $p_{\gamma_*} \log \frac{\eta}{\alpha} = o(n)$, for $\gamma \in \mathcal{M}_n \setminus \gamma_*$, we have

$$\begin{aligned}
\frac{p_{\gamma_*}}{p_{\gamma} - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \alpha, \eta)}{p(\hat{\beta}_{\gamma}|\gamma, \alpha, \eta)} &= p_{\gamma_*} \log 2 + p_{\gamma_*} \log \frac{\eta}{\alpha} + \frac{p_{\gamma_*} O_p(n^c)}{(p_{\gamma} - p_{\gamma_*})\eta/\alpha} \\
&= o(n),
\end{aligned}$$

which satisfies Condition 2 that $\frac{p_{\gamma_*}}{p_{\gamma} - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} = o(n)$ for $\gamma \in \mathcal{M}_n \setminus \gamma_*$.

There are many ways to set up η/α to satisfy both conditions. For example, we can set $\eta/\alpha = p_n^{2+\delta}$ for some $\delta > 0$, we then have $\eta/\alpha > p_n^2/\sqrt{n}$ and $p_{\gamma_*} \log \frac{\eta}{\alpha} = o(n)$. Thus, according to Theorem 3.1, the model selection consistency holds for the generalized double Pareto prior.

□

Proof of Horseshoe prior. By Lemma B.7, for any $\gamma \in \mathcal{M}_n$, we first have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \tau)}{p(\hat{\beta}_{\gamma}|\gamma, \tau)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_i^2}{2\tau^2}) E_1(\frac{\hat{\beta}_i^2}{2\tau^2})}{\prod_{j=1}^{p_{\gamma}} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_j^2}{2\tau^2}) E_1(\frac{\hat{\beta}_j^2}{2\tau^2})} \\
&> \log \frac{\prod_{i=1}^{p_{\gamma_*}} \frac{K_0}{2}(\tau^2)^{-1/2} \log\left(1 + \frac{4\tau^2}{\hat{\beta}_i^2}\right)}{\prod_{j=1}^{p_{\gamma}} K_0(\tau^2)^{-1/2} \log\left(1 + \frac{2\tau^2}{\hat{\beta}_j^2}\right)} \\
&= (p_\gamma - p_{\gamma_*}) \log \frac{2}{K_0} - p_\gamma \log 2 + \frac{p_\gamma - p_{\gamma_*}}{2} \log \tau^2 + \sum_{i=1}^{p_{\gamma_*}} \log\left[\log\left(1 + \frac{4\tau^2}{\hat{\beta}_i^2}\right)\right] \\
&- \sum_{j=1}^{p_\gamma} \log\left[\log\left(1 + \frac{2\tau^2}{\hat{\beta}_j^2}\right)\right].
\end{aligned}$$

We also have

$$\begin{aligned}
\log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \tau)}{p(\hat{\beta}_{\gamma}|\gamma, \tau)} &= \log \frac{\prod_{i=1}^{p_{\gamma_*}} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_i^2}{2\tau^2}) E_1(\frac{\hat{\beta}_i^2}{2\tau^2})}{\prod_{j=1}^{p_{\gamma}} K_0(\tau^2)^{-1/2} \exp(\frac{\hat{\beta}_j^2}{2\tau^2}) E_1(\frac{\hat{\beta}_j^2}{2\tau^2})} \\
&< \log \frac{\prod_{i=1}^{p_{\gamma_*}} K_0(\tau^2)^{-1/2} \log\left(1 + \frac{2\tau^2}{\hat{\beta}_i^2}\right)}{\prod_{j=1}^{p_{\gamma}} \frac{K_0}{2}(\tau^2)^{-1/2} \log\left(1 + \frac{4\tau^2}{\hat{\beta}_j^2}\right)} \\
&= (p_\gamma - p_{\gamma_*}) \log \frac{2}{K_0} + p_{\gamma_*} \log 2 + \frac{p_\gamma - p_{\gamma_*}}{2} \log \tau^2 + \sum_{i=1}^{p_{\gamma_*}} \log\left[\log\left(1 + \frac{2\tau^2}{\hat{\beta}_i{}^2}\right)\right] \\
&- \sum_{j=1}^{p_\gamma} \log\left[\log\left(1 + \frac{4\tau^2}{\hat{\beta}_j{}^2}\right)\right].
\end{aligned}$$

For $\gamma \in \mathcal{M}_1$, when $\tau^2 \geq p_n^4/n$, as $n \to \infty$, we have

$$
\begin{aligned}
&\frac{1}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \tau)}{p(\hat{\beta}_\gamma|\gamma, \tau)} \\
&> \log \frac{2}{K_0} - \frac{p_\gamma}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{1}{2} \log \tau^2 \\
&\quad + \frac{1}{p_\gamma - p_{\gamma_*}} \left\{ \sum_{i=1}^{p_{\gamma_*}} \log \left[\log \left(1 + \frac{4\tau^2}{\hat{\beta}_i^{\,2}} \right) \right] - \sum_{j=1}^{p_\gamma} \log \left[\log \left(1 + \frac{2\tau^2}{\hat{\beta}_j^2} \right) \right] \right\} \\
&= \log \frac{2}{K_0} - \frac{p_\gamma}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{1}{2} \log \tau^2 + \frac{1}{p_\gamma - p_{\gamma_*}} \left\{ p_{\gamma_*} \log \left[o(\tau^2) \right] - p_\gamma \log \left[o(\tau^2)) \right] \right\} \\
&= \log \frac{2}{K_0} - \frac{p_\gamma}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{1}{2} \log \tau^2 - \log \left[o(\tau^2) \right] \{1 + o(n)\} \\
&= \frac{1}{2} \log \tau^2 \{1 + o(n)\} \\
&\geq \log p_n^2/\sqrt{n},
\end{aligned}
$$

which satisfies Condition 1 that $\frac{1}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} > \log p_n^2/\sqrt{n}$ for $\gamma \in \mathcal{M}_1$.

