PRIORS FOR BAYESIAN SHRINKAGE AND HIGH-DIMENSIONAL MODEL SELECTION

A Dissertation

by

MINSUK SHIN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Valen E. Johnson
Co-Chair of Committee	Anirban Bhattacharya
Committee Members,	Jianhua Huang
	Byung-Jun Yoon
Head of Department,	Valen E. Johnson

August 2017

Major Subject: Statistics

ABSTRACT

This dissertation focuses on the choice of priors in Bayesian model selection and their applied, theoretical and computational aspects. As George Box famously said “all models are wrong, but some are useful"; many statisticians and scientists are aware of the importance of model selection. In a Bayesian perspective, however, it is challenging to choose the prior on the parameters involved in model selection or how to evaluate the criterion on the prior, especially when the number of models to be compared is massive or when a nonparametric model is considered.

For high-dimensional Bayesian model selection for linear models, my dissertation studies theoretical perspectives of the choice of the prior on the regression coefficient. Especially, I consider the nonlocal prior densities that assign zero density around the null value, which is typically 0 in model selection settings. When certain regularity conditions apply, I demonstrate that the model selection procedure based on the nonlocal priors is consistent for linear models even when the number of covariates p increases sub-exponentially with the sample size n. I investigate the asymptotic form of the marginal likelihood based on the nonlocal priors and show that it attains a unique penalty term that adapts to the strength of signal corresponding variable in the model, and I remark that this term cannot be attained from local priors such as Gaussian prior densities.

Another topic of my dissertation is about computational aspects of Bayesian model selection under high-dimensional settings. A full posterior sampling using existing Markov chain Monte Carlo (MCMC) algorithms to explore high-dimensional model space is highly inefficient and often not feasible from a practical perspective. To overcome this issue, I propose a scalable stochastic search algorithm called Simplified Shotgun Stochastic Search with Screening (S5), which efficiently explores the model space. The S5 algorithm dramatically reduces

the computational burden to search the neighborhood of a model by considering a screening step within the algorithm. Its empirical performance is examined in several examples, and it outperforms existing algorithms in the sense that S5 is computationally fast while it efficiently searches the model space. S5 is used to implement the model selection procedures introduced in this dissertation, including linear and nonparametric model selection. The computing functions are provided in the `R` package `BayesS5` in `CRAN` (`https://cran.r-project.org`).

For nonparametric regression models, I introduce a new shrinkage prior on function spaces, the functional horseshoe prior, that encourages shrinkage towards parametric classes of functions. When the true underlying function is in the parametric class, improved estimation performance is obtained relative to classical nonparametric procedures. The proposed prior also provides a natural penalization interpretation, and casts light on a new class of penalized likelihood methods for function estimation. I theoretically exhibit the efficacy of the proposed approach by showing an adaptive posterior concentration property.

The last topic of the dissertation is about a novel extension of the nonlocal idea to functional spaces, called the nonlocal functional prior, which is suitable for nonparametric Bayesian hypothesis testing (model selection) problems. I illustrate the asymptotic rate of the Bayes factor defined by the proposed prior for nonparametric hypothesis testing problems. I apply the proposed prior densities for high-dimensional model selection of nonparametric additive models, and investigate the model selection consistency of the resulting model selection procedure. I provide some simulation studies and real data examples that show that the proposed model selection procedure outperforms state-of-the-art methods in finite samples.

DEDICATION

To my lovely wife Mirae, my adorable daughter Jane and son Sungmin (Daniel).

ACKNOWLEDGMENTS

First of all, I would like to express my deep gratitude to my advisors Valen E. Johnson and Anirban Bhattacharya. Because of their thoughtful advice, I was able to complete my dissertation, and they helped me a lot to grow academically. Val taught me the right attitude and passion towards science and he has alway given me a piece of warm hearted advice. Anirban is not only a good friend, but he is also a good partner to discuss ideas. Val and Anirban, I have been incredibly inspired by you, and no word of thanks is enough for your wonderful mentorship.

I would also like to extend my gratitude to the members of my dissertation committee, Dr. Jianhua Huang and Dr. Byung-Jun Yoon for their support and encouragement. Thanks also to Dr. Irina Gaynanova and Dr. Xianyang Zhang for giving me helpful advice to apply for academic jobs and to improve my presentation skills. Dr. David Rossell has been extremely encouraging and I have thoroughly enjoyed the discussions with him. It has been a great pleasure working with Dr. Naveen N. Narisetty, even though our work has been delayed so long. I believe that we can finish the project soon. Amir Nikooienejad is also a good friend of mine, and I really appreciate his constant encouragement. I am also grateful to my friend Sangyoon Yi for buying me Starbucks coffee many times.

Dr. Ersen Arseven, who has worked in the clinical trial industry for 30 years, has been a good mentor. He gave me a lot of supports and warm-hearted advice. In particular, his advice for presentation and job interview was really useful and practical.

Finally, I devote my dissertation to my family: my beloved wife Mirae, my daughter Jane, my son Sungmin (Daniel), my parents and my parents-in-law. Especially, my mother Eulsun Kim and my mother-in-law Soonja Lee crossed the Pacific Ocean from South Korea and came to America to help us when my daughter and son were newborn. I would also like to thank my

grandmother and grandfather for their dedicated support. Without supports from my family, none of this work would have been possible. I love you!

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Valen E. Johnson (co-advisor), Dr. Anirban Bhattacharya (co-advisor) and Dr. Jianhua Huang of the Department of Statistics and Dr. Byung-Jun Yoon of the Department of Electrical Engineering.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a teaching assistantship from Texas A&M University and from the National Institute of Health (NIH) R01 CA.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 A Brief Review of Bayesian Model Selection

Suppose that a set of H models $\mathcal{M} = \{M_1, \ldots, M_H\}$ is considered for observed data $\mathbf{y}$. Under a model $M_h \in \mathcal{M}$, the density function of $\mathbf{y}$ is $L(\mathbf{y} \mid \theta_h, M_h)$, where θ_h is a vector of unknown parameters under model M_h. For Bayesian inference, priors should be fully specified by assigning a prior distribution $\pi(\theta_m \mid M_h)$ to the parameters of each model and a model prior $\pi(M_h)$ to each model. The posterior probability of model M_h conditionally on the observed $\mathbf{y}$ can be expressed as

$$\pi(M_h \mid \mathbf{y}) = \frac{m_{M_h}(\mathbf{y})\pi(M_h)}{\sum_{h'} m_{M_{h'}}(\mathbf{y})\pi(M_{h'})}, \tag{1.1}$$

where

$$m_{M_h}(\mathbf{y}) = \int L(\mathbf{y} \mid \theta_h, M_h)\pi(\theta_h \mid M_h)d\theta_h$$

is the marginal likelihood of a model M_h. Based on these posterior model probabilities, pairwise comparison of models M_1 and M_2 is conducted by the posterior odds that can be expressed as a product between the ratio of marginal likelihoods and the model prior odds; i.e.,

$$\frac{\pi(M_1 \mid \mathbf{y})}{\pi(M_2 \mid \mathbf{y})} = \frac{m_{M_1}(\mathbf{y})}{m_{M_2}(\mathbf{y})} \times \frac{\pi(M_1)}{\pi(M_2)}.$$

In particular, the ratio of marginal likelihoods $m_{M_1}(\mathbf{y})/m_{M_2}(\mathbf{y})$ is called the *Bayes factor* and it determines the decision rules for Bayesian hypothesis testing problems as discussed in Kass and Raftery (1995) and Jeffreys (1961). The higher the Bayes factor value supports, the more evidence in favor of M_1. In Kass and Raftery (1995), a rough descriptive statement of

some decision rules regarding Bayes factors was empirically addressed as

$\log BF_{10}$	Evidence against H_0
0 to 1	Not worth more than a bare mention
1 to 3	Positive
3 to 5	Strong
> 5	Very strong.

More discussions and empirical examples regarding Bayes factor are provided in Kass and Raftery (1995).

Throughout this dissertation, I assume that one of considered models is the true model that represents the data-generating process, which is a setting called the “M-closed" framework as proposed in Bernardo and Smith (1994). This in itself is somewhat controversial, because the true model might not exist or it might not be one of those under consideration. However, it is a helpful viewpoint for at least thinking through the consequences of a true Bayesian model selection procedure and desirable qualities.

I now introduce a desirable asymptotic property for Bayesian model selection procedures called *model selection consistency* that can be defined as follows.

Definition 1. *(Model Selection Consistency) Suppose that* $\mathbf{t}$ *is the true data-generating model. Then, if*

$$\pi(\mathbf{t} \mid \mathbf{y}) \xrightarrow{p} 1,$$

as $n \to \infty$*, the Bayesian model selection procedure is called “consistent".*

From a theoretical point of view, when the number of models is fixed regardless of the sample size n, Schwarz (1978) showed that the model selection procedures defined by some general classes of priors; e.g. Gaussian priors achieve the model selection consistency. How-

ever, when I allow the number of models to increase at a certain rate of n, the asymptotic behavior of the posterior probability of the true model is not clear. This situation is commonly faced in high-dimensional variable selection problems for regression models due to the fact that the total number of models for variable selection is 2^p, where $p(\gg n)$ is the total number of variables .

For the Bayesian framework, the uncertainty of the model space can be represented by the posterior model distribution $\pi(M_1 \mid \mathbf{y}), \ldots, \pi(M_H \mid \mathbf{y})$. By considering $\pi(M_h \mid \mathbf{y})$ as a measure of the "truth" of model M_h, a natural strategy for model selection is to choose the model that attains the largest posterior model probability. This model is called the *maximum a posteriori* (MAP) model, i.e. $\widehat{M}^{MAP} = \text{argmax}_h \, \pi(M_h \mid \mathbf{y})$. These posterior probabilities are also important for full posterior inference in prediction using *Bayesian model averaging* (Raftery et al., 1997), which is quantified by the posterior predictive distribution as $p(y^{pred} \mid \mathbf{y}) = \sum_{h'} \pi(y^{pred} \mid M_{h'}, \mathbf{y}) \pi(M_{h'} \mid \mathbf{y})$ for a future observation y^{pred}.

1.1.1 Bayesian Model Selection for the Linear Regression Model

Consider the standard setup of a Gaussian linear regression model with a univariate response and p candidate predictors. Let $\mathbf{y} = \{y_1, \ldots, y_n\}^{\mathrm{T}}$ denote a vector of responses for n individuals and X an $n \times p$ matrix of covariates. Let $\beta = \{\beta_1, \ldots, \beta_p\}^{\mathrm{T}}$ denote the regression coefficients. The linear regression model for the data is given by

$$\mathbf{y} = X\beta + \epsilon, \tag{1.2}$$

where $\epsilon \sim N_n(0, \sigma^2 \mathrm{I}_n)$. However, in high-dimensional settings ($n \ll p$), the unique MLE does not exist and a MLE fails to achieve consistency of estimation. To overcome this issue, from a Bayesian perspective, one can consider a sparsity inducing prior (Castillo et al., 2015) that restricts the size of a given model $\mathbf{k}$ and puts zero prior probability on other parameters that are

not in model $\mathbf{k}$. More precisely, for a given model $\mathbf{k}$, the prior is

$$\pi(\beta \mid \mathbf{k}) \propto \pi_{\mathbf{k}}(\beta_{\mathbf{k}})\delta_0(\beta_{\mathbf{k}^c}) \tag{1.3}$$

where the term $\delta_0(\beta_{\mathbf{k}^c})$ implies the coordinates $\beta_{\mathbf{k}^c} = \{\beta_j : j \in \mathbf{k}^c\}$ being zero and $\pi_{\mathbf{k}}(\beta_{\mathbf{k}})$ is a prior on the nonzero regression coefficients $\beta_{\mathbf{k}} = \{\beta_j : j \in \mathbf{k}\}$. This class of priors includes many instances such as Zellner's g-prior (Zellner, 1986), mixtures of g-priors (Liang et al., 2008) and discrete mixtures of spike and slab priors (Ishwaran and Rao, 2005). With a slight abuse of the notation, I denote the prior on the model space as $\pi(\mathbf{k})$ for a model $\mathbf{k}$. By following the definition of the posterior model probability in (1.1), the resulting posterior probability of model $\mathbf{k}$ is defined as

$$\pi(\mathbf{k} \mid \mathbf{y}) = \frac{m_{\mathbf{k}}(\mathbf{y})\pi(\mathbf{k})}{\sum_{\mathbf{l}} m_{\mathbf{l}}(\mathbf{y})\pi(\mathbf{l})},$$

where the marginal likelihood of a model $\mathbf{k}$ is given by

$$m_{\mathbf{k}}(\mathbf{y}) = \int L(\mathbf{y} \mid \beta, \mathbf{k})\pi(\beta \mid \mathbf{k})d\beta,$$

for the likelihood function for model $\mathbf{k}$, $L(\mathbf{y} \mid \beta, \mathbf{k})$. More practically, when σ^2 is unknown, a prior on σ^2 can be deployed and the corresponding marginal likelihood can be defined by integrating with respect to the prior on σ^2. The posterior model probability can be used to select variables that are associated with the response. The simplest approach to the best model is to consider the MAP model that maximizes the posterior model probability. An other option is to utilize the marginal inclusion probabilities $\{q_j : j = 1, \ldots, p\}$, where $q_j = \sum_{\mathbf{k}: j \in \mathbf{k}} \pi(\mathbf{k} \mid \mathbf{y})$ for $j = 1, \ldots, p$. The median probability model is defined as the set of variables whose marginal inclusion probability is larger than 0.5. Barbieri and Berger (2004) showed that the median probability model is optimal in a predictive sense when only a single model is considered .

Some desirable theoretical properties of the posterior inference with some choices of the prior specification (1.3) have been discussed in high-dimensional settings. Johnson and Rossell (2010) proposed a class of prior densities called *nonlocal prior densities* for $\pi_{\mathbf{k}}(\beta_{\mathbf{k}} \mid \mathbf{k})$. Nonlocal prior densities are density functions that are identically zero whenever a model parameter is equal to its null value, which is typically 0 in model selection settings. More formal definition is as follows:

Definition 2. *Suppose that θ is a parameter supported in Θ and θ_0 is the null value. Consider a hypothesis test $H_0 : \theta = \theta_0 \quad versus \quad H_1 : \theta \neq \theta_0$. Under the alternative hypothesis, a prior density π is nonlocal, if for every $\epsilon > 0$, there is $\delta > 0$ such that $\pi(\theta) < \epsilon$ for all $\theta \in \Theta$ such that $|\theta - \theta_0| < \delta$.*

Conversely, local prior densities are positive at the null parameter value. In Johnson and Rossell (2012), it was shown that Bayesian model selection procedures based on nonlocal priors achieve model selection consistency when $p = O(n)$. However, when p increases at a sub-exponential rate of n, i.e. $\log p = O(n^c)$ for some $0 \leq c < 1$, its posterior model consistency has not been derived.

Also, under increasing p at a sub-exponential rate of n, Narisetty and He (2014) investigated the asymptotic behavior of model selection procedure based on a Gaussian prior with diverging variance as n grows. Castillo et al. (2015) discussed some general conditions on priors on the coefficient and models for the optimal rate of posterior contraction and model selection consistency. All priors considered in the literature were local priors.