For $\gamma \in \mathcal{M}_n \setminus \gamma_*$, when τ^2 grows with n and $p_{\gamma_*} \log \tau^2 = o(n)$, as $n \to \infty$ we have

$$
\begin{aligned}
&\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \tau)}{p(\hat{\beta}_\gamma|\gamma, \tau)} \\
&> p_{\gamma_*} \log \frac{2}{K_0} - \frac{p_\gamma p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{p_{\gamma_*}}{2} \log \tau^2 \\
&\quad + \frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \left\{ \sum_{i=1}^{p_{\gamma_*}} \log \left[\log \left(1 + \frac{4\tau^2}{\hat{\beta}_i^2} \right) \right] - \sum_{j=1}^{p_\gamma} \log \left[\log \left(1 + \frac{2\tau^2}{\hat{\beta}_j^2} \right) \right] \right\} \\
&= p_{\gamma_*} \log \frac{2}{K_0} - \frac{p_\gamma p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{p_{\gamma_*}}{2} \log \tau^2 + \frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} [p_{\gamma_*} \log o(\tau^2) - p_\gamma \log o(\tau^2)] \\
&= p_{\gamma_*} \log \frac{2}{K_0} - \frac{p_\gamma p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{p_{\gamma_*}}{2} \log \tau^2 - p_{\gamma_*} \log o(\tau^2) \{1 + o(n)\} \\
&= \frac{p_{\gamma_*}}{2} \log \tau^2 \{1 + o(n)\} \\
&= o(n),
\end{aligned}
$$

and we also have

$$
\begin{aligned}
&\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*, \tau)}{p(\hat{\beta}_\gamma|\gamma, \tau)} \\
&< p_{\gamma_*} \log \frac{2}{K_0} + \frac{p_{\gamma_*}^2}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{p_{\gamma_*}}{2} \log \tau^2 \\
&\quad + \frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \left\{ \sum_{i=1}^{p_{\gamma_*}} \log \left[\log \left(1 + \frac{2\tau^2}{\hat{\beta}_i^2} \right) \right] - \sum_{j=1}^{p_\gamma} \log \left[\log \left(1 + \frac{4\tau^2}{\hat{\beta}_j^2} \right) \right] \right\} \\
&= p_{\gamma_*} \log \frac{2}{K_0} + \frac{p_{\gamma_*}^2}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{p_{\gamma_*}}{2} \log \tau^2 + \frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} [p_{\gamma_*} \log o(\tau^2) - p_\gamma \log o(\tau^2)] \\
&= p_{\gamma_*} \log \frac{2}{K_0} + \frac{p_{\gamma_*}^2}{p_\gamma - p_{\gamma_*}} \log 2 + \frac{p_{\gamma_*}}{2} \log \tau^2 - p_{\gamma_*} \log o(\tau^2) \{1 + o(n)\} \\
&= \frac{p_{\gamma_*}}{2} \log \tau^2 \{1 + o(n)\} \\
&= o(n),
\end{aligned}
$$

which satisfies Condition 2 that $\frac{p_{\gamma_*}}{p_\gamma - p_{\gamma_*}} \log \frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_\gamma|\gamma)} = o(n)$ for $\gamma \in \mathcal{M}_n \setminus \gamma_*$.

There are many ways to set up τ^2. For example, we set $\tau^2 = p_n^{4+\delta}$ for some $\delta > 0$. We then have $\tau^2 > p_n^4/n$ and $p_{\gamma_*} \log \tau^2 = o(n)$. Thus, according to Theorem 3.1, the model selection consistency holds for the Horseshoe prior. □

B.9

Under the sufficient condition of Assumption 3.7, for any $\gamma \in \mathcal{M}_n$, the Gaussian prior

$$
p(\beta_\gamma \mid \gamma) = \prod_{j \in \gamma} \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{1}{2s^2} \beta_j^2 \right)
$$

satisfies Assumptions 3.7 and 3.8 if $s^2 \geq d'_\gamma n^c$ for some constant $d'_\gamma > 0$.

Proof. First, by the sufficient condition of Assumption 3.7, we have

$$\begin{aligned}
|\log p(\beta'_\gamma|\gamma) - \log p(\beta_\gamma|\gamma)| &= |\log \frac{p(\beta'_\gamma|\gamma)}{p(\beta_\gamma|\gamma)}| \\
&= |\log \frac{\prod_{j\in\gamma} \frac{1}{(2\pi s^2)^{1/2}} \exp(-\frac{1}{2s^2}\beta'^2_j)}{\prod_{j\in\gamma} \frac{1}{(2\pi s^2)^{1/2}} \exp(-\frac{1}{2s^2}\beta_j^2)}| \\
&= |\frac{1}{2s^2}(\sum_{j\in\gamma}\beta_j^2 - \sum_{j\in\gamma}\beta'^2_j)| \\
&= |\frac{1}{2s^2}(\beta_\gamma + \beta'_\gamma)^\top(\beta_\gamma - \beta'_\gamma)| \\
&\leq \frac{1}{2s^2}\|\beta_\gamma + \beta'_\gamma\|_2\|\beta_\gamma - \beta'_\gamma\|_2 \\
&\leq \frac{1}{2s^2}(\|\beta_\gamma\|_2 + \|\beta'_\gamma\|_2)\|\beta_\gamma - \beta'_\gamma\|_2 \\
&< |\frac{2(1+d_\gamma n^c)}{2s^2}|\|\beta'_\gamma - \beta_\gamma\|_2 \\
&\leq |\frac{1+d_\gamma n^c}{d'_\gamma n^c}|\|\beta'_\gamma - \beta_\gamma\|_2 \\
&< F_1\|\beta'_\gamma - \beta_\gamma\|_2,
\end{aligned}$$

for some constants $F_1 \in (0, \infty)$.