1.1.2 Bayesian Model Selection in the Nonparametric Regression

Consider a simple nonparametric regression model defined according to a response $\mathbf{y} = \{y_1, \ldots, y_n\}$ and a univariate predictor $X = \{x_1 \ldots, x_n\}$

$$\mathbf{y} = F + \epsilon, \tag{1.4}$$

where $\epsilon \sim N(0, \sigma^2 \mathrm{I}_n)$ and $F = \{f(x_1), \ldots, f(x_n)\}$ with the unknown regression function f. From a practical point of view, a practitioner should decide whether the shape of f is specified by a parametric from such as linear or quadratic function, or a nonparametric representation using splines, wavelet, Gaussian processes etc. This argument can be formalized as a model selection problem by writing

$$H_0 : F \in \mathfrak{L} \qquad versus \qquad H_1 : F \notin \mathfrak{L}, \tag{1.5}$$

where $\mathfrak{L}$ is a class of the parametric functions that are specified in advance. For example, if a practitioner is interested in whether F is linear or not, $\mathfrak{L}$ can be defined as $\mathfrak{L} = \{\beta_0 + \beta_1 X : \beta_0,\ \beta_1 \in \mathbb{R}\}$. Under H_0, the resulting model is simply a univariate linear regression model. Under H_1, one can model the unknown function f as spanned by a set of pre-specified basis functions $\{\phi_j\}_{1 \leq j \leq K_n}$, where K_n is the number of basis functions, as follows:

$$f(x) = \sum_{k=1}^{K_n} \beta_k \phi_k(x). \tag{1.6}$$

I shall work with the B-spline basis (De Boor, 1978) in the sequel, although the methodology generalizes to a larger class of basis functions. The B-splines basis functions can be constructed in a recursive way. Let the positive integer q denote the degree of the B-spline basis functions satisfying $K_n > q + 1$. Without loss of generality, assume that $x_i \in [0, 1]$ for $i = 1, \ldots, n$. Define a sequence of knots $0 = t_0 < t_1 < \cdots < t_{K_n - q} = 1$. In addition, define q knots $t_{-q} = \cdots = t_{-1} = t_0$ and another set of q knots $t_{K_n - q} = \cdots = t_{K_n}$. As in De Boor

(1978), the B-spline basis functions are defined as

$$\begin{aligned}\phi_{k,1}(x) &= \begin{cases} 1, & t_k \le x < t_{k+1}, \\ 0, & \text{otherwise}, \end{cases} \\ \phi_{k,q+1}(x) &= \frac{x - t_k}{t_{k+q} - t_k}\phi_{k,q}(x) + \frac{t_{k+q+1} - x}{t_{k+q+1} - t_{k+1}}\phi_{k+1,q}(x),\end{aligned}$$

for $k = -q, \ldots, K_n - q - 1$. I reindex $k = -q, \ldots, K_n - q - 1$ to $k = 1, \ldots, K_n$ and the number of basis functions is K_n. Letting $\beta = (\beta_1, \ldots, \beta_{K_n})^{\mathrm{T}}$ denote the vector of basis coefficients and $\Phi = \{\phi_k(x_i)\}_{1 \le i \le n, 1 \le k \le K_n}$ denote the $n \times K_n$ matrix of basis functions evaluated at the observed covariates, I model $F = \Phi\beta$.

Even though parametric models might fail to capture important features of the data when they do not fit into the parametric form, the asymptotic behavior of the parametric model is superior to the nonparametric counterparts when the data-generating model is in the class of the parametric models or it is close enough to the class. In Ghosal and van der Vaart (2007), it was shown that the contraction rate of the posterior distribution defined by isotropic Gaussian priors for the nonparametric regression model in (1.4) is $n^{-\alpha/(1+2\alpha)}$, where $\alpha > 0$ quantifies the smoothness of the function, which is slower than the parametric optimal rate $n^{-1/2}$. In practice, nonparametric models also require extra steps to choose the tuning parameter that controls the smoothness of the estimated function, and they are usually challenging in computational and practical senses. Furthermore, in many cases, parametric shapes of F have advantages for interpretation of the regression function. For example, the slope parameter of the linear regression model represents the linear association between the response and the covariate.

In the Bayesian paradigm, the evidence in favor of each model in (1.5) is naturally quantified by Bayes factor that was introduced in Jeffreys (1961) and defined as

$$BF_{10} = \frac{m_1(\mathbf{y})}{m_0(\mathbf{y})},$$

where $m_1(\mathbf{y})$ and $m_0(\mathbf{y})$ are the marginal likelihoods under H_1 and H_0; i.e. $m_1(\mathbf{y}) = \int L(\mathbf{y} \mid \beta, H_1) d\pi_{NP}(\beta)$ and $m_0(\mathbf{y}) = \int L(\mathbf{y} \mid \theta, H_0) d\pi_P(\theta)$, where π_{NP} is a prior on the B-spline coefficient β and π_P is a prior on the parameter $\theta \in \mathbb{R}^{d_0}$ for the parametric model in $\mathfrak{L}$.

For the hypothesis test in (1.5), Choi et al. (2009) considered a semiparametric model that has an additive form between a parametric function and a nonparametric function. They investigated the asymptotic behavior of the Bayes factor defined by Gaussian priors on the coefficients of the basis functions. More general theoretical results regarding Bayes factor were provided in Choi and Rousseau (2015), which showed that the resulting Bayes factor achieves consistency in the sense that BF_{10} converges to zero in probability when the true data-generating process supports H_0 and BF_{10} diverges to infinity in probability, otherwise.

When multiple predictors are considered, the nonparametric additive model (Hastie and Tibshirani, 1986) can be considered, which is expressible as

$$\mathbf{y} = \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{1.7}$$

where $\epsilon \sim N(0, \sigma^2 \mathrm{I}_n)$ and f_j is the j-th marginal regression function. Also, X_j is the j-th covariate among p covariates. The setting for (1.4) can be naturally extended to the additive model by modeling each component function as a linear combination of the B-spline basis functions, i.e. $f_j(X_j) = \sum_{k=1}^{K_n} \phi_k(X_j)\beta_{jk} = \Phi_j \beta_j$, where $\beta_j = \{\beta_{j1}, \ldots, \beta_{jK_n}\}$ and $\Phi_j = \{\phi_1(X_j), \ldots, \phi_{K_n}(X_j)\}$ for $1 \leq j \leq p$. Similar to the model selection procedures discussed in Section 1.1.1 for linear models, I am interested in selecting variables that are associated with the response $\mathbf{y}$, and the uncertainty identification of the model space is also my concern.

From a frequentist perspective, there have been several studies regarding the additive model selection in high-dimensional settings, including Ravikumar et al. (2009), Meier et al. (2009), and Huang et al. (2010). Many procedures use the group Lasso penalty proposed in Yuan and Lin (2006) to induce the sparse representation of the component function. Theoretical proper-

ties of associated estimation and model selection properties have been investigated in Raskutti et al. (2012) and Yuan and Zhou (2016). In Bayesian frameworks, Shang and Li (2014) considered the Bayesian additive model in high-dimensional settings and provided some conditions on the prior on the basis coefficient necessary to achieve model selection consistency. However, in Shang and Li (2014), the practical guideline regarding the choice of prior on the spline coefficients is unclear, and the computational challenges are not resolved, since the proposed algorithm is based on an MCMC algorithm that is inefficient in high-dimensional settings.

1.2 Research Challenges and Main Contributions

1.2.1 Linear Model Selection in High-dimensional Settings

The Choice of Priors

For high-dimensional linear model selection problems, there is a rich literature regarding the choice of prior on the regression coefficient and the model space. In Castillo et al. (2015), a class of model priors called *complexity priors* was defined as

$$\pi(\mathbf{k}) \propto \binom{p}{|\mathbf{k}|}^{-1} a^{-|\mathbf{k}|} p^{-b|\mathbf{k}|},$$

for some constants $a, b > 0$. I note that the Bernoulli-uniform prior discussed in Scott and Berger (2010), $\pi(\mathbf{k}) \propto \binom{p}{|\mathbf{k}|}^{-1}$, is a special case of the complexity prior with $a = 1$ and $b = 0$. Castillo et al. (2015) provided tail conditions on the prior on the coefficients and sufficient conditions on the hyperparameter of a class of model priors to guarantee the optimal posterior concentration rate and model selection ocnsistency. Narisetty and He (2014) investigated the asymptotic behavior of model selection procedures based on the Bernoulli-uniform model prior and a Gaussian prior with variance that increases faster than $p^{2+\epsilon}$ for any small $\epsilon > 0$.

For linear models, the posterior model probability based on priors discussed in Castillo et al. (2015) and Narisetty and He (2014) (or Zellner's g-prior Zellner (1986)) can be asymptotically

expressed as

$$\log \pi(\mathbf{k} \mid \mathbf{y}) \approx l_{\mathbf{k}}(\widehat{\beta}_{\mathbf{k}}) - |\mathbf{k}| c_{n,p} + C, \tag{1.8}$$

where $l_{\mathbf{k}}$ is the logarithm of the likelihood function under a model $\mathbf{k}$ and $\widehat{\beta}_{\mathbf{k}}$ is the maximum likelihood estimator of $\beta_{\mathbf{k}}$ under model $\mathbf{k}$ for some sequence $c_{n,p} > 0$ and a constant C. For example, as shown in Narisetty and He (2014), if $\beta_{\mathbf{k}} \mid \mathbf{k} \sim N(0, p^c)$ for some constant $c > 0$ and the Bernoulli-uniform prior is imposed on the model space, then the logarithm of the resulting posterior model probability is asymptotically equivalent to $l_{\mathbf{k}}(\widehat{\beta}_{\mathbf{k}}) - (1 + c/2)|\mathbf{k}| \log p + C$. It is interesting to note that this expression is exactly the same as penalized likelihood procedures with a L_0 penalty (e.g., Zhang et al. (2010), Chen and Chen (2008), Kim et al. (2012)).

The main property of the form in (1.8) is that the penalty strength on model $\mathbf{k}$ is determined solely by its size $|\mathbf{k}|$, regardless of the marginal strength of the regression coefficient in the model. For example, suppose that two different models with the same model size are considered. One model consists of predictors that are strongly associated with the response, and the other model contains only some of strongly associated variables and the rest of variables in the model are weakly associated with the response. Under objective function in (1.8), two models would be penalized by the same amount, because the model size is the same. Even though the model with weakly associated variables will be strongly penalized by the log-likelihood function, the model selection criterion with the penalty that only depends on the model size might not be able to select important variables and might fail to control the multiplicity in a practical sense since there are too many models to be compared in the model space in high-dimensional settings.

In this dissertation, certain sufficient conditions on the nonlocal priors defined in Definition 2 will be provided to allow the resulting model selection procedure to achieve the model selection consistency when $\log p = O(n^c)$ for some $0 \leq c < 1$. Also, the asymptotic form

of the posterior model probability defined by the nonlocal priors will be discussed, and I will point out that the asymptotic form of the posterior model probability contains a unique form of penalty on the regression coefficient. The form of penalty cannot be attained by local priors that have been used previously in the literature.

The asymptotic form of the logarithm of the posterior model probability defined by the nonlocal prior on the regression coefficients can be expressed as

$$\log \pi(\mathbf{k} \mid \mathbf{y}) \approx l_{\mathbf{k}}(\widehat{\beta}_{\mathbf{k}}) - \sum_{j=1}^{|\mathbf{k}|} \frac{\tau}{\widetilde{\beta}^2_{\mathbf{k},j}} + \log \pi(\mathbf{k}) + C', \tag{1.9}$$

where $\widetilde{\beta}_{\mathbf{k},j} = \widehat{\beta}_{\mathbf{k},j} + O_p\big((n/\tau)^{-1/4}\big)$ and τ is the hyperparameter for the nonlocal prior, which controls the parsimony of model selection. Also, $\widehat{\beta}_{\mathbf{k},j}$ denotes the j-th element of $\widehat{\beta}_{\mathbf{k}}$.

While model selection procedures defined by local priors penalize a model only by the size of the model, the penalty $\sum_{j=1}^{|\mathbf{k}|} \tau/\widetilde{\beta}^2_{\mathbf{k},j}$ in the objective function in (1.9) is adaptively determined by the strength of the marginal signal that is measured by the $\widetilde{\beta}^2_{\mathbf{k},j}$ term and imposes different penalties on each predictor in the given model. This adaptive term encourages the model selection procedure to select variables with strong signals. This property has not been previously discussed in the original literature (Johnson and Rossell, 2010), and it explains why the nonlocal prior shows empirically outstanding performance in model selection.

A Scalable Computation

Even though sparsity inducing priors in (1.3) enjoy desirable theoretical properties, the practical implementation of Bayesian model selection procedures based on these priors is computationally challenging due to the discrete nature of the prior. Since the total number of possible models is enormous (2^p) even for a moderate dimension p, it is not computationally practical to calculate all possible marginal likelihoods to evaluate the exact posterior model probabilities. Thus, algorithms to efficiently explore the model space to reduce the computa-

tional burden are needed. One might consider reversible jump Markov chain Monte Carlo proposed in Green (1995) for posterior inference, but that algorithm is inefficient and impractical, especially in high-dimensional settings. A Gibbs sampling based algorithm called the Stochastic Search Variable Selection (SSVS) was proposed in George and McCulloch (1993), but its computational efficiency decreases as the number of predictors increases. Besides Markov Chain Monte Carlo (MCMC) approaches, Hans et al. (2007) introduced Shotgun Stochastic Search (SSS) to efficiently search the model space and approximate posterior model probabilities. However, the computational demands of SSS significantly increases as dimension grows. More recently, some deterministic approaches, such as Rockova and George (2014) and Carbonetto and Stephens (2012), were used to find the MAP model. Those algorithms only find a single model and do not provide posterior model probabilities, so it is challenging to quantify the uncertainty on the model space.

To ameliorate these computational issues, several continuous shrinkage priors have been proposed, including the Bayesian Lasso (Hans, 2009; Park and Casella, 2008), the horseshoe prior (Carvalho et al., 2010), the generalized double Pareto shrinkage prior (Armagan et al., 2013) and the Dirichlet-Laplace prior (Bhattacharya et al., 2015). Those priors can be expressed as scale mixtures of Gaussian distributions, so the resulting marginal priors are continuous. By avoiding the structure of the discrete mixtures, those continuous shrinkage priors provide a computational advantage, and efficient MCMC algorithms are available for sampling from the corresponding posterior distribution; e.g. Bhattacharya et al. (2016). However, posterior inferences obtained under these continuous priors do not induce any posterior model probabilities. Nor is it straightforward to choose a model or select variables.

In this dissertation, a scalable stochastic model search algorithm called *Simplified Shotgun Stochastic Search with Screening* (S5) is proposed, and its empirical performance is examined. S5 is a simplified version of SSS and it utilizes a screening step embedded in the algorithm to reduce the model space to be searched. Even though S5 is motivated by SSS, its efficiency

in searching interesting regions in the model space is remarkably improved by adopting a screening algorithm. For linear model selection in high-dimensional settings, the S5 algorithm often finds the MAP model hundreds of times faster than SSS does, but it identifies the same MAP model as SSS in all data sets examined in this dissertation. Furthermore, S5 accurately approximates posterior model probabilities and approximated posterior model probabilities are almost identical to those obtained from SSS. This algorithm is applicable to any variable selection procedures as long as a sound screening procedure is available. That is, it can be used for logistic regression models and nonparametric additive models. I extend the S5 algorithm to search the space of nonparametric additive models by adding a nonparametric screening step in the algorithm, and used it to implement the nonparametric model selection procedure that is described in the following sections.

An `R` functions that implements S5 is available in the `R` package `BayesS5` in `CRAN` (`https://cran.r-project.org`). This package includes a parallelized version of the code, which lets multiple independent chains search the model space simultaneously. This allows the algorithm explore a wider range of interesting regions in the model space. Simple tutorials about the package are also provided in this dissertation.