Second, since the density of Gaussian prior maximizes at $\beta_\gamma = 0$. Thus we have

$$\log p(\beta_\gamma|\gamma) - \log p(0|\gamma) \leq F_2,$$

for some constant $F_2 \in (0, \infty)$. □

Under the Assumptions 3.3, the Laplace prior

$$p(\beta_\gamma|s,\gamma) = \prod_{j=1}^{p_\gamma} \frac{1}{2s} \exp\left(-\frac{|\beta_j|}{s}\right)$$

satisfies Assumptions 3.7 and 3.8, if $s > d'_\gamma n^{b/2}$ for some constant $d'_\gamma \in (0, \infty)$.

Proof. First, by Assumption 3.3, we have

$$\begin{aligned}
|\log p(\beta'_\gamma|\gamma) - \log p(\beta_\gamma|\gamma)| &= |\log \frac{p(\beta'_\gamma|\gamma)}{p(\beta_\gamma|\gamma)}| \\
&= |\log \frac{\prod_{j=1}^{p_\gamma} \frac{1}{2s} \exp\left(-\frac{|\beta'_j|}{s}\right)}{\prod_{j=1}^{p_\gamma} \frac{1}{2s} \exp\left(-\frac{|\beta_j|}{s}\right)}| \\
&= \frac{1}{s}|\sum_{j=1}^{p_\gamma}(\frac{|\beta_j|}{s} - \frac{|\beta'_j|}{s})| \\
&= \frac{1}{s}|(|\beta'_\gamma| - |\beta_\gamma|)^\top J_{p_\gamma,1}| \\
&\leq \frac{1}{s}\|\beta'_\gamma - \beta_\gamma\|_2 \|J_{p_\gamma,1}\|_2 \\
&= \frac{p_\gamma^{1/2}}{s}\|\beta'_\gamma - \beta_\gamma\|_2 \\
&\leq \frac{n^{b/2}}{d'_\gamma n^{b/2}}\|\beta'_\gamma - \beta_\gamma\|_2 \\
&\leq F_1\|\beta'_\gamma - \beta_\gamma\|_2,
\end{aligned}$$

if $s > d'_\gamma n^{b/2}$ for some constant $d'_\gamma \in (0, \infty)$, where $J^\top_{p_\gamma,1} = (1, 1, \ldots, 1)$ and F_1 is a positive constant.

Second, since the density of Laplace prior maximizes at $\beta_\gamma = 0$, we have

$$\log p(\beta_\gamma|\gamma) - \log p(0|\gamma) \leq F_2,$$

for some constant $F_2 \in (0, \infty)$.

□

Under the sufficient condition of Assumption 3.7, the Scaled Student's t prior

$$p(\beta_\gamma|s, d, \gamma) = \prod_{j=1}^{p_\gamma}[sd^{1/2}B(d/2, 1/2)]^{-1}(1 + \frac{\beta_j^2}{sd})^{-(d+1)/2}$$

satisfies Assumption 3.7 and Assumption 3.8, if d grows with n and $s \geq d'_\gamma n^c$ for some constant $d'_\gamma > 0$.

Proof. First, by the sufficient condition of Assumption 3.7, when n is sufficiently large, we have

$$\begin{aligned}
|\log p(\beta'_\gamma|\gamma) - \log p(\beta_\gamma|\gamma)| &= |\log \frac{\prod_{j=1}^{p_\gamma}[sd^{1/2}B(d/2,1/2)]^{-1}(1+\frac{\beta_j'^2}{sd})^{-(d+1)/2}}{\prod_{j=1}^{p_\gamma}[sd^{1/2}B(d/2,1/2)]^{-1}(1+\frac{\beta_j^2}{sd})^{-(d+1)/2}}| \\
&= |\frac{d+1}{2}\left[\sum_{j=1}^{p_\gamma}\log(1+\frac{\beta_j^2}{sd}) - \sum_{j=1}^{p_\gamma}\log(1+\frac{\beta_j'^2}{sd})\right]| \\
&= |\frac{1}{2s}\left[\sum_{j=1}^{p_\gamma}\beta_j^2 - \sum_{i=1}^{p_\gamma}\beta_j'^2\right]| \\
&= \frac{1}{2s}|(\beta'_\gamma - \beta_\gamma)^\top(\beta'_\gamma + \beta_\gamma)| \\
&\leq \frac{1}{2s}\|\beta'_\gamma - \beta_\gamma\|_2\|\beta'_\gamma + \beta_\gamma\|_2 \\
&\leq \frac{1+d_\gamma n^c}{d'_\gamma n^c}\|\beta'_\gamma - \beta_\gamma\|_2 \\
&\leq F_1\|\beta'_\gamma - \beta_\gamma\|_2,
\end{aligned}$$

if d grows with n and $s \geq d'_\gamma n^c$ for some constant $d'_\gamma > 0$, where F_1 is a positive constant.