1.2.2 Nonparametric Model Selection in High-dimensional Settings

A Novel Shrinkage Prior for Nonparametric Regression

Frequently, practitioners face the problem of choosing between a parametric model and a nonparametric model, where the parametric model is nested within a more general class of functions. For example, a simple linear regression model or a nonparametric regression model might be considered for a data set, and the linear regression model is a special case of the nonparametric model. However, sometimes building a reasonable criterion for the choice between the parametric form and nonparametric form of the function is not evident, especially when multiple functions are involved in the model.

From a frequentist perspective, there has been a surge of interest in solving this problem using various forms of penalized estimation via the group Lasso (Yuan and Lin, 2006). Variable selection based on the group Lasso for partially linear additive models was studied in Zhang et al. (2011), where it was shown that the resulting procedure identifies the underlying true model structure correctly and at the same time estimates the multivariate regression function consistently. For variable selection problems in high-dimensional additive models, several penalized likelihood approaches using the group Lasso penalty have been proposed in Ravikumar et al. (2009); Meier et al. (2009); Huang et al. (2010). These approaches force the objective function to shrink only towards the zero function, and cannot impose shrinkage towards a more general class of functions, which is useful in many practical examples. For example, in log-density estimation problems, when the logarithm of a density function is quadratic, the resulting density function is Gaussian. This means that if we let the log-density function shrink towards a class of quadratic functions, the resulting estimated density can converge a Gaussian density. However, shrinkage procedures that accomplish this more general form of shrinkage have not been investigated in either Bayesian and frequentist frameworks.

In this dissertation, I propose a new shrinkage prior called *functional horseshoe prior* (fHS) that encourages shrinkage of the function towards a general class of functions including zero, constant, linear and quadratic functions. The resulting posterior mean of the function obtained from the fHS prior is expressed as a mixture of nonparametric and parametric estimators. Hence, by using the fHS prior, when the true function is in a class of parametric functions that are specified in advance, the posterior distribution of the function behaves as if the parametric model is used, and when the true function is strictly separated form the class of parametric functions, the resulting posterior distribution holds its nonparametric properties.

To construct the fHS prior, I introduce a novel semi-norm that measures the discrepancy between a function and a class of parametric functions. For the nonparametric regression model in (1.4), the semi-norm is $F^{\mathrm{T}}(I-Q_0)F$, where Q_0 is the projection matrix of the null covariates

that span the class of parametric functions. For example, the semi-norm can be interpreted as

$$F \text{ is linear } \iff F^{\mathrm{T}}(I - Q_0)F = 0, \tag{1.10}$$

by setting Q_0 to be the projection matrix of $\{\mathbf{1}, X\}$. The above relation is natural, because any linear function that is expressed as $a + bX$ for some $a, b \in \mathbb{R}$ must have zero sum of square residuals from a linear model, which is $F^{\mathrm{T}}(I - Q_0)F = 0$. Unlike existing shrinkage priors for shrinkage on a parameter towards zero such as the horseshoe prior (Carvalho et al., 2010), the shrinkage of the fHS prior acts on this semi-norm of the function and compels shrinkage towards the class of parametric functions (linear functions in the above example). Further, the proposed prior provides a natural connection to a new class of penalized likelihood methods which can be interpreted from a frequentist perspective.

Theoretical properties of the fHS priors are studied, and it is shown that under some mild conditions the posterior contraction rate achieves the parametric optimal rate $n^{-1/2}$ under the L_2 norm when the true function lies on the class of pre-specified parametric functions. That is, resulting inferences maintain the optimal nonparametric rate up to a logarithm term of n when the underlying function is not parametric. This result suggests that the use of the fHS prior can improve the estimation performance when the underlying function is parametric, and it does not degrade the estimation when the underlying model is nonparametric. The product of the fHS priors can be applied to the additive models in (1.7) to select variables by letting each component function shrink towards the zero function ($Q_0 = 0$). I evaluate the performance of this methodology through multiple real and simulated data sets. In terms of estimation and model selection, the proposed prior outperforms the state-of-the-art alternative methods including the standard horseshoe prior and the penalized likelihood procedure using the group Lasso.

A Novel Nonlocal Prior for Nonparametric Model Selection

As briefly discussed in Section 1.1.2, for nonparametric hypothesis testing problems in (1.5), Choi et al. (2009) and Choi and Rousseau (2015) have shown that Bayes factors defined by certain classes of priors achieve consistency. Even though these approaches showed that the convergence rate of Bayes factors in favor of alternative hypotheses increases at exponential rate of n under a true alternative hypothesis, they did not address the asymptotic behavior of the Bayes factor under true null hypothesis. This asymptotic study of Bayes factors under true null hypothesis is important, especially when the number of functions to be tested increases as sample size n grows.

I show that local prior densities, which assign positive probability around a null function in nonparametric Bayesian hypothesis tests, provide exponential accumulation of evidence in favor of an alternative hypothesis under a true alternative hypothesis, but only a polynomial rate of accumulation in favor of null hypothesis under true null. This imbalanced behavior has been noted also in parametric hypothesis testing problems as discussed in Johnson and Rossell (2010).

For parametric hypothesis testing problems (Johnson and Rossell, 2010), the nonlocal prior densities defined in Definition 2 were proposed to improve the convergence rate of the Bayes factor under a true null. These priors ameliorates the imbalanced behavior of the convergence rate of Bayes factor. Johnson and Rossell (2012) showed that the application of nonlocal priors to linear model selection procedures resulted in consistency in high-dimensional settings, whereas procedures based on local priors failed to be consistent.

To improve the convergence rate of nonparametric Bayes factor and pursue a consistent model selection procedure for nonparametric models in high-dimensional settings, the same strategy as nonlocal priors seems compelling in nonparametric settings. However, the application of the nonlocal idea to nonparametric models has been challenging. Unlike the null

hypothesis for scalar valued parameters, the nonparametric null hypothesis in (1.5) is composite. This means that the null hypothesis does not define a unique density for generating the data. Thus, the null space of functions is difficult to be parameterized, and this has hindered a consideration of an extension of nonlocal prior densities to nonparametric models.

In this dissertation, by using a novel semi-norm $F^{\mathrm{T}}(I-Q_0)F$ introduced in (1.10), I define the null space of functions in (1.5) as $\{F : F^{\mathrm{T}}(I-Q_0)F = 0\}$. I then construct a new class of nonlocal priors called *nonlocal functional prior* densities for nonparametric hypothesis testing and model selection problems. I provide the convergence rate of Bayes factors based on the nonlocal functional priors. When the true data-generating process is from the null model, I show that the convergencerate is much faster than that obtained from local priors. Finally, I apply the nonlocal functional prior to variable selection problems for the additive model in (1.7). Under mild regularity conditions, the consistency of the resulting model selection procedure is shown in high-dimensional settings where the number of predictors p increases at sub-exponential rate of n. A wide range of simulated and real data sets are considered to examine the model selection performance of the nonlocal functional prior, showing that it has better or comparable performance compared to all of its current competitors.

1.3 Outline

In Chapter 2, I consider model selection consistency for nonlocal prior densities in high-dimensional settings where the dimensionality p is allowed to increase at sub-exponential rate in n. Under suitable regularity conditions, the asymptotic form of the logarithm of the posterior model probability based on the nonlocal prior is illustrated. I show that it contains a unique form of adaptive penalty that cannot be derived from local priors.

In Chapter 3, I provide a detailed description of the S5 algorithm. Its efficiency is examined by using simulated and real data sets. Also, I provide examples of the implementation of the S5 algorithm using the `R` package `BayesS5`.

In Chapter 4, I propose a new class of shrinkage densities called the fHS prior for nonparametric models. These shrinkage priors bridge the gap between parametric functions and nonparametric functions. Under mild conditions, I show that when the true underlying function has a parametric form that is pre-specified in advance, the resulting posterior distribution contracts at the parametric optimal rate $n^{-1/2}$ under the L_2 norm, and that it achieves the optimal nonparametric rate when the true function is strictly separated from the class of parametric functions. I apply the fHS prior to additive models to improve estimation and select variables. For several real and simulated data sets, it shows outstanding performance in both estimation and model selection.

In Chapter 5, the nonlocal functional prior is proposed for nonparametric hypothesis testing (or model selection). I show that local prior densities, which assign positive probability around a null function in nonparametric Bayesian tests, provide exponential accumulation of evidence in favor of an alternative hypothesis under a true alternative hypothesis, but only a polynomial rate of accumulation in favor of a null hypothesis under a true null. This imbalanced behavior of the convergence rate of Bayes factor can be ameliorated by nonlocal I functional priors, and the resulting hypothesis testing procedures strongly penalize cases where the null hypotheses are rejected. I apply the proposed prior densities for high-dimensional model selection of nonparametric additive models and investigate model selection consistency of the resulting model selection procedures. I provide simulation studies and real data examples wher the proposed model selection procedure outperforms state-of-the-art methods.

The proofs of theoretical results in this dissertation appear in Appendix A, while the technical details of computation are presented in Appendix B.

2. NONLOCAL PRIOR DENSITIES FOR HIGH-DIMENSIONAL LINEAR MODEL SELECTION

2.1 Introduction

In the context of hypothesis testing, Johnson and Rossell (2010) defined nonlocal (alternative) priors as densities that are exactly zero whenever a model parameter equals its null value. Nonlocal priors were extended to model selection problems in Johnson and Rossell (2012), where product moment (pMoM) prior and product inverse moment (piMoM) prior densities were introduced as priors on a vector of regression coefficients. In $p \leq n$ settings, model selection procedures based on these priors were demonstrated to have a model selection property: the posterior probability of the true model converges to 1 as the sample size n increases. More recently, Rossell et al. (2013) and Rossell and Telesca (2017) proposed product exponential moment (peMoM) prior densities that have similar behavior to piMoM densities near the origin. However, the behavior of nonlocal priors in $p \gg n$ settings remains understudied to date (particularly in comparison to other commonly used variables selection procedures), which serves as the motivation for this dissertation.

I undertook a detailed simulation study to compare the performance of nonlocal priors in $p \gg n$ settings under sparsity with a host of penalization methods including the least absolute shrinkage and selection operator (Lasso; Tibshirani (1996)), smoothly clipped absolute deviation (Scad; Fan and Li (2001)), adaptive Lasso (Zou, 2006), minimum convex penalty (MCP; Zhang (2010)), and the reciprocal Lasso (rLasso), recently proposed by Song and Liang (2015). The penalty function of the rLasso is equivalent to the negative log-kernel of nonlocal prior densities; further connections are described in Section 2.5. As a natural Bayesian competitor, I also considered the widely used g-prior (Zellner, 1986; Liang et al., 2008), which is a local prior in the sense of Johnson and Rossell (2010). I used precision-recall curves (Davis

and Goadrich, 2006) as a basis for comparison between methods. These curves eliminate the effect of the choice of tuning parameters for each method so that the comparison across different methods is transparent. In cases where only a tiny proportion of variables are significant, precision-recall curves are more appropriate tools for comparison than are the more widely used receiver operating characteristic curves (Davis and Goadrich, 2006). While ROC curves present a trade-off between the type I error and the power of a decision procedure, precision-recall curves examine the trade-off between the power and the false discovery rate.

My studies indicate that Bayesian procedures based on nonlocal priors and the g-prior perform better than penalized likelihood approaches in the sense that they achieve a lower false discovery rate while maintaining a given level of statistical power. Furthermore, I find that posterior distributions on the model space based on nonlocal priors are more tightly concentrated around the maximum a posteriori model than the posterior based on g-priors, implying that they have a faster rate of posterior concentration. I also identified the oracle hyperparameter that maximizes the posterior probability of the true model for the Bayesian procedures. The growth-rate of these oracle hyperparameters with p also offers an interesting contrast between nonlocal and local priors. In the case of g-priors, the oracle value of g varied between 7.83×10^8 and 4.29×10^{13} as p ranged between 1000 and 20000 in a variety of simulation settings. For the same range of p, the oracle value of τ varied between 1.97 and 3.60, where τ is the tuning parameter for nonlocal priors described in Section 2. George and Foster (2000) argued from a minimax perspective that the g parameter should satisfy $g \asymp p^2$, which explains the large values of the optimal g. However, using asymptotic arguments to obtain default hyperparameters is difficult because the constant of proportionality is typically unknown. Moreover, when g is very large, the g-prior assigns negligible prior mass at the origin, essentially resulting in a nonlocal like prior. A similar point can be made about the recently proposed Bayesian shrinking and diffusing (BASAD) priors Narisetty and He (2014). On the other hand, the optimal hyperparameter value for the nonlocal priors is stable with increasing p, growing at a very slow

rate.

Motivated by this empirical finding, I studied properties of two classes of nonlocal priors allowing the hyperparameter τ to scale with p. Using a fixed value of τ, it seems that model selection consistency is possible only when $p \leq n$ (Johnson and Rossell (2012)). In this article, I establish that nonlocal priors can achieve model selection consistency even when the number of variables p increases sub-exponentially in the sample size n, provided that the hyperparameter τ is asymptotically larger than $\log p$. This theoretical result is consistent with my empirical finding.

2.2 Nonlocal Prior Densities for Regression Coefficients

I consider the standard setup of a Gaussian linear regression model with a univariate response and p candidate predictors. Let $\mathbf{y} = \{y_1, \ldots, y_n\}^{\mathrm{T}}$ denote a vector of responses for n individuals and X an $n \times p$ matrix of covariates. I denote a model by an index set of variables $\mathbf{k} = \{k_1, \ldots, k_{|\mathbf{k}|}\}$, with $1 \leq k_1 < \ldots < k_{|\mathbf{k}|} \leq p$. Given a model $\mathbf{k}$, let $X_{\mathbf{k}}$ denote the design matrix formed from the columns of X corresponding to the model $\mathbf{k}$ and $\beta_{\mathbf{k}} = \{\beta_{k,1}, \ldots, \beta_{k,|\mathbf{k}|}\}^{\mathrm{T}}$ the regression coefficient for the model $\mathbf{k}$. Under each model $\mathbf{k}$, the linear regression model for the data is

$$\mathbf{y} = X_{\mathbf{k}}\beta_{\mathbf{k}} + \epsilon, \tag{2.1}$$

where $\epsilon \sim N_n(0, \sigma^2 \mathrm{I}_n)$. Let $\mathbf{t}$ denote the true, or data-generating model and let $\beta^0_{\mathbf{t}}$ be the true regression coefficient under model $\mathbf{t}$. I assume that the true model is fixed but unknown.

Given a model $\mathbf{k}$, the product exponential moment (peMoM) prior density (Rossell et al., 2013; Rossell and Telesca, 2017) for the vector of regression coefficients $\beta_{\mathbf{k}}$ is defined as

$$\pi(\beta_{\mathbf{k}} \mid \sigma^2, \tau, \mathbf{k}) = C^{-|\mathbf{k}|} \prod_{j=1}^{|\mathbf{k}|} \exp\{-\beta^2_{\mathbf{k},j}/(2\sigma^2\tau) - \tau/\beta^2_{\mathbf{k},j}\}. \tag{2.2}$$

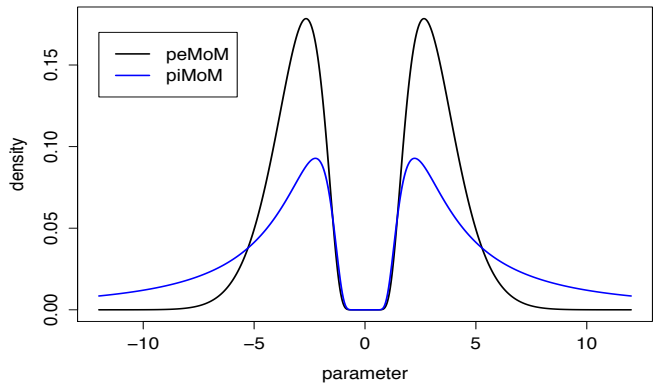


Figure 2.1: Nonlocal prior density functions for a single regression coefficient with $\tau = 5$; for the piMoM prior, $r = 1$.

The normalizing constant C can be explicitly calculated as

$$C = \int_{-\infty}^{\infty} \exp\{-t^2/(2\sigma^2\tau) - \tau/t^2\}dt = (2\pi\sigma^2\tau)^{1/2} \exp\{-(2/\sigma^2)^{1/2}\}, \tag{2.3}$$

since $\int \exp\{-\mu/t^2 - \zeta t^2\}dt = (\pi/\zeta)^{1/2} \exp\{-2(\mu\zeta)^{1/2}\}$.

Second, for a fixed positive integer r, the product inverse-moment (piMoM) prior density (Johnson and Rossell, 2012) for $\beta_{\mathbf{k}}$ is given by

$$\pi(\beta_{\mathbf{k}} \mid \sigma^2, \tau, \mathbf{k}) = C^{*-|\mathbf{k}|} \prod_{j=1}^{|\mathbf{k}|} [(\beta_{\mathbf{k},j})^{-2r} \exp\{-\tau/\beta_{\mathbf{k},j}^2\}], \tag{2.4}$$

where $C^* = \tau^{-r+1/2}\Gamma(r - 1/2)$ for $r > 1/2$ and $\Gamma(\cdot)$ is the gamma function.

The piMoM and peMoM prior densities are nonlocal in the sense that the density value at the origin is exactly zero. This feature of the densities for a single regression coefficient is illustrated in Figure 2.1. Since the piMoM prior densities and the peMoM prior densities have the same term $\exp\{-\tau/\beta^2\}$ that controls the behavior of the density function around the origin, they attain almost the same shape of the density function at the origin, which yields the

similar properties. Further details regarding this point are discussed in Section 2.4.

I focus on these two classes of nonlocal priors in the sequel. Note that in both (2.2) and (2.4), $\pi(\beta_{\mathbf{k}}) = 0$ when $\beta_{\mathbf{k}} = 0$; a defining feature of nonlocal priors. The distinction between the peMoM and the piMoM priors mainly involves their tail behavior. Whereas peMoM priors possess Gaussian tails, the piMoM prior densities have inverse polynomial tails. For example, piMoM densities with $r = 1$ have Cauchy-like tails, which has implications for their finite sample consistency and asymptotic bias in posterior mean estimates of regression coefficients. Because similar constraints are imposed on the hyperparameter τ appearing in both (2.2) and (2.4), at the risk of some ambiguity I use the same symbol for the two hyperparameters in these equations.

In addition to imposing priors on the regression parameters given a model, I need to place a prior on the space of models to complete the prior specification. I consider a uniform prior on the model space restricted to models having size less than or equal to q_n, with $q_n < n$, i.e.,

$$\pi(\mathbf{k}) \propto I(|\mathbf{k}| \leq q_n), \tag{2.5}$$

where $I(\cdot)$ denotes the indicator function and with a slight abuse of notation, I denote the prior on the space of models by π as well. Similar priors have been considered in the literature by Jiang (2007) and Liang et al. (2013). Since the peMoM and piMoM priors already induce a strong penalty on the size of the model space (see Section 2.4), I do not need to additionally penalize larger models using, for example, model space priors of the type discussed in Scott and Berger (2010).

Under a peMoM prior (2.2) on the regression coefficients, the marginal likelihood $m_{\mathbf{k}}(y)$ under model $\mathbf{k}$ given σ^2 can be obtained by integrating out $\beta_{\mathbf{k}}$, resulting in
$m_{\mathbf{k}}(y) = (2\pi\sigma^2)^{-\frac{n}{2}}\, C^{-|\mathbf{k}|}\, Q_{\mathbf{k}}\, \exp\{-\widetilde{R}_{\mathbf{k}}/(2\sigma^2)\}$,

where

$$\begin{aligned}
\widetilde{R}_{\mathbf{k}} &= y^T(\mathrm{I}_n - \widetilde{\mathrm{P}}_{\mathbf{k}})y, \quad \widetilde{\mathrm{P}}_{\mathbf{k}} = X_{\mathbf{k}}(X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{k}} + 1/\tau \mathrm{I}_{\mathbf{k}})^{-1}X_{\mathbf{k}}^{\mathrm{T}}, \\
Q_{\mathbf{k}} &= \int \exp\{-(\beta_{\mathbf{k}} - \widetilde{\beta}_{\mathbf{k}})^{\mathrm{T}}\widetilde{\Sigma}_{\mathbf{k}}^{-1}(\beta_{\mathbf{k}} - \widetilde{\beta}_{\mathbf{k}})/(2\sigma^2) - \sum_{j=1}^{|\mathbf{k}|}\tau/\beta_{\mathbf{k},j}^2\}d\beta_{\mathbf{k}}, \qquad (2.6) \\
\widetilde{\beta}_{\mathbf{k}} &= (X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{k}} + 1/\tau \mathrm{I}_{\mathbf{k}})^{-1}X_{\mathbf{k}}^{\mathrm{T}}y, \quad \widetilde{\Sigma}_{\mathbf{k}} = (X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{k}} + 1/\tau \mathrm{I}_{\mathbf{k}})^{-1}.
\end{aligned}$$

Similarly, the marginal likelihood using the piMoM prior densities (2.4) can be expressed as $m_{\mathbf{k}}(y) = (2\pi\sigma^2)^{-\frac{n}{2}}\, C^{*-|\mathbf{k}|}\, Q_{\mathbf{k}}^*\, \exp\{-R_{\mathbf{k}}^*/(2\sigma^2)\}$, where

$$\begin{aligned}
R_{\mathbf{k}}^* &= y^{\mathrm{T}}(\mathrm{I}_n - \mathrm{P}_{\mathbf{k}})y, \quad \mathrm{P}_{\mathbf{k}} = X_{\mathbf{k}}\left(X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{k}}\right)^{-1}X_{\mathbf{k}}^{\mathrm{T}}, \\
Q_{\mathbf{k}}^* &= \int \prod_{j=1}^{|\mathbf{k}|}\beta_{\mathbf{k},j}^{-2r}\exp\{-(\beta_{\mathbf{k}} - \widehat{\beta}_{\mathbf{k}})^{\mathrm{T}}\Sigma_{\mathbf{k}}^{*-1}(\beta_{\mathbf{k}} - \widehat{\beta}_{\mathbf{k}})/(2\sigma^2) - \sum_{j=1}^{|\mathbf{k}|}\tau/\beta_{\mathbf{k},j}^2\}d\beta_{\mathbf{k}}, \qquad (2.7) \\
\widehat{\beta}_{\mathbf{k}} &= (X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{k}})^{-1}X_{\mathbf{k}}^{\mathrm{T}}y, \quad \Sigma_{\mathbf{k}}^* = (X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{k}})^{-1}.
\end{aligned}$$

The integrals for $Q_{\mathbf{k}}$ and $Q_{\mathbf{k}}^*$ cannot be obtained in closed form, so for computational purposes I make Laplace approximations to $m_{\mathbf{k}}(y)$. The expressions for the marginal likelihood derived here is nevertheless important for theoretical studies in Section 2.4.

2.3 Numerical Results

2.3.1 Simulation Studies Using Precision-Recall Curves

To illustrate the performance of nonlocal priors in ultrahigh-dimensional settings and to compare their performance with other methods, I calculated precision-recall curves Davis and Goadrich (2006) for all selection procedures. A precision-recall curve plots the precision = TP/(TP + FP) versus recall (or sensitivity) = TP/(TP + FN), where TP, FP and FN respectively denote the number of true positives, false positives, and false negatives, as the tuning parameter is varied. The efficacy of a procedure can be measured by the area under the precision-recall

curve; the greater the area, the more accurate the method. Since both precision and recall take values in $[0, 1]$, the area under the curve for an ideal precision-recall curve is 1. I used two (n, p) combinations, namely $(n, p) = (400, 10000)$ and $(n, p) = (400, 20000)$, and plotted the average of the precision-recall curves obtained from 100 independent replicates of each procedure. To evaluate the marginal likelihood of each model, I used the Laplace approximation method.

I compared the performance of peMoM and piMoM priors to a number of frequentist penalized likelihood methods: Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006), Scad (Fan and Li, 2001), and Minimax Concave Penalty (MCP) (Zhang, 2010). I used the `R` package *ncvreg* to fit these penalized likelihood methods. I also included the reciprocal Lasso in the simulation studies. However, due to computational constraints involved in implementing the full rLasso procedure, I followed the recommendation in Song and Liang (2015) and instead implemented the reduced rLasso. The reduced rLasso procedure is a simplified version of rLasso that uses the least square estimators of β when minimizing the rLasso objective function.

I considered Zellner's g-prior Zellner (1986); Liang et al. (2008) as a competing Bayesian method, with $\beta_{\mathbf{k}} \mid \mathbf{k}, \sigma^2 \sim N(0, g\sigma^2(X_{\mathbf{k}}^{\mathrm{T}} X_{\mathbf{k}})^{-1})$ and g a tuning parameter. With the prior $\pi(\sigma^2) \propto 1/\sigma^2$, the marginal likelihood $m_{\mathbf{k}}(y) \propto (1+g)^{-|\mathbf{k}|/2}\{1 + g(1 - D_{\mathbf{k}}^2)\}^{-(n-1)/2}$ can be obtained in a closed form; see for example, (Liang et al., 2008, pp 412), where $D_{\mathbf{k}}^2$ is the ordinary coefficient of determination for the model $\mathbf{k}$.

A uniform model prior (2.5) was considered for all Bayesian procedures. This prior was chosen for several reasons. First, construction of the PR curves requires maximization over model hyperparameters, which is most easily achieved if there is only one unknown hyperparameter. I also wished to avoid providing an advantage to the Bayesian methods by introducing additional tuning parameters into these methods that were not present in the penalized likelihood methods. Furthermore, the use of non-uniform priors on the model space introduces (at least) one more degree of freedom into the comparisons between methods, and my intent was

to compare the effects of the penalties imposed on regression coefficients by both penalized likelihood and Bayesian methods. At first blush, this might appear to put Bayesian methods like those based on the g-prior at a disadvantage, since such methods do not yield consistent variable selection even in $p < n$ settings without prior sparsity penalties on the model space (when g is held fixed as n increases). However, in the construction of PR curves, I allowed prior hyperparameters to increase with n, which effectively allowed the Bayesian methods to impose additional sparseness penalties through the introduction of large hyperparameter values.

I arbitrarily fixed $r = 1$ for the piMoM prior (2.4) and used an inverse-gamma prior on σ^2 with parameters $(0.1, 0.1)$ for the peMoM, piMoM priors, and g-priors. Posterior computations for the peMoM, piMoM and g-priors were implemented using the Simplified Shotgun Stochastic Search with Screening (S5) algorithm described in Chapter 3. The maximum a posteriori model was used in each case to summarize the model selection performance. The precision-recall curves are drawn by varying the hyperparameters (τ for the nonlocal priors and g for the g-priors) so the comparison between the model selection based on the nonlocal priors and the g-prior is free of the choice of hyperparameters. Because of their high computational burden, I could not include BASAD Narisetty and He (2014) in the comparisons.

For each simulation setting, I simulated data according to a Gaussian linear model as in (2.1) with the fixed true model $\mathbf{t} = \{1, 2, 3, 4, 5\}$ with the true regression coefficient $\beta^0_{\mathbf{t}} = (0.50, 0.75, 1.00, 1.25, 1.50)^{\mathrm{T}}$ and $\sigma = 1.5$. Also, the signs of the true regression coefficients were randomly determined with probability one-half. Each row of X was independently generated from a $N(0, \Sigma)$ distribution with one of the following covariance structures:

Case (1): compound symmetry design; $\Sigma_{jj'} = 0.5$, if $j \neq j'$ and $\Sigma_{jj} = 1$, $1 \leq j, j' \leq p$.

Case (2): autoregressive correlated design; $\Sigma_{jj'} = 0.5^{|j-j'|}$, $1 \leq j, j' \leq p$.

Case (3): isotropic design; $\Sigma = \mathrm{I}_p$.

Figure 2.2 plots the precision-recall curves averaged over 100 simulation replicates for the

different methods across the two (n,p) pairs and the three covariate designs. From Figure 2.2, it is evident that the precision-recall curves for the peMoM and piMoM priors have an overall better performance than the penalized likelihood methods Lasso, adaptive Lasso, Scad, and MCP. For decision procedures having the same power, this implies that the nonlocal priors achieve lower false discovery rates. As discussed in Section 2.5, since the reduced rLasso shares the same nonlocal kernel as the nonlocal priors, it has a similar selection performance. The figure also shows that Zellner's g-prior attains comparable performance with the nonlocal priors in terms of the precision-recall curves.

2.3.2 Further Comparison with Zellner's g-prior

The similarity of the performances of the g-prior and the nonlocal priors in terms of precision-recall curves begs for closer comparisons of these procedures. For this reason, I also investigated the concentration of the posterior densities around their maximum models. To this end, I fixed $p = 20,000$ and varied n from 150 to 400; the data generating mechanism was exactly the same as in Section 2.3.1. The left column of Figure 2.3 displays the posterior probability of the true model under the peMoM, piMoM and g-prior models versus n for the three covariate designs in Section 2.3.1. The plot shows that the posterior probability of the true model increases with n for all three methods, with the peMoM and piMoM priors almost uniformly dominating the g-prior, implying a higher concentration of the posterior around the true model for the nonlocal priors.