Second, since the density of Scaled Student's t prior maximizes at $\beta_\gamma = 0$, thus we have

$$\log p(\beta_\gamma|\gamma) - \log p(0|\gamma) \leq F_2,$$

for some constant $F_2 \in (0, \infty)$. □

Under Assumptions 3.3, the Generalized Double Pareto prior

$$p(\beta_\gamma|s,\gamma) = \prod_{j=1}^{p_\gamma}\frac{\alpha}{2\eta}\left(1+\frac{|\beta_{j,\gamma}|}{\eta}\right)^{-(\alpha+1)}$$

satisfies Assumptions 3.7 and 3.8, if α, η grow with n, and we choose $\eta/\alpha > d'_\gamma n^{b/2}$ for some $d'_\gamma \in (0, \infty)$.

Proof. First, under Assumptions 3.3, when n is sufficiently large, we have

$$\begin{aligned}
|\log p(\beta'_\gamma|\gamma) - \log p(\beta_\gamma|\gamma)| &= \log \frac{\prod_{j=1}^{p_\gamma} \frac{\alpha}{2\eta}\left(1+\frac{|\beta'_j|}{\eta}\right)^{-(\alpha+1)}}{\prod_{j=1}^{p_\gamma} \frac{\alpha}{2\eta}\left(1+\frac{|\beta_j|}{\eta}\right)^{-(\alpha+1)}} \\
&= |(\alpha+1)\left[\sum_{j=1}^{p_\gamma}\log(1+\frac{|\beta_j|}{\eta}) - \sum_{j=1}^{p_\gamma}\log(1+\frac{|\beta'_j|}{\eta})\right]| \\
&= |\frac{\alpha}{\eta}\left[\sum_{j=1}^{p_\gamma}|\beta_j| - \sum_{j=1}^{p_\gamma}|\beta'_j|\right]| \\
&= \frac{\alpha}{\eta}|(|\beta'_\gamma| - |\beta_\gamma|)^\top J_{p_\gamma,1}| \\
&\leq \frac{\alpha}{\eta}\|\beta'_\gamma - \beta_\gamma\|_2 \|J_{p_\gamma,1}\|_2 \\
&= \frac{p_\gamma^{1/2}}{\eta/\alpha}\|\beta'_\gamma - \beta_\gamma\|_2 \\
&\leq \frac{n^{b/2}}{d'_\gamma n^{b/2}}\|\beta'_\gamma - \beta_\gamma\|_2 \\
&\leq F_1\|\beta'_\gamma - \beta_\gamma\|_2,
\end{aligned}$$

if α, η grow with n, and we choose $\eta/\alpha > d'_\gamma n^{b/2}$ for some $d'_\gamma \in (0,\infty)$, where $J_{p_\gamma,1} = (1,1,\ldots,1)^\top$ and F_1 is a positive constant.

Second, since the density of Generalized Double Pareto prior maximizes at $\beta_\gamma = 0$, we have

$$\log p(\beta_\gamma|\gamma) - \log p(0|\gamma) \leq F_2,$$

for some constant $F_2 \in (0,\infty)$. □

Under the sufficient condition of Assumption 3.7, the Horseshoe prior

$$p(\beta_\gamma|\gamma,\tau) = \prod_{i=1}^{p_\gamma} K(\tau^2)^{-1/2}\exp(\frac{\beta_i^2}{2\tau^2})E_1(\frac{\beta_i^2}{2\tau^2}),$$

satisfies Assumption 3.7 and 3.8 if $\tau^2 \geq d'_\gamma n^c$ for some constant $d'_\gamma \in (0,\infty)$.

Proof. First, when n is sufficiently large, by the sufficient conditions of Assumption 3.7, we

have

$$
\begin{aligned}
|\log p(\beta'_\gamma|\gamma,\tau) - \log p(\beta_\gamma|\gamma,\tau)| &= |\log \frac{\prod_{j=1}^{p_\gamma} K(\tau^2)^{-1/2}\exp(\frac{\beta_j'^2}{2\tau^2})E_1(\frac{\beta_j'^2}{2\tau^2})}{\prod_{j=1}^{p_\gamma} K(\tau^2)^{-1/2}\exp(\frac{\beta_j^2}{2\tau^2})E_1(\frac{\beta_j^2}{2\tau^2})}| \\
&= |\sum_{j=1}^{p_\gamma}(\frac{\beta_j'^2-\beta_j^2}{2\tau^2} + \log\frac{E_1(\frac{\beta_j'^2}{2\tau^2})}{E_1(\frac{\beta_j^2}{2\tau^2})})| \\
&= |\sum_{j=1}^{p_\gamma}(\frac{\beta_j'^2-\beta_j^2}{2\tau^2} + o(1))| \\
&= \frac{1}{2\tau^2}|(\beta'_\gamma-\beta_\gamma)^\top(\beta'_\gamma+\beta_\gamma)| \\
&\le \frac{1}{2\tau^2}\|\beta'_\gamma+\beta_\gamma\|_2\|\beta'_\gamma-\beta_\gamma\|_2 \\
&\le \frac{1+d_\gamma n^c}{d'_\gamma n^c}\|\beta'_\gamma-\beta_\gamma\|_2 \\
&\le F_1\|\beta'_\gamma-\beta_\gamma\|_2,
\end{aligned}
$$

if $\tau^2 \ge d'_\gamma n^c$ for some constant $d'_\gamma \in (0,\infty)$, where F_1 is a positive constant.