This tendency is confirmed in the right panel of Figure 2.3, where I plot the number of models $\mathbf{k}$ which achieve a posterior odds ratio $\pi(\mathbf{k} \mid y)/\pi(\widehat{\mathbf{k}} \mid y) > 0.001$, where $\widehat{\mathbf{k}}$ is the maximum a posteriori model. This plot clearly shows that the posterior distribution on the model space from the g-priors is more diffuse than those obtained using the nonlocal prior methods. These comparisons were based on fitting the hyperparameters g and τ at their oracle value, i.e., the value which maximized the posterior probability of the true model for a given

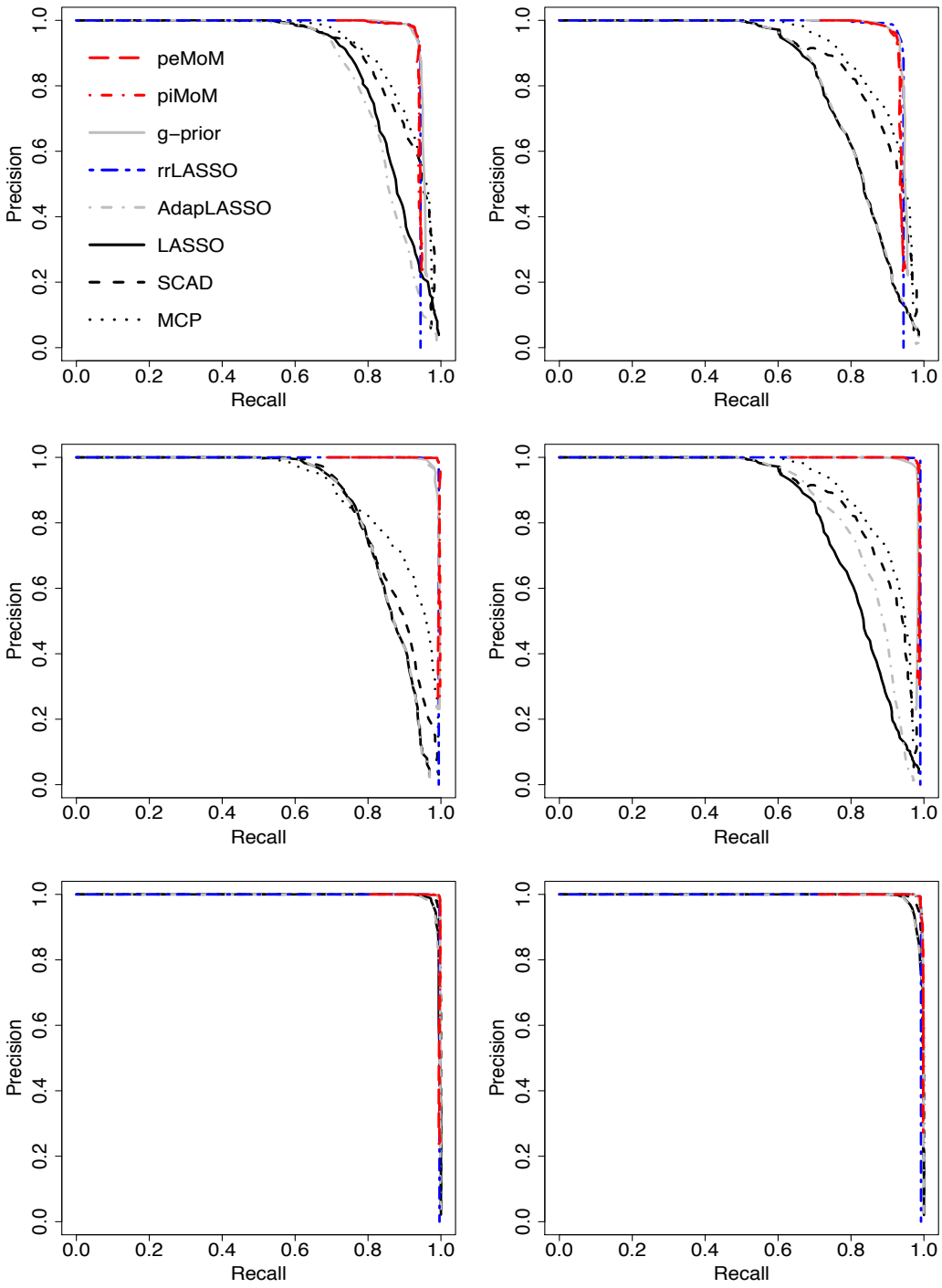


Figure 2.2: Plot of the mean precision-precision curves over 100 datasets with $(n, p) = (400, 10000)$(first column) and $(n, p) = (400, 20000)$(second column). Top: case (1); middle: case (2); bottom: case (3).

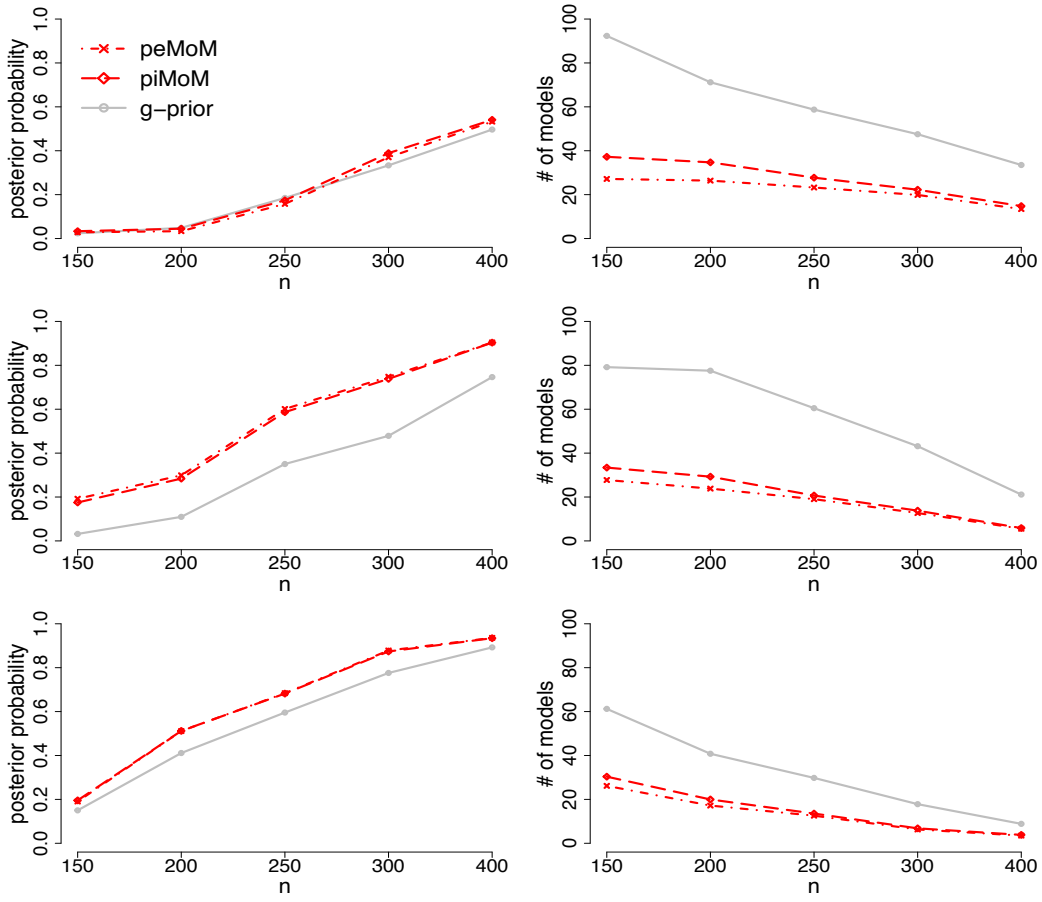


Figure 2.3: Averaged posterior true model probability and the number of models which attain the posterior odds ratio, with respect to the maximum a posteriori model, larger than 0.001 with the fixed $p = 20000$ and varying n. Top: case (1); middle: case (2); bottom: case (3).

Table 2.1: Optimal hyperparameters for Bayesian model selection methods

		The number of predictors				
Case		$p = 1000$	$p = 2000$	$p = 5000$	$p = 10000$	$p = 20000$
(1)	peMoM	2.24	2.72	2.88	3.32	3.60
	piMoM	2.16	2.59	2.70	3.04	3.26
	g-prior	7.83×10^8	2.87×10^9	3.05×10^9	9.66×10^9	1.70×10^{10}
(2)	peMoM	1.97	2.29	2.34	2.75	3.00
	piMoM	1.97	2.20	2.32	2.66	2.86
	g-prior	8.56×10^9	2.55×10^{10}	2.62×10^{10}	6.58×10^{10}	1.25×10^{11}
(3)	peMoM	2.66	3.00	3.00	3.10	3.60
	piMoM	2.61	2.94	2.94	2.94	3.46
	g-prior	1.26×10^{12}	8.84×10^{12}	9.67×10^{12}	6.81×10^{12}	4.29×10^{13}

value of n.

The magnitudes of the oracle hyperparameters under each model also present an interesting contrast between the local and nonlocal priors. I observed that the oracle value of g increased rapidly with p, whereas the oracle value of τ was much more stable. This phenomenon is illustrated in Table 2.1, which shows the oracle hyperparameter value averaged over 100 replicates for the three different covariate designs in Section 2.3.1. For this comparison, I fixed $n = 400$ and varied p between 1000 and $20,000$; five representative values are displayed. The oracle values for g are on a completely different scale from the oracle values of τ, and they vary more with p. This table confirms the recommendations in George and Foster (2000) for setting $g = p^2$ based on minimax arguments. However, the finite sample behavior of the optimal choice of g is unclear, which means that the large variance of the optimal hyperparameter value is likely to hinder the selection of g in real applications. Finally, I note that such large values of g effectively convert the local g-priors into nonlocal priors by collapsing the g-prior density to 0 at the origin.

2.4 Model Selection Consistency

The empirical performance of the peMoM and piMoM priors suggests that the hyperparameter τ should be increased slowly with p. While Johnson and Rossell (2012) were able to show strong selection consistency with a fixed value of τ, it is not clear whether their proof can be extended to $p \gg n$ cases. Motivated by the empirical findings of the last section, I next investigated the strong consistency properties of peMoM and piMoM priors when τ was allowed to grow at a logarithmic rate in p. I found that in such cases, both peMoM and piMoM priors achieve model selection consistency under standard regularity assumptions when p increases sub-exponentially with n, i.e., $\log p = O(n^{\alpha})$ for $\alpha \in (0, 1)$.

Henceforth, I use $\tau_{n,p}$ instead of τ to denote the hyperparameter in the peMoM and piMoM priors in (2.2) and (2.4) respectively. The normalizing constants for these priors are now denoted by $C_{n,p}$ and $C^*_{n,p}$, respectively. Before providing my theoretical results, I first state a number of regularity conditions. Let $\nu_j(A)$ denote the j-th largest nonzero eigenvalue of an arbitrary matrix A, and let

$$\nu_{\mathbf{k}*} = \min_{1 \le j \le min(n, |\mathbf{k}|)} \nu_j(X_{\mathbf{k}}^{\mathrm{T}} X_{\mathbf{k}}/n), \quad \nu^*_{\mathbf{k}} = \max_{1 \le j \le min(n, |\mathbf{k}|)} \nu_j(X_{\mathbf{k}}^{\mathrm{T}} X_{\mathbf{k}}/n). \tag{2.8}$$

For sequences a_n and b_n, $a_n \succeq b_n$ indicates $b_n = O(a_n)$, and $a_n \succ b_n$ indicates $b_n = o(a_n)$. With this notation, I assume that the following regularity conditions apply.

Assumption 1. There exists $\alpha \in (0, 1)$ such that $\log p = O(n^{\alpha})$.

Assumption 2. $\log p \prec \tau_{n,p} \prec n$.

Assumption 3. $|\mathbf{k}| \le q_n$, where $q_n \prec \frac{\tau_{n,p}}{\log p}$.

Assumption 4. $\min_{\mathbf{k}: |\mathbf{k}| \le q_n} \nu_{\mathbf{k}*} \succ \frac{\tau_{n,p}}{n}$.

Assumption 5. $C_1 < \nu_{\mathbf{t}*} \leq \nu^*_{\mathbf{t}} < C_2$ for some positive constants C_1 and C_2.

Several comments regarding these conditions are worth making. Assumption 1 allows p to grow sub-exponentially with n. My theoretical results continue to hold when p grows polynomially in n, i.e., at the rate $O(n^\gamma)$ for some $\gamma > 1$. Assumption 2 reflects the empirical findings about the oracle $\tau \equiv \tau_{n,p}$ in Section 2.3.1, which was observed to grow slowly with p. I need the bound on q_n in Assumption 3 to ensure that the least square estimator of a model is consistent when a model contains the true model. In the $p \leq n$ setting, Johnson and Rossell (2012) assumed that all eigenvalues of the Gram matrix $(X^{\mathrm{T}}_{\mathbf{k}} X_{\mathbf{k}})/n$ are bounded above and below by global constants for all $\mathbf{k}$. However, this assumption is no longer viable when $p \gg n$ and I replace that by Assumption 4, where the minimum of the minimum eigenvalue of $(X^{\mathrm{T}}_{\mathbf{k}} X_{\mathbf{k}})/n$ over all submodels $\mathbf{k}$ with $|\mathbf{k}| \leq q_n$ is allowed to decrease with increasing n and p. Assumption 4 is called the sparse Riesz condition and is also used in Chen and Chen (2008) and Kim et al. (2012). Narisetty and He (2014) showed that Assumption 4 holds with overwhelmingly large probability when the rows of the design matrix are independent with an isotropic sub-Gaussian distribution. Even though the assumption of sub-Gaussian tails on the covariates is difficult to verify, the results in Narisetty and He (2014) show that Assumption 4 can be satisfied for some sequence of design matrices.

I now state a Theorem that demonstrates that model selection procedures based on the peMoM and piMoM nonlocal prior densities achieve strong consistency under the proposed regularity conditions. A proof of the Theorem is provided in the Appendix.

Theorem 1. *Suppose σ^2 is known and that Assumptions 1 – 5 hold. Let $\pi(\mathbf{t} \mid \mathbf{y})$ denote the posterior probability of the true model obtained under a peMoM prior* (2.2)*. Also, assume a uniform prior on all models of size less than or equal to q_n, i.e., $\pi(\mathbf{k}) \propto I(|\mathbf{k}| \leq q_n)$. Then, $\pi(\mathbf{t} \mid \mathbf{y})$ converges to one in probability as n goes to ∞.*

Corollary 2. *Assume the conditions of the preceding Theorem apply. Let* $\pi(\mathbf{t} \mid \mathbf{y})$ *denote the posterior probability of the true model obtained under a piMoM prior density* (2.4)*. Then,* $\pi(\mathbf{t} \mid \mathbf{y})$ *converges to one in probability as* n *goes to* ∞*.*

I note that these results apply also if a beta-Bernoulli prior is imposed on the model space as in Scott and Berger (2010), because the effect of that prior is asymptotically negligible when $|\mathbf{k}| \leq q_n \prec n$.

In most applications, σ^2 is unknown, and it is thus necessary to specify a prior density on it. By imposing a proper inverse gamma prior density on σ^2, I can obtain the model consistency result stated in the Theorem below. The proof is again deferred to the Appendix.

Theorem 3. *Suppose* σ^2 *is unknown and a proper inverse gamma density with parameters* (a_0, b_0) *is assumed for* σ^2*. Also, let* $\pi(\mathbf{t} \mid \mathbf{y})$ *denote the posterior probability of the true model evaluated using peMoM priors. Then if Assumptions 1 – 5 are satisfied,* $\pi(\mathbf{t} \mid \mathbf{y})$ *converges to one in probability as* n *goes to* ∞*.*

Corollary 4. *Suppose the conditions of the preceding Theorem apply, but that* $\pi(\mathbf{t} \mid \mathbf{y})$ *now denotes the posterior probability of the true model obtained under a piMoM prior density. Then* $\pi(\mathbf{t} \mid \mathbf{y})$ *converges to one in probability as* n *goes to* ∞*.*

2.5 Connections Between Nonlocal Priors and Reciprocal Lasso

In this section, I highlight the connection between the rLasso in Song and Liang (2015) and Bayesian variable selection procedures based on nonlocal priors. I begin by noting that the objective function $g(\beta_{\mathbf{k}}; \mathbf{k})$ of rLasso on a model $\mathbf{k}$ can be expressed as follows:

$$g(\beta_{\mathbf{k}}; \mathbf{k}) = \left\| y - X_{\mathbf{k}} \beta_{\mathbf{k}} \right\|_2^2 + \sum_{j=1}^{|\mathbf{k}|} \tau_{n,p} / |\beta_{\mathbf{k},j}|. \tag{2.9}$$

The optimal model is selected by minimizing this objective function with respect to $\beta_{\mathbf{k}}$ and $\mathbf{k}$. It is clear that the penalty function $\sum_{j=1}^{|\mathbf{k}|} \tau_{n,p} / |\beta_{\mathbf{k},j}|$ in (2.9) is similar to the negative log-

density of piMoM nonlocal priors as proposed in (Johnson and Rossell, 2012, pp 659) and (Johnson and Rossell, 2010, pp 149). The main difference between the nonlocal prior version of rLasso and the piMoM-type prior densities proposed in the previous section is the power of β in the exponential kernels. For the rLasso penalty this power is 1, while for piMoM-type prior densities it is 2. The implications of this difference are apparent from the following proposition.