Second, it follows Lemma B.3 that $p(\beta_\gamma|\gamma,\tau)\to\infty$ if $\beta_\gamma \to 0$. Thus, we have

$$
\begin{aligned}
\log p(\beta_\gamma|\gamma,\tau) - \log p(0|\gamma,\tau) &= \log\frac{p(\beta_\gamma|\gamma,\tau)}{p(0|\gamma,\tau)} \\
&= -\infty \\
&< F_2,
\end{aligned}
$$

for some constant $F_2 \in (0,\infty)$. □

B.10 Proof of Theorem 3.2

Proof of Theorem 3.2. Without the lost of generality, we assume $\sigma^2 = 1$, and we prove that, for every step within $Kn^{c+4c_{\min}}$ steps, a predictor belongs to the true model will be selected. Since the prove of each step is similar, we only prove the case starting from $\gamma^{(0)}$. Assume the predictor selected by the $k+1$ step is still an irrelevant predictor variable, where

$k < Kn^{c+4c_{\min}}$. We have the following relationship

$$\begin{aligned}\Omega(k) &= RSS(\gamma^{(k)}) - RSS(\gamma^{(k+1)}) \\ &= \|H^{(k)}_{a_{k+1}}\{I_n - H_{\gamma^{(k)}}\}y\|^2.\end{aligned}$$

Since we assume $a_{k+1} \notin \gamma_*$, we must have

$$\begin{aligned}\Omega(k) &\geq \max_{j\in\gamma_*}\|H^{(k)}_j\{I_n - H_{\gamma^{(k)}}\}y\|^2 \\ &\geq \|H^{(k)}_j\{I_n - H_{\gamma^{(k)}}\}y\|^2 \\ &\geq \|H^{(k)}_j\{I_n - H_{\gamma^{(k)}}\}X_{\gamma_*}\beta^0_{\gamma_*}\|^2 - \|H^{(k)}_j\{I_n - H_{\gamma^{(k)}}\}\epsilon\|^2 \\ &\geq \max_{j\in\gamma_*}\|H^{(k)}_j\{I_n - H_{\gamma^{(k)}}\}X_{\gamma_*}\beta^0_{\gamma_*}\|^2 - \max_{j\in\gamma_*}\|H^{(k)}_j\{I_n - H_{\gamma^{(k)}}\}\epsilon\|^2 \\ &= S_1 + S_2\end{aligned}$$

We first consider S_1, define $Q_{\gamma^{(k)}} = I_n - H_{\gamma^{(k)}}$. Then we have

$$\begin{aligned}&\max_{j\in\gamma_*}\|H^{(k)}_j\{I_n - H_{\gamma^{(k)}}\}X_{\gamma_*}\beta^0_{\gamma_*}\|^2 = \max_{j\in\gamma_*}\|H^{(k)}_j Q_{\gamma^{(k)}}X_{\gamma_*}\beta^0_{\gamma_*}\|^2 \\ &= \max_{j\in\gamma_*}\left\{\|X^{(k)}_j\|^{-2}|{X^{(k)}_j}^\top Q_{\gamma^{(k)}}X_{\gamma_*}\beta^0_{\gamma_*}|^2\right\} \\ &\geq \|X^{(k)}_{j_*}\|^{-2}|{X^{(k)}_{j_*}}^\top Q_{\gamma^{(k)}}X_{\gamma_*}\beta^0_{\gamma_*}|^2 \\ &\geq \min_{j\in\gamma_*}\left\{\|X^{(k)}_{j_*}\|^{-2}\right\}|{X^{(k)}_{j_*}}^\top Q_{\gamma^{(k)}}X_{\gamma_*}\beta^0_{\gamma_*}|^2 \\ &\geq \left\{\max_{j\in\gamma_*}\|X^{(k)}_{j_*}\|^2\right\}^{-1}\left\{\max_{j\in\gamma_*}|{X^{(k)}_{j_*}}^\top Q_{\gamma^{(k)}}X_{\gamma_*}\beta^0_{\gamma_*}|^2\right\} \qquad \text{(B.18)}\end{aligned}$$

Note that

$$\begin{aligned}
\|Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*}\|^2 &= {\beta^0_{\gamma_*}}^\top X_{\gamma_*}^\top Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*} \\
&= \sum_{j \in \gamma_*} \beta^0_j X_j^\top Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*} \\
&\leq \left(\sum_{j \in \gamma_*} \beta_j^2 \right)^{1/2} \left\{ \sum_{j \in \gamma_*} (X_j^\top Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*})^2 \right\}^{1/2} \\
&\leq \|\beta^0_{\gamma_*}\| \max_{j \in \gamma_*} |X_j^\top Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*}| \sqrt{p_{\gamma_*}} \qquad \text{(B.19)}
\end{aligned}$$

Apply (B.19) to (B.18), and note that $\max_{j \in \gamma_*} \|X_j\|^2/n \leq \lambda_{\max}$ with probability tending to one, by Condition 9 and 10. The we have

$$\max_{j \in \gamma_*} \|H_j^{(k)} Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*}\|^2 \geq \frac{\|Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*}\|^4}{n \lambda_{\max} p_{\gamma_*} \|\beta^0_{\gamma_*}\|^2} \qquad \text{(B.20)}$$

Defining $\xi_{\gamma^{(k)}} = \left(X_{\gamma^{(k)}}^\top X_{\gamma^{(k)}} \right)^{-1} X_{\gamma^{(k)}}^\top X_{\gamma_*} \beta^0_{\gamma_*}$, we obtain

$$\|Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*}\|^2 = \|X_{\gamma_*} \beta_{\gamma_*} - X_{\gamma^{(k)}} \xi_{\gamma^{(k)}}\|^2.$$

Under the assumption that no additional relevant predictor has been identified by the procedure. Then we have

$$\|Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*}\|^2 \geq n \lambda_{\min} \beta^2_{\min},$$

withe the probability tending to one. We find that

$$\begin{aligned}
\max_{j \in \gamma_*} \|H_j^{(k)} Q_{\gamma^{(k)}} X_{\gamma_*} \beta^0_{\gamma_*}\|^2 &\geq n \lambda_{\max}^{-1} p_{\gamma_*}^{-1} \|\beta^0_{\gamma_*}\|^{-2} \lambda^2_{\min} \beta^4_{\min} \\
&\geq \lambda_{\max}^{-1} \nu^{-1} C^{-2}_{\beta^0_{\gamma_*}} \lambda^2_{\min} \nu^4_{\beta^0_{\gamma_*}} n^{1-c_0-4c_{\min}}
\end{aligned}$$

Next, we consider S_2, we have

$$\begin{aligned}\|H_j^{(k)}\{I_n - H_{\gamma^{(k)}}\}\epsilon\|^2 &= \|X_j^{(k)}\|^{-2}(X_j^\top \epsilon - X_j^\top H_{\gamma^{(k)}}\epsilon)^2 \\ &\leq \lambda_{\min}^{-1} n^{-1}(X_j^\top Q_{\gamma^{(k)}}\epsilon)^2 \\ &\leq \lambda_{\min}^{-1} n^{-1} \max_{j\in\gamma_*} \max_{p_\gamma \leq m^*} (X_j^\top Q_\gamma \epsilon)^2\end{aligned}$$

where $m^* = K\nu n^{2c_0+4c_{\min}}$. Note that $X_j^\top Q_\gamma \epsilon$ is a normal random variable with mean 0 and variance $\|Q_\gamma X_j\|^2 \leq \|X_j\|^2$. Thus, we have

$$\lambda_{\min}^{-1} n^{-1} \max_{j\in\gamma_*} \max_{p_\gamma \leq m^*} (X_j^\top Q_\gamma \epsilon)^2 \leq \lambda_{\min}^{-1} n^{-1} \max_{j\in\gamma_*} \|X_j\|^2 \max_{j\in\gamma_*} \max_{p_\gamma \leq m_*} \chi_1^2$$

where χ_1^2 is a chi-square random variable with one degree of freedom. Then, we have

$$\begin{aligned}\max_{j\in\gamma_*} \max_{p_\gamma \leq m_*} \chi_1^2 &\leq 2m^* \log p_n \\ &\leq 3K\nu n^{2c+4c_{\min}} \nu n^{c_0} \\ &= 3K\nu^2 n^{c_0+2c+4c_{\min}}\end{aligned}$$

as $n \to \infty$.

Combing S_1 and S_2, we find

$$\begin{aligned}n^{-1}\Omega(k) &\geq \lambda_{\max}^{-1}\nu^{-1} C_{\beta^0_{\gamma_*}}^{-2} \lambda_{\min}^2 \nu_{\beta^0_{\gamma_*}}^4 n^{1-c_0-4c_{\min}} - \lambda_{\min}^{-1}\lambda_{\max} 3Kv^2 n^{c_0+2c+4c_{\min}-1} \\ &= \lambda_{\max}^{-1}\nu^{-1} c_{\beta^0_{\gamma_*}}^{-2} \lambda_{\min}^2 v_\gamma^4 n^{-c-4c_{\min}}\{1 - \lambda_{\max}^2 \nu^3 c_{\beta^0_{\gamma_*}}^2 \lambda_{\min}^{-3} \nu_{\beta^0_{\gamma_*}}^{-4} 3Kn^{c_0+3c+8c_{\min}-1}\}\end{aligned}$$

uniformly for every $k \leq Kn^{c+4c_{\min}}$. We then have

$$\begin{aligned}n^{-1}\|(I_n - H_{\gamma^{(k)}})y\|^2 &\geq n^{-1} \sum_{k=1}^{Kn^{c+4c_{\min}}} \Omega(k) \\ &\geq 2\{1 - \lambda_{\max}^2 \nu^3 c_{\beta^0_{\gamma_*}}^2 \lambda_{\min}^{-3} \nu_{\beta^0_{\gamma_*}}^{-4} 3Kn^{c_0+5c+8c_{\min}-1}\} \\ &\to 2\end{aligned}$$

as $n \to \infty$. On the other hand, under the assumption $\sigma^2 = 1$, we have $n^{-1}\|(I_n - H_{\gamma^{(k)}})y\|^2 \to 1$ in probability. Thus, it's impossible to have $\gamma^{(k)} \cup \gamma_* = \emptyset$ for every $1 \leq k \leq Kn^{c+4c_{\min}}$, which implies that at least one relevant variable will be discovered within $Kn^{c+4c_{\min}}$ steps. $\square$

Appendix C

Chapter 4 Preliminaries

C.1 Proof of Lemma 4.1

Proof of Lemma 4.1. For any $\gamma \in \mathcal{M}_1$, by Assumption 4.2, we have

$$-2log\frac{p(y|X\hat{\beta}_{\gamma_*})}{p(y|X\hat{\beta}_{\gamma})} \to \chi^2_{p_\gamma - p_{\gamma_*}},$$

as $n \to \infty$.