Proposition 5. *For a given model* $\mathbf{k}$*, suppose that* $\widetilde{\beta}^*_{\mathbf{k}}$ *is the minimizer of the objective function* (2.9)*, and again let* $\widehat{\beta}_{\mathbf{k}}$ *denote the least square estimator of* β *under model* $\mathbf{k}$*. Assume that* $\tau_{n,p} \prec n$*, and there exist strictly positive contants* C_L *and* C_U *such that* $C_L < \nu_{\mathbf{k}*} \leq \nu^*_{\mathbf{k}} < C_U$*. Then, for any* $\epsilon^*_n \succ (\tau_{n,p}/n)^{1/3}$*,*

$$P\left[\widetilde{\beta}^*_{\mathbf{k}} \notin R\big(\widehat{\beta}_{\mathbf{k}}; \epsilon^*_n\big)\right] \to 0,$$

where $R(u;\epsilon) = \{\mathbf{x} \in \mathbb{R}^{|\mathbf{k}|} : |x_j - u_j| \leq \epsilon, j = 1, \ldots, |\mathbf{k}|\}$.

This proposition shows that under standard conditions on the eigenvalues of the Gram matrix $X^T_{\mathbf{k}} X_{\mathbf{k}}/n$, the estimator derived from (2.9) is asymptotically within $(\tau_{n,p}/n)^{1/3}$ distance of the least squares estimator $\widehat{\beta}_{\mathbf{k}}$. On the other hand, results cited in the previous section show that maximum a posteriori estimators obtained from the piMoM-type prior densities reside at an asymptotic distance of $(\tau_{n,p}/n)^{1/4}$ from the least squares estimator. Variable selection procedures based on both forms of piMoM priors thus achieve adaptive penalties on the regression coefficients in the sense described in Song and Liang (2015).

Although rLasso is proposed as a penalized likelihood approach, the computational procedure to optimize its objective function is quite different from the other penalized likelihood methods. The resulting computational complexity of this optimization procedure, which contains a discontinuous penalty function, is NP-hard. This suggests that the formulation of this nonlocal penalty in a penalized likelihood framework is unlikely to provide significant compu-

tational advantages over related Bayesian model selection procedures, even though the inferential advantages of the Bayesian framework are lost.

2.6 An Adaptive Form of Asymptotic Marginal Likelihoods Based on Nonlocal Priors

From Lemma A.1.1 in the Appendix, it follows that the asymptotic log-marginal likelihood of a model $\mathbf{k}$ based on a peMoM or piMoM prior density can be expressed as

$$\begin{aligned} \log \pi(\mathbf{k} \mid y) &= l(\widehat{\beta}_{\mathbf{k}}) + \log Q_{\mathbf{k}} - |\mathbf{k}| \log C_{n,p} \\ &\approx l(\widehat{\beta}_{\mathbf{k}}) - \sum_{j=1}^{|\mathbf{k}|} p_{\tau_{n,p}}\big(\widehat{\beta}_{\mathbf{k},j}\big) + C, \end{aligned}$$

for some constant C, where $\widehat{\beta}_{\mathbf{k}}$ is the maximum likelihood estimator under model $\mathbf{k}$, *i.e.* $\widehat{\beta}_{\mathbf{k}} = (X_{\mathbf{k}}^T X_{\mathbf{k}})^{-1} X_{\mathbf{k}}^T y$, and

$$p_{\tau_{n,p}}\big(\widehat{\beta}_{\mathbf{k},j}\big) \approx \begin{cases} (n\tau_{n,p} u_{\mathbf{k}})^{1/2}, & \text{if } |\widehat{\beta}_{\mathbf{k},j}| < c\big(\frac{nu_{\mathbf{k}}}{\tau_{n,p}}\big)^{-1/4} \\ \tau_{n,p}/\widehat{\beta}_{\mathbf{k},j}^2, & \text{if } |\widehat{\beta}_{\mathbf{k},j}| \geq c\big(\frac{nu_{\mathbf{k}}}{\tau_{n,p}}\big)^{-1/4}, \end{cases} \tag{2.10}$$

for some constant c and some arbitrary sequence $u_{\mathbf{k}}$ with $\nu_{\mathbf{k}*} \leq u_{\mathbf{k}} \leq \nu_{\mathbf{k}}^*$. I note that the strength of the correlation between the variables in model $\mathbf{k}$ affects the behavior of $u_{\mathbf{k}}$, and $(nu_{\mathbf{k}}/\tau_{n,p})^{-1/4}$ converges to zero as n tends to infinity due to Assumption 4 described in Section 2.4.

On the other hand, the penalty term in the other Bayesian model selection approaches is quite different from that of the nonlocal priors as in (2.10). The marginal likelihood based on the g-prior when σ^2 is known can be expressed as

$$l(\widehat{\beta}_{\mathbf{k}}) - |\mathbf{k}| \log(1+g)/2.$$

Narisetty and He(2014) demonstrated that BASAD achieves model selection consistency.

This consistency follows from that the fact that the BASAD "penalty" is asymptotically equivalent to

$$l(\widehat{\beta}_{\mathbf{k}}) - c|\mathbf{k}|\log(p), \tag{2.11}$$

where c is some constant. Yang et al. (2016) and Castillo and van der Vaart (2012) also considered a similar penalty term on the model space, which implies that the posterior probability for their procedures can be expressed in the same form as (2.11). When $g = p^{2c}$, the marginal likelihood based on a g-prior is asymptotically equivalent to (2.11).

This asymptotic term of the marginal likelihoods is quite different from that of the nonlocal priors, since the penalty terms in the other Bayesian approaches only focus on the model size without considering the different weights on variables in the model. The marginal likelihoods based on nonlocal priors, however, impose different penalties on each predictor in the given model. When the MLE of the regression coefficient in the model is asymptotically close to zero ($|\widehat{\beta}_{\mathbf{k},j}| < c(nu_{\mathbf{k}}/\tau_{n,p})^{-1/4}$), the model that contains the corresponding variable would be strongly penalized by $(n\tau_{n,p}u_{\mathbf{k}})^{1/2}$. In contrast, when the MLE is asymptotically significant ($|\widehat{\beta}_{\mathbf{k},j}| \geq c(nu_{\mathbf{k}}/\tau_{n,p})^{-1/4}$), the penalty attains a different weight based on the MLE ($p_{\tau_{n,p}}(\widehat{\beta}_{\mathbf{k},j}) \approx \tau_{n,p}/\widehat{\beta}^2_{\mathbf{k},j}$).

This analysis highlights the fact that the nonlocal priors are able to adapt their penalty for the inclusion of covariates based on the observed data, whereas local priors must instead rely on a prior penalty that encourages non-sparse models.

2.7 Real Data Analysis

2.7.1 Analysis of Polymerase Chain Reaction (PCR) data

Lan et al. (2006) studied coordinated regulation of gene expression levels on 31 female and 29 male mice ($n = 60$). A number of psychological phenotypes, including numbers of

stearoyl-CoA desaturase 1 (SCD1), glycerol-3-phosphate acyltransferase (GPAT) and phosphoenopyruvate carboxykinase (PEPCK), were measured by quantitative real-time RT-PCR, along with 22,575 gene expression values. The resulting data set is publicly available at `http://www.ncbi.nlm.nih.gov/geo` (accession number GSE3330).

Zhang et al. (2009) used penalized orthogonal components regression to predict the three phenotypes mentioned above based on the high-dimensional gene expression data. Bondell and Reich (2012) also used the same data set to examine their model selection procedure based on penalizing regression coefficients within a (marginal or joint) credible interval obtained from a ridge-type prior. For brevity, I restrict attention here to SCD1 as the response variable.

Since the ground truth regarding the true significant variables is not known for this data, I compared my approach with a host of competitors on predictive accuracy and parsimony of the selected model.

Prior to analyses, I standardized the covariates and randomly split the data set into 5 test samples and 55 training samples to evaluate the out-of-sample mean square prediction error (MSPE)

$$\text{MSPE} = \sum_{i \in T_{test}} (y_i - X_i^T \widehat{\beta}^{tr}_{\widehat{\mathbf{k}}})^2 / |T_{test}|,$$

where T_{test} is the index set of the test samples and $\widehat{\beta}^{tr}_{\widehat{\mathbf{k}}}$ is the least square estimator under the estimated model $\widehat{\mathbf{k}}$ based on the training samples. To avoid sensitivity to a particular split, I considered 100 replications of the training and test sample generation. To measure the stability of model selection, I considered the number of variables that were (i) selected at least 95 times, and (ii) at least once, out of the 100 replicates.

Due to the high-computational burden of the penalized credible interval approach (Bondell and Reich, 2012), I followed the pre-processing step suggested in that article to marginally screen variables to reduce to 2000 variables (1999 genes and gender). For all the other approaches, all 22,575 genes were used. For the nonlocal prior method, I considered both the

MAP estimator and the least squares (LS) estimator from the MAP model. For the g-prior, I set $g = p^2$ as recommended in George and Foster (2000). For the penalized likelihood procedures, I used ten-fold cross validation to choose the tuning parameter.

To choose the hyperparameter $\tau_{n,p}$ for the nonlocal priors, I used a procedure proposed by Nikooienejad et al. (2016). That procedure sets the hyperparameter so that the L_1 distance between the posterior distribution on the regression parameters under the null distribution (i.e., $\beta = 0$) and the nonlocal prior distributions on these parameters is constrained to be less than a specified value (e.g., $p^{-1/2}$). The average value of the hyperparameter values chosen by this procedure were $\tau_{n,p} = 1.12$ and $\tau_{n,p} = 1.16$ for piMoM and peMoM priors, respectively.

To make the comparison between the nonlocal priors and the g-prior more transparent, I used the same beta-binomial prior on the model space in both models, rather than the uniform prior on the model space described previously. The form of the beta-binomial prior was given by

$$\pi(\mathbf{k}) \propto \rho^{|\mathbf{k}|}(1-\rho)^{p-|\mathbf{k}|} I(|\mathbf{k}| \leq q_n), \tag{2.12}$$

with a uniform prior on ρ and $q_n = 40$. I note that this prior does not strongly induce sparsity as does, for example, the prior obtained by imposing a $Beta(1, p^u)$, $u > 1$ prior on ρ, as suggested in Castillo et al. (2015).

Table 2.2 summarizes the results from the analysis of the gene expression data set. On average, the nonlocal priors simultaneously produced the lowest MSPE and the most parsimonious model. The other model selection methods selected a wide array of different variables for different splits of the data set. In particular, Lasso and the penalized credible region approach selected more than 180 different variables from 100 repeated splits, while the average size of the selected model was less than 20 and the number of frequently selected variables was only zero or one, indicating a potentially large number of false positives picked up by these

Method	MSPE	MS	FS	TS
piMoM(MAP)	0.283 (0.17)	1.00 (0.00)	1	1
piMoM(LS)	**0.282** (0.17)	1.00 (0.00)	1	1
peMoM(MAP)	0.291 (0.18)	1.02 (0.14)	1	2
peMoM(LS)	0.287 (0.17)	1.02 (0.14)	1	2
g-prior	0.368 (0.20)	4.07 (0.56)	1	133
Lasso	0.542 (0.39)	17.97 (8.62)	1	211
Scad	0.308 (0.23)	12.66 (7.62)	2	163
MCP	0.308 (0.21)	2.20 (0.94)	0	29
Marginal($p = 2000$)	0.456 (0.40)	17.47 (11.16)	0	273
Joint($p = 2000$)	0.440 (0.40)	16.42 (11.06)	1	185

Table 2.2: Analysis of the PCR data. Marginal and Joint refer to the variable selection procedures Bondell and Reich (2012) based on Bayesian marginal credible set and Bayesian joint credible set, respectively. MS is the average size of the selected model. FS is the number of frequently selected variables, i.e., that were selected at least 95 times in 100 repetitions. TS refers to the total number of variables selected at least once from 100 repetitions. Standard errors are provided in parenthesis.

methods.

2.7.2 A Simulation Study Based on the Boston Housing Data

I next examined the Boston housing data set that contains the median value of owner-occupied homes in the Boston area, together with several variables that might be associated with their median value. There were $n = 506$ median values in the data set, and I considered 10 continuous variables as the predictor variables: `crim`, `indus`, `nox`, `rm`, `age`, `dis`, `tax`, `ptratio`, `b`, and `lstat`. This data set has been used to validate a variety of variable selection methods; some recent examples include Radchenko and James (2011), Yuan and Lin (2005), and Rockova and George (2014).

To examine the model selection performance in high-dimensional settings, I added 1,000 noise variables that were generated independently from a standard Gaussian distribution ($p = 1,010$). The same competitors from the previous subsection were used with the aforementioned choice of hyperparameters. For nonlocal priors, the hyperparameter value was chosen by the

aforementioned procedure Nikooienejad et al. (2016); the average of the chosen hyperparameter values were $\tau_{n,p} = 2.01$ and $\tau_{n,p} = 0.47$ for piMoM and peMoM priors, respectively. Prior to analyses, I standardized the covariates and considered a simulation test size of 100 samples.

Methods	MSPE	MS-O	MS-N	FS-O	TS-O
piMoM(MAP)	24.281 (9.01)	5.05 (0.22)	0.01 (0.10)	5	6
piMoM(LS)	24.265 (9.04)	5.05 (0.22)	0.01 (0.10)	5	6
peMoM(MAP)	**24.156** (9.02)	5.02 (0.14)	**0.00** (0.00)	5	6
peMoM(LS)	24.165 (9.00)	5.02 (0.14)	**0.00** (0.00)	5	6
g-prior	26.314 (9.87)	3.10 (0.44)	**0.00** (0.00)	3	5
Lasso	30.243 (11.82)	5.07 (0.87)	7.77 (11.16)	4	8
Scad	33.993 (10.66)	5.39 (0.57)	31.60 (28.28)	5	7
MCP	26.191 (9.87)	4.66 (0.74)	0.54 (1.04)	3	6
Marginal	26.612 (10.16)	3.74 (0.88)	0.41 (0.72)	3	7
Joint	26.385 (10.25)	3.77 (0.94)	0.02 (0.20)	3	6

Table 2.3: The Boston Housing data set: MS-O and MS-N refer to the average number of selected original variables and selected noise variables, respectively. FS-O is the number of original variables that are frequently selected at least 95 times out of 100 repetitions. TS-O refers to the number of original variables selected at least once from 100 repetitions.

The results of are analysis are summarized in Table 2.3. The conclusions are similar to those reported in Section 8.1; the nonlocal priors consistently choose more parsimonious models and had better predictive performance. The model selection procedure resulting from the nonlocal prior selects almost the same variables across the 100 repetitions. The average number of the original variables selected more than 95 times over 100 repetitions is 5, which is close to the average model size. It is also reliable in the sense that the average number of the original variables that are selected at least once across the repetitions is only 6. This means that model selection based on the nonlocal prior selects the same model in most data splits. On the other hand, penalized likelihood methods such as Lasso and Scad tend to select a large number of noise variables.