Let $r_\gamma = p_\gamma - p_{\gamma_*}$. Note that $\text{pr}\{\chi^2_r \geq 2x + 2(rx)^{1/2} + r\} \leq \exp(-x)$ by Laurent and Massart (2000). Then, as $n \to \infty$, we have

$$\begin{aligned}
&\text{pr}(\chi^2_{r_\gamma} \geq 2r_\gamma m_n + 2(m_n r_\gamma^2)^{1/2} + r_\gamma, \textit{some } 1 \leq r_\gamma < K_n - p_{\gamma_*}) \\
&\leq \sum_{j=1}^{K_n - p_{\gamma_*}} \binom{p_n}{j} \exp(-jm_n) \\
&\leq \sum_{j=1}^{K_n} p_n^j \exp(-jm_n) \\
&= \sum_{j=1}^{K_n} \exp\{-j(m_n - \log p_n)\} \\
&= \exp\{-(m_n - \log p_n)\} \\
&\times \frac{1 - \exp\{-K_n(m_n - \log p_n)\}}{1 - \exp\{-(m_n - \log p_n)\}},
\end{aligned}$$

which goes to 0 if $m_n = \log p_n + \delta_n$ with $\delta_n \to \infty$ as n increases.

Thus, for any $\gamma \in \mathcal{M}_1$, we have

$$\begin{aligned} -2\log\frac{p(y|\hat{\beta}_{\gamma_*})}{p(y|\hat{\beta}_{\gamma})} &< 2r_\gamma(\log p_n + \delta_n) + 2\sqrt{r_\gamma^2(\log p_n + \delta_n)} + r_\gamma \\ &\equiv r_\gamma \Lambda_n, \end{aligned} \tag{C.1}$$

in probability as $n \to \infty$, where $\Lambda_n = 2(\log p_n + \delta_n) + 2\sqrt{\log p_n + \delta_n} + 1$. Since the above result holds for an arbitrary δ_n, by Assumption 4.5, we let $\delta_n = o(\log n)$.

By Assumption 4.3 and Eq.(C.1), for any $\gamma \in \mathcal{M}_1$, as $n \to \infty$, we have

$$\begin{aligned} &-2\log\frac{p(y|\gamma_*)}{p(y|\gamma)} \\ &= -2\log\frac{p(y|\hat{\beta}_{\gamma_*})}{p(y|\hat{\beta}_{\gamma})} - 2\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} - r_\gamma \log n + p_{\gamma_*}c_{\gamma_*} - p_\gamma c_\gamma \\ &\le r_\gamma \Lambda_n - 2\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} - r_\gamma \log n - r_\gamma c_0 \\ &= r_\gamma \left\{ 2\log p_n + \delta_n - \frac{2}{r_\gamma}\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} - \log n - c_0 \right\} \\ &\le -r_\gamma c_0, \end{aligned}$$

when $\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} \ge r_\gamma \log\frac{p_n}{\sqrt{n}}$ for a constant c_0 which depends on γ.

Therefore, as $n \to \infty$, we have

$$\min_{\gamma \in \mathcal{M}_1} \log\frac{p(y|\gamma_*)}{p(y|\gamma)} \ge \frac{c_0}{2} r_\gamma.$$

The proof is completed. □

C.2 Proof of Lemma 4.2

Proof of Lemma 4.2. For any $\gamma \in \mathcal{M}_2$, by Assumption 4.3, we have

$$-2\log\frac{p(y|\gamma)}{p(y|\gamma_*)} = -2\log\frac{p(y|\hat{\beta}_\gamma)}{p(y|\hat{\beta}_{\gamma_*})} - 2\log\frac{p(\hat{\beta}_\gamma|\gamma)}{p(\hat{\beta}_{\gamma_*}|\gamma_*)} + r_\gamma \log n + p_\gamma c_\gamma - p_{\gamma_*} c_{\gamma_*},$$

as $n \to \infty$.

Let γ' be a model such that

$$\gamma' = arg\min_{\gamma \in \mathcal{M}_2} -2\log\frac{p(y|\hat{\beta}_\gamma)}{p(y|\hat{\beta}_{\gamma_*})}.$$

It follows that, for any $\gamma \in \mathcal{M}_2$, we have

$$\min_{\gamma \in \mathcal{M}_2} -2\log\frac{p(y|\hat{\beta}_\gamma)}{p(y|\hat{\beta}_{\gamma_*})} \geq -2\log\frac{p(y|\hat{\beta}_{\gamma'})}{p(y|\hat{\beta}_{\gamma_*})}.$$

Let $\gamma_{**} = \gamma' \cup \gamma_*$, we have

$$-2\log\frac{p(y|\hat{\beta}_{\gamma'})}{p(y|\hat{\beta}_{\gamma_*})} = -2\log\frac{p(y|\hat{\beta}_{\gamma'})}{p(y|\hat{\beta}_{\gamma_{**}})} - 2\log\frac{p(y|\hat{\beta}_{\gamma_{**}})}{p(y|\hat{\beta}_{\gamma_*})}.$$

By Assumption 4.4, we have

$$-2\log\frac{p(y|\hat{\beta}_{\gamma'})}{p(y|\hat{\beta}_{\gamma_{**}})} > a_{\gamma',\gamma_*} n,$$

in probability, where a_{γ',γ_*} is a positive constant depends on γ' and γ_*.