2.8 Conclusion

This dissertation described theoretical properties of peMoM and piMoM priors for variable selection in ultrahigh-dimensional linear model settings. In terms of identifying a "true" model, selection procedures based on peMoM priors are asymptotically equivalent to piMoM priors in Johnson and Rossell (2012) because they share the same kernel, $\exp\{-\tau_{n,p}/\beta^2\}$. I demonstrated that model selection procedures based on peMoM priors and piMoM priors achieve model selection consistency in $p \gg n$ settings.

In Section 2.3.1, precision-recall curves were used to show that the model selection procedure based on a g-prior can achieve nearly the same performance in identifying the MAP model as nonlocal priors when an optimal value for the hyperparameter g is chosen. However, as shown in Section 2.3.2, the value of the hyperparameter that maximizes the posterior probability of the true model is very large and has high variability, which may limit the practical application of this method. To overcome this problem, one can consider mixtures of g-prior as in Liang et al. (2008), but the asymptotic behavior of Bayes factor and model selection consistency in ultrahigh-dimensional settings have not been examined for hyper-g priors, and they are difficult to implement computationally.

In Section 5.5.3, I proposed an efficient and scalable model selection algorithm called S5. By incorporating the SSS with a screening idea and a temperature control, S5 was able to accelerate the computation speed without losing the capacity to explore the interesting region in the model space. Under some simulation settings, it outperformed the SSS in a sense that not only did S5 search the MAP model much faster than the SSS, but it also found exactly the same MAP model that was identified by the SSS.

Because the explicit form of the marginal likelihood of the nonlocal priors is not available, I used the Laplace approximation throughout this chapter. Barber et al. (2016) studied the accuracy of the approximation in Bayesian high-dimensional variable selection, especially when

the dimension of the approximation (which is q_n) and n are both increasing. However, their results do not apply to the case of the nonlocal priors, since the nonlocal priors violate their regularity condition (nonzero density at the origin). While empirical results in this chapter and Johnson and Rossell (2012) suggest that the use of the Laplace approximation is reasonable, in future work it is still worth paying attention to the approximation error of the Laplace approximation to the marginal likelihood of the nonlocal priors.

The close connection between my methods and the reduced rLasso procedures provides a useful contrast between Bayesian and penalized likelihood methods for variable selection procedures. According to the evaluation criteria proposed in Section 2.5, the two classes of methods appear to perform quite similarly. A potential advantage of the reduced rLasso procedure, and to the lesser extent the rLasso procedure, is reduced computation cost. This advantage accrues primarily because the reduced rLasso can be computed from the least squares estimate of each model's regression parameter, whereas the Bayesian procedures require numerical optimization to obtain the maximum a posteriori estimate used in the evaluation of the Laplace approximation to the marginal density of each model visited. However, the procedures used to search the model space, given the value of a marginal density or objective function, are approximately equally complex for both classes of procedures. There are also potential advantages of the Bayesian methods. For example, it is possible to approximate the normalizing constant of the posterior model probability from the models visited by S5 algorithm, and to use this normalizing constant to obtain an approximation to the posterior probability assigned to each model. In so doing, the Bayesian procedures provide a natural estimate of uncertainty associated with model selection. These posterior model probabilities can also be used in Bayesian modeling averaging procedures, which have been demonstrated to improve prediction accuracy (e.g., Raftery et al. (1997)) over prediction procedures based on maximum a posteriori estimates. Finally, the availability of prior densities may prove useful in setting model hyperparameters (i.e., $\tau_{n,p}$) in actual applications, where scientific knowledge is typically available

to guide the definition of the magnitude of substantively important regression parameters.

I also developed an `R` package `BayesS5` that provides all computational functions used in this dissertation, including a support of parallel computing environments. It is available on the author's website and on `CRAN` (`https://cran.r-project.org`).

3. SIMPLIFIED SHOTGUN STOCHASTIC SEARCH WITH SCREENING ALGORITHM FOR HIGH-DIMENSIONAL BAYESIAN MODEL SELECTION

3.1 Introduction

In $p \gg n$ settings, full posterior sampling using existing Markov chain Monte Carlo (MCMC) algorithms is highly inefficient and often not feasible from a practical perspective. Due to this limitation, several deterministic approaches to find the maximum a posteriori (MAP) model have been proposed, e.g. Carbonetto and Stephens (2012), Liu and Ihler (2013) and Rockova and George (2014). However, those procedures only provide a single model without considering uncertainty on the model space. This lack of assessment of model uncertainty can be problematic, particularly if one wishes to average over models to improve prediction performance (Raftery et al., 1997). To overcome this issue and approximate full posterior model probabilities, I propose a scalable stochastic search algorithm aimed at rapidly identifying regions of high posterior probability and finding the MAP) model for linear model selection problems. My main innovation is to develop a stochastic search algorithm combining isis-like screening techniques (Fan and Lv, 2008) and temperature control procedure similar to those used in global optimization algorithms like simulated annealing (Kirkpatrick and Vecchi, 1983).

To describe my proposed algorithm, note that the MAP model $\widehat{\mathbf{k}}$ that can be expressed as

$$\widehat{\mathbf{k}} = \underset{\mathbf{k} \in \Gamma^*}{\operatorname{argmax}} \{\pi(\mathbf{k} \mid y)\}, \tag{3.1}$$

where Γ^* is the set of all models assigned non-zero prior probability.

3.2 Shotgun Stochastic Search Algorithm (SSS)

Hans et al. (2007) proposed the shotgun stochastic search (SSS) algorithm in an attempt

to efficiently navigate through very large model spaces and identify global maxima. Letting $\text{nbd}(\mathbf{k}) = \{\Gamma^+, \Gamma^-, \Gamma^0\}$, where $\Gamma^+ = \{\mathbf{k} \cup \{j\} : j \in \mathbf{k}^c\}$, $\Gamma^- = \{\mathbf{k} \setminus \{j\} : j \in \mathbf{k}\}$, and $\Gamma^0 = \{[\mathbf{k} \setminus \{j\}] \cup \{l\} : l \in \mathbf{k}^c, j \in \mathbf{k}\}$, the SSS procedure is described in **Algorithm 1**.

Algorithm 1 Shotgun Stochastic Search (SSS)

Choose an initial model $\mathbf{k}^{(1)}$
For $i = 1$ to $i = N - 1$
  Compute $\pi(\mathbf{k} \mid y)$ for all $\mathbf{k} \in \text{nbd}(\mathbf{k}^{(i)})$
  Sample $\mathbf{k}^+$, $\mathbf{k}^-$, and $\mathbf{k}^0$, from Γ^+, Γ^-, and Γ^0, with probabilities proportional to $\pi(\mathbf{k} \mid y)$
  Sample $\mathbf{k}^{(i+1)}$ from $\{\mathbf{k}^+, \mathbf{k}^-, \mathbf{k}^0\}$, with probability proportional to
   $\{\pi(\mathbf{k}^+ \mid y), \pi(\mathbf{k}^- \mid y), \pi(\mathbf{k}^0 \mid y)\}$

The estimated MAP model is defined as the model that achieves the largest posterior probability among those searched models only.

3.3 Simplified Shotgun Stochastic Search Algorithm with Screening (S5)

SSS is effective in exploring regions of high posterior model probability, but its computational cost is still expensive because it requires the evaluation of marginal probabilities for models in Γ^+, Γ^-, and Γ^0 at each iteration. The largest computational burden occurs for the evaluation of marginal likelihood for models in Γ^0, since $|\Gamma^0| = |\mathbf{k}|(p - |\mathbf{k}|)$. To improve the computational efficiency of SSS, I propose a modified version which only examines models in Γ^+ and Γ^-. These sets have cardinality $p - |\mathbf{k}|$ and $|\mathbf{k}|$, respectively. However, ignoring Γ^0 in the sampling updates can make the algorithm less likely to explore "interesting" regions of high posterior model probability. The algorithm would therefore be more likely to get stuck in local maxima. To counter this problem, I introduce a "temperature parameter" analogous to simulated annealing that allows the algorithm to explore a broader spectrum of models.

Even though ignoring models in Γ^0 reduces the computational burden of the SSS algorithm, the calculation of p posterior model probabilities in every iteration is still computationally prohibitive when p is very large. To further reduce the computational burden, I borrow ideas from the Iterative Sure Independence Screening (isis; Fan and Lv (2008)) and consider only those variables that have a large correlation with the residuals of the current model. More precisely, I examine the products $|r_{\mathbf{k}}^T X_j|$, where $r_{\mathbf{k}}$ is the residual of the model $\mathbf{k}$, for $j = 1, \ldots, p$, after every iteration of the modified shotgun stochastic search algorithm, and then restrict attention to variables for which $\{|r_{\mathbf{k}}^T X_j| : j = 1, \ldots, p\}$ is large (I assume that the columns of X have been standardized). This yields a scalable algorithm even when the number of variables p is large.

With these ingredients, I propose a new stochastic model search algorithm called Simplified Shotgun Stochastic Search with Screening (S5). This algorithm is described in **Algorithm 2**.

Algorithm 2 Simplified Shotgun Stochastic Search with Screening (S5)

Set a temperature schedule $t_1 > t_2 > \ldots > t_L > 0$
Choose an initial model $\mathbf{k}^{(1,1)}$ and a set of variables after screening $\mathbf{S}_{\mathbf{k}^{(1,1)}}$ based on $\mathbf{k}^{(1,1)}$
For $l = 1$ in $l = L$
  For i in $1, \ldots, J-1$
    Compute all $\pi(\mathbf{k} \mid y)$ for all $\mathbf{k} \in \text{nbd}_{scr}(\mathbf{k}^{(i,l)})$
    Sample $\mathbf{k}^+$ and $\mathbf{k}^-$, from Γ_{scr}^+ and Γ^-, with probabilities proportional to $\pi(\mathbf{k} \mid y)^{1/t_l}$
    Sample $\mathbf{k}^{(i+1,l)}$ from $\{\mathbf{k}^+, \mathbf{k}^-\}$, with probability proportional to
    $\{\pi(\mathbf{k}^+ \mid y)^{1/t_l}, \pi(\mathbf{k}^- \mid y)^{1/t_l}\}$
    Update the set of considered variables $\mathbf{S}_{\mathbf{k}^{(i+1,l)}}$ to be the union of variables in $\mathbf{k}^{(i+1,l)}$ and the top M_n variables according to $\{|r_{\mathbf{k}^{(i+1,l)}}^T X_j| : j = 1, \ldots, p\}$

In S5, $\mathbf{S}_{\mathbf{k}}$ is the union of variables in $\mathbf{k}$ and the top M_n variables obtained by screening using the residuals from model $\mathbf{k}$. The screened neighborhood of model $\mathbf{k}$ can be defined as $\text{nbd}_{scr}(\mathbf{k}) = \{\Gamma_{scr}^+, \Gamma^-\}$, where $\Gamma_{scr}^+ = \{\mathbf{k} \cup \{j\} : j \in \mathbf{k}^c \cap \mathbf{S}_{\mathbf{k}}\}$.

Even though this algorithm is designed to identify the MAP model, it also provides an approximation to the posterior model probability of each sampled model. The uncertainty of the model space can be measured by approximating the normalizing constant from the (unnormalized) posterior probabilities of the models explored by the algorithm.

Denoting the computational complexity of the evaluation of the unnormalized posterior model probability of the largest model among searched models by E_n, the computational complexity of the SSS algorithm can be expressed as the product of the number of models explored by the algorithm and E_n, which is $[O\{Np\} + O\{Nq_n\} + O\{N(p-q_n)q_n\}] \times E_n$, where q_n is the maximum size of model among searched models and $q_n < n \ll p$.

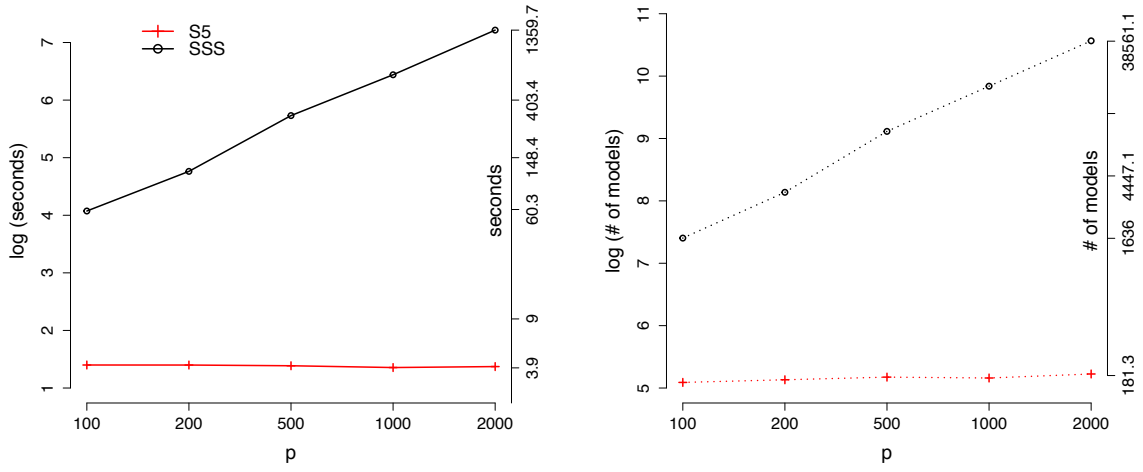
S5 only considers M_n variables after the screening step in each iteration, which dramatically reduces the number of models to be considered in constructing the neighborhood, $O\{JL(M_n - q_n)\} + O(JLM_n)$. Therefore, the resulting computational complexity is

$$[O\{JL(M_n - q_n)\} + O(JLM_n)] \times E_n + O(JLnp),$$

where $q_n < M_n$. When the computational complexity for screening steps, $O(JLnp)$, is dominated by the other terms, the computational complexity is almost independent of p. As a result, the proposed algorithm is scalable in the sense that the resulting computational complexity is typically robust to the size of p.

3.4 Performance Comparisons Between S5 and SSS

I examined the computational efficiency of S5 to SSS in identifying the MAP model under a piMOM prior with $\tau_{n,p} = \log n \log p$ and $r = 1$. I generated data according to Case (1) in Section 2.3 with a fixed sample size ($n = 200$) and a varying number of covariates p. I set $M_n = 20$, $L = 20$, and $J = 20$ for S5. To match the total number of iterations between S5 and SSS, I set $N = 400$ for SSS. All computations were implemented in `R` on a machine containing 16 CPU cores (Intel(R) Xeon(R) CPU E5-2690 @ 2.90GHz with 64GB of DDR3



(a) (b)

Figure 3.1: (a) Average computation time to first find the MAP model; (b) Average number of models searched before hitting the MAP model. The left y-axis is on a logarithmic scale and the right y-axis is on the raw scale.

@ 1600Mhz).

Figure 3.1 shows the average computation time and the number of models searched before hitting the MAP model for the first time for the S5 and SSS algorithms. All averages were based on 100 simulated datasets, and both algorithms found the same MAP model in all data sets. Panel (a) shows that the computation time of SSS increases roughly at a p^2 rate, but that the computation time for S5 was nearly independent of the number of covariates p (about 4 seconds). For example, when $p = 2,000$, SSS first found the MAP model in an average of 1,360 seconds (about 23 minutes), whereas S5 hit the MAP model after about only 4 seconds. Interestingly, panel (b) of Figure 3.1 shows that the S5 algorithms explored only 181 models on average before finding the MAP model, whereas SSS typically visited slightly more than 38,000 models. Thus, not only is S5 much faster than SSS in identifying the MAP model, but it also visited far fewer models before visiting the MAP model.