Recall Eq. (C.1) in the proof of Lemma 1, we have

$$\begin{aligned} -2\log\frac{p(y|\hat{\beta}_{\gamma_{**}})}{p(y|\hat{\beta}_{\gamma_*})} &> -r_{\gamma_{**}}\Lambda_n \\ &\geq -p_{\gamma'}\Lambda_n, \end{aligned}$$

where $r_{\gamma_{**}} = p_{\gamma_{**}} - p_{\gamma_*} \leq p_{\gamma'}$ and $\Lambda_n = 2(\log p_n + \delta_n) + 2\sqrt{\log p_n + \delta_n} + 1$ with $\delta_n \to \infty$, as

$n \to \infty$.

Thus, for γ', we have

$$-2\log\frac{p(y|\hat{\beta}_{\gamma'})}{p(y|\hat{\beta}_{\gamma_*})} \quad > \quad a_{\gamma',\gamma_*}n - p_{\gamma'}\Lambda_n$$

For any $\gamma \in \mathcal{M}_2$, under Assumption 4.1, 4.3, 4.5 and 4.6, it follows that

$$\begin{aligned}
-2\log\frac{p(y|\gamma)}{p(y|\gamma_*)} &= -2\log\frac{p(y|\hat{\beta}_{\gamma})}{p(y|\hat{\beta}_{\gamma_*})} - 2\log\frac{p(\hat{\beta}_{\gamma}|\gamma)}{p(\hat{\beta}_{\gamma_*}|\gamma_*)} + r_\gamma \log n + p_\gamma c_\gamma - p_{\gamma_*}c_{\gamma_*} \\
&> a_{\gamma',\gamma_*}n - p_{\gamma'}\Lambda_n - 2\log\frac{p(\hat{\beta}_{\gamma}|\gamma)}{p(\hat{\beta}_{\gamma_*}|\gamma_*)} + r_\gamma \log n + p_\gamma c_\gamma - p_{\gamma_*}c_{\gamma_*} \\
&= a_{\gamma',\gamma_*}n\{1 + o(1)\},
\end{aligned}$$

if $\log\frac{p(\hat{\beta}_{\gamma_*}|\gamma_*)}{p(\hat{\beta}_{\gamma}|\gamma)} = o(n)$, as $n \to \infty$.

Thus, we have $\min_{\gamma\in\mathcal{M}_2} \log\frac{p(y|\gamma_*)}{p(y|\gamma)} \geq a_{\gamma',\gamma_*}n\{1+o(1)\}$ as $n \to \infty$. The proof is completed. □

C.3 Lemma C.1

Lemma C.1. *For model $\gamma \in \mathcal{M}_n$, the univariate Horseshoe density $p(\beta_i|\tau)$ satisfies the following: (a) $\lim_{\beta_i\to 0} p(\beta_i|\tau) = \infty$. (b) For $\beta_i \neq 0$, we have*

$$\frac{K_0}{2}(\tau^2)^{-1/2}\log\left(1+\frac{4\tau^2}{\beta_i^2}\right) < p(\beta_i|\tau) < K_0(\tau^2)^{-1/2}\log\left(1+\frac{2\tau^2}{\beta_i^2}\right)$$

where $K_0 = 1/(2\pi^3)^{1/2}$ and $i = 1, \ldots, p_\gamma$.

Proof of Lemma C.1. First, for model γ, we have

$$p(\beta_i|\tau) = \int_0^\infty (2\pi\lambda_i^2)^{-1/2}\exp\left(-\beta_i^2/2\lambda_i^2\right)\frac{2}{\pi\tau[1+(\lambda_i/\tau)^2]}d\lambda_i$$

Let $u = (\tau/\lambda_i)^2$, then we have

$$\begin{aligned} p(\beta_i|\tau) &= \int_0^\infty (2\pi\tau^2)^{-1/2} u^{1/2} \exp\left(-\frac{\beta_i^2}{2\tau^2}u\right) \frac{u^{-3/2}\tau}{\pi\tau(1+1/u)} du \\ &= \frac{K_0}{(\tau^2)^{1/2}} \int_0^\infty \frac{\exp\left(-\frac{\beta_i^2}{2\tau^2}u\right)}{1+u} du \end{aligned}$$

where $K_0 = 1/(2\pi^3)^{1/2}$. Let $z = u + 1$, then we have

$$\begin{aligned} p(\beta_i|\tau) &= \frac{K_0}{(\tau^2)^{1/2}} \int_1^\infty \exp\left[-\frac{\beta_i^2}{2\tau^2}(z-1)\right] \frac{1}{z} dz \\ &= \frac{K_0}{(\tau^2)^{1/2}} \exp(\frac{\beta_i^2}{2\tau^2}) \int_1^\infty \frac{1}{z} \exp\left[-\frac{\beta_i^2}{2\tau^2}z\right] dz \\ &= \frac{K_0}{(\tau^2)^{1/2}} \exp(\frac{\beta_i^2}{2\tau^2}) E_1(\frac{\beta_i^2}{2\tau^2}) \end{aligned}$$

where $E_1(\cdot)$ is the exponential integral function. This function satisfies tight upper and lower bounds:

$$\frac{\exp(-t)}{2} \log\left(1 + \frac{2}{t}\right) < E_1(t) < \exp(-t) \log\left(1 + \frac{1}{t}\right)$$

for all $t > 0$, thus we have

$$\frac{K_0}{2}(\tau^2)^{-1/2} \log\left(1 + \frac{4\tau^2}{\beta_i^2}\right) < p(\beta_i|\tau) < K_0(\tau^2)^{-1/2} \log\left(1 + \frac{2\tau^2}{\beta_i^2}\right)$$

which proves Part (b) and Part (a) then follows from the lower bound approaches ∞ as $\beta_i \to 0$.

□