To see how sensitive the efficiency of the S5 algorithm is to the choice of the screening set

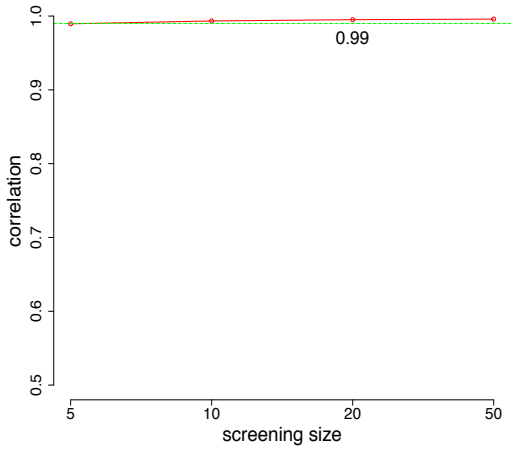


Figure 3.2: Correlation between the top 10 posterior model probabilities estimated from SSS and S5 with different screening set sizes.

size, Figure 3.2 reports the average correlation between the top 10 posterior model probabilities approximated from S5 and SSS with varying screening set sizes. This figure shows that even when the screening set size is small in comparison to the true model size ($|\mathbf{t}| = 5$), the correlation of the top 10 posterior model probabilities from S5 and SSS is at least 0.99 (the horizontal green line in the figure is located on 0.99). Thus, the resulting posterior model probabilities are almost same as those of SSS. At least in this example, S5 is not sensitive to the choice of the screening set size. However, for real data sets, I recommend examining output from multiple screening set sizes.

3.4.1 Application to Real Data Examples

In this subsection, I apply the S5 algorithm to Bardet-Biedl syndrome gene expression data that was first reported in Scheetz et al. (2006). The data set contains microarrays expression values from eye tissue of 120 twelve-week old male rats. a total of 31,042 different probe sets were used to analyze the RNA values from the tissue. The intensity values were normalized using the robust multi-chip averaging method (Irizarry et al., 2003). This microarray data set

has been considered in multiple papers, including Huang et al. (2008), Kim et al. (2008) and Fan et al. (2011). As in those papers I am interested in finding a subset of the probe sets that are associated with the probe set is *1389163_at*, which corresponds to the expression of gene *TRIM32*. This gene is related to Bardet-Biedl syndrome, a hereditary disease of the retina. The data set was first ranked all other probes according to the absolute value of the marginal correlation to *1389163_at* and selected the top 200 probes ($n = 120$ and $p = 200$). The screened data set is available in the `R` package `flare`.

I also considered a simulated data set based on the Boston housing data set that was used in Section 2.7.2 by adding 1000 spurious variables to the original Boston housing data set ($n = 506$ and $p = 1010$).

For S5 and SSS, the settings usedin the previous simulation study section were again used. I repeatedly ran S5 and SSS for 30 replicates starting from different initial models. Table 3.1 reports the average time of the computation (Time) over 30 replicates, the logarithm of the (unnormalized) posterior probability of the MAP model found by each algorithm (Log-post), and the average number of models searched to find the MAP model by each algorithm (Avg.#models). S5 found exactly the same MAP model searched by SSS for each data set, and the computation time of S5 is much shorter than SSS. For the Bardet-Biedl syndrome data, S5 is 29 times faster than SSS and 84 times faster for the Boston housing data. Moreover, S5 finds the MAP model after visiting far fewer models than SSS. This shows that S5 very efficiently explores the model space.

	Bardet-Biedl Syndrome Data			Boston Housing Data		
Method	Time (sec)	Log-post	Avg.#models	Time (sec)	Log-post	Avg.#models
S5	61.4	199.746	1198.0	54.6	-1115.34	442.9
SSS	1767.4	199.746	21423.8	4569.8	-1115.34	40406.9

Table 3.1: Comparisons between S5 and SSS using the Bardet-Biedl syndrome data and the Boston housing data.

3.5 R Package: BayesS5

In this section, I provide tutorials about how to use an R package called BayesS5 (Shin and Tian, 2017) to implement the S5 algorithm for high-dimensional Bayesian model selection problems. This package is available from CRAN (https://cran.r-project.org). In the following subsections, I provide examples that demonstrate how to use the package in several model selection problems.

To illustrate the use of the package, I consider the Boston data set that was used in Section 2.7.2. To examine high-dimensional settings, I added 500 spurious variables to the original data set, so that $p = 510$ and $n = 506$. The following code imports the data set in R.

```
R> library(BayesS5); library(MASS)
R> data(Boston); attach(Boston)
R> X = cbind(crim,indus,nox,rm,age,dis,tax,ptratio,black,lstat)
R> X = scale(X)
R> y = medv; y = y-mean(y)
R> n = nrow(X)
R> set.seed(291287)
R> X = cbind(X,matrix(rnorm(500*n),n,500)); X = scale(X)
R> p = ncol(X)
```

3.5.1 S5 Function

Without specifying the priors, the simplest implementation with the default setting can be conducted with the following command:

```
R> fit_default = S5(X,y)
```

The default setting is the piMOM prior in (2.4) for regression coefficients and the beta-uniform prior in (2.12). The hyperparameter value of the piMOM prior is automatically chosen

by the procedure proposed in Nikooienejad et al. (2016). The default setting for iterations is $M_n = 20$, $L = 20$, and $J = 20$, and the default choice of temperature schedule is the square inverse of the equi-spaced sequence from 0.4 to 1 with size 20. For every transition between temperatures, the function prints out the current status of the model selection. An example of this output looks like this:

```
[1] "################################"
[1] "Inverse Temperature"
[1] 0.16
[1] "The Selected Variables in the Searched MAP Model"
[1]  3  4  6  8 10
[1] "The Evaluated Object Value at the Searched MAP Model"
[1] -1111.231
[1] "Current Model"
[1]  4  8 10
[1] "The Evaluated Object Value at the Current Model"
[1] -1118.82
[1] "The Number of Total Searched Models"
[1] 341
```

During the run, the `S5` function outputs the inverse of the current temperature used in the algorithm, and it provides the MAP model and its (unnormalized) log-posterior model probability, $\log m_{\mathbf{k}}(\mathbf{y}) + \log \pi(\mathbf{k})$. For example, in the above output, the model of 3, 4, 6, 8 and 10 indicates the model defined by the third, fourth, sixth, eighth and tenth covariates, and -1111.231 is the (unnormalized) log-posterior model probability of the model $\{3, 4, 6, 8, 10\}$. The `"Current Model"` is the model that the algorithm is currently visiting. Its log-posterior model probability is also listed. The final line of the output is the number of models that have been searched by the algorithm.

The object of `S5` (e.g. `fit_default` in the above code) contains the list of the searched models that are identified by binary vectors and their (unnormalized) log-posterior model probabilities. To approximate the full posterior model probabilities of each model, one extra step is required:

```
R> res_default = result(fit_default)
[1] "# of Searched Models by S5"
[1] 1291
[1] "The MAP model is "
[1]  3  4  6  8 10
[1] "with posterior probability 0.739"
```

The result shows that the MAP model is $\{3, 4, 6, 8, 10\}$ and its posterior model probability is 73.9%. In this case, the selected model does not include any of the 500 spurious variables that are generated independently from the response variable. The total number of models explored by S5 was 1,291. The `result` function also provides the marginal inclusion probabilities for each variable. These probabilities are defined as

$$q_j = \sum_{\mathbf{k}: j \in \mathbf{k}} \pi(\mathbf{k} \mid \mathbf{y}),$$

for $j = 1, \ldots, p$. The command to generate these values is given by

```
R> mar_default = res_default$marg.prob
R> print(which(mar_default>0.5))
[1]  3  4  6  8 10
R> plot(mar_default, ylim=c(0,1), xlab="covariate index",
 ylab="marginal inclusion prob", pch=3)
```

Figure 3.3 is the output of the above command presenting the marginal inclusion probabil-

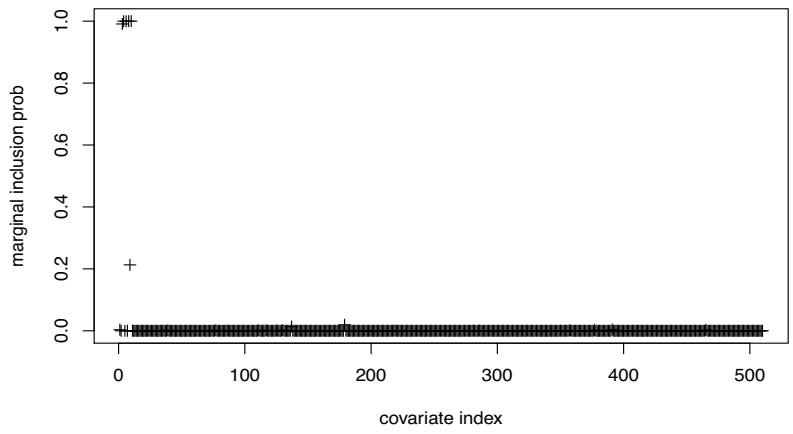


Figure 3.3: Marginal inclusion probabilities approximated by S5 for the synthesized Boston housing data set.

ities.

I note that full posterior model probabilities searched by the S5 algorithm can be calculated by using the `result` function. The below code provides the top three models that have highest posterior model probabilities.

```
R> gam_default = res_default$gam
R> post_default = res_default$post
R> round(post_default[1:3], 3)
[1] 0.739 0.192 0.012
R> which(gam_default[,1] == 1)
[1]  3  4  6  8 10
R> which(gam_default[,2] == 1)
[1]  3  4  6  8  9  10
R> which(gam_default[,3] == 1)
[1]  3  4  6  8  9  10  179
```

In the above code, `gam_default` contains the binary vectors that identify the corresponding models (1 indicates the corresponding variable is in the model, and 0 means it is not in the model). The object `post_default` stores the posterior probabilities of models identified by binary vectors in `gam_default`. As shown in the previous code, the MAP model is $\{3, 4, 6, 8, 10\}$ with 73.9% posterior probability, and the second most significant model is $\{3, 4, 6, 8, 9, 10\}$ with probability 19.2%. The third highest posterior model probability model is $\{3, 4, 6, 8, 10, 179\}$ with 1.2% posterior probability. This model includes a spurious variable X_{179}. The S5 algorithm searched a total $1,291$ models; the posterior probabilities of models not visited are approximate by 0.

The `S5` package also provides other priors for Bayesian model selection procedure. These includes the peMOM priors in (2.2) and Zellner's g-prior. For example, the g-prior can be applied to S5 by the following code:

```
R> tuning = p^2 # tuning parameter g for g-prior
R> ind_fun = ind_fun_g # choose g-prior for the regression coef
R> model = Uniform #choose the uniform model prior
R> fit_g = S5(X,y,ind_fun=ind_fun,model=model,tuning=tuning)
```

3.5.2 `S5_parallel` Function for Parallel Computing Environments

The S5 algorithm is efficient and fast in exploring the model space. However, it may not be fast enough to implement in practice when the data set is high-dimensional and variables are highly correlated. To overcome this problem, it is reasonable to use multiple independent chains to search the model space. The S5 parallel function permits S5 to be ran in parallel computing environments. The following command can be used to implement the parallel version of S5 using 20 cores.

```
R> NC = 20 # the number of cores that will be used
R> fit_parallel = S5_parallel(NC=NC, X, y)
R Version:  R version 3.3.3 (2017-03-06)
snowfall 1.84-6.1 initialized (using snow 0.4-2):
parallel execution on 20 CPUs.
Library Matrix loaded.
Library Matrix loaded in cluster.
   user  system elapsed
  0.090   0.004 153.129
Stopping cluster
R> res_parallel = result(fit_parallel)
[1] "# of Searched Models by S5"
[1] 6840
[1] "The MAP model is "
[1]  3  4  6  8 10
[1] "with posterior probability 0.736"
```

In the single processor version of S5, $1,291$ models were visited. In this 20 CPU application, $6,840$ models were visited. This is more than five times the number of visited models using the same amount of real time. The MAP model found by the parallel version is exactly the same with the MAP model by the standard S5, and its posterior probability is 73.6%; this is slightly smaller than the 73.9% that was estimated from the single chain. The code to extract these results follows:

```
R> gam_parallel = res_parallel$gam
R> post_parallel = res_parallel$post
R> round(post_parallel[1:3], 3)
[1] 0.736 0.191 0.012
```

```
R> which(gam_parallel[,1] == 1)
[1]  3  4  6  8 10
R> which(gam_parallel[,2] == 1)
[1]  3  4  6  8  9  10
R> which(gam_parallel[,3] == 1)
[1]  3  4  6  8  9  10  179
```

4. FUNCTIONAL HORSESHOE PRIOR FOR NONPARAMETRIC SUBSPACE SHRINKAGE

4.1 Introduction

Since the seminal work of James and Stein (1961), shrinkage estimation has been immensely successful in various statistical disciplines and continues to enjoy widespread attention. Many shrinkage estimators have a natural Bayesian flavor. For example, one obtains the ridge regression estimator as the posterior mean arising from an isotropic Gaussian prior on the vector of regression coefficients (Jeffreys, 1961; Hoerl and Kennard, 1970). Along similar lines, an empirical Bayes interpretation of the (positive part) James–Stein estimator can be obtained (Efron and Morris, 1973). Such connections have been extended to the semiparametric regression context, with applications to smoothing splines and penalized splines (Wahba, 1990; Ruppert et al., 2003). Over the past decade and a half, a number of second-generation shrinkage priors have appeared in the literature for application in high-dimensional sparse estimation problems. Such priors can be almost exclusively expressed as global-local scale mixtures of Gaussians (Polson and Scott, 2010a); examples include the relevance vector machine (Tipping, 2001), normal/Jeffrey's prior (Bae and Mallick, 2004), the Bayesian Lasso (Park and Casella, 2008; Hans, 2009), the horseshoe priors (Carvalho et al., 2010), normal/gamma and normal/inverse-Gaussian priors (Caron and Doucet, 2008; Griffin and Brown, 2010), generalized double Pareto priors (Armagan et al., 2013) and Dirichlet–Laplace priors (Bhattacharya et al., 2015). These priors typically have a large spike near zero with heavy tails, thereby providing an approximation to the operating characteristics of sparsity inducing discrete mixture priors (George and McCulloch, 1997; Johnson and Rossell, 2012). For more on connections between Bayesian model averaging and shrinkage, refer to Polson and Scott (2010a).

A key distinction between ridge-type shrinkage priors and the global-local priors is that

while ridge-type priors typically shrink towards a fixed point–most commonly the origin– the global-local priors shrink towards the union of subspaces consisting of sparse vectors. The degree of shrinkage to sparse models is controlled by certain hyperparameters (Bhattacharya et al., 2015). In this dissertation, I further enlarge the scope of shrinkage prior by imposing a class of functional shrinkage priors, called the functional horseshoe priors (fHS). fHS priors facilitate shrinkage towards pre-specified subspaces. The shrinkage factor (defined in Section 3) is assigned a Beta(a, b) prior with $a, b < 1$, which has the shape of a horseshoe prior (Carvalho et al., 2010). While the horseshoe prior shrinks towards sparse vectors, the proposed fHS prior enforces functions to shrink towards arbitrary subspaces.

To illustrate the proposed methodology, consider a nonparametric regression model with unknown regression function $f : \mathcal{X} \to \mathbb{R}$ given by

$$Y = F + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n), \tag{4.1}$$

where $Y = \{y_1, \ldots, y_n\}^{\mathrm{T}}$, $F = \{f(x_1), \ldots, f(x_n)\}^{\mathrm{T}} = \mathbb{E}(Y \mid \mathbf{x})$, and the covariates $x_i \in \mathcal{X} \subset \mathbb{R}$.

In (4.1), one can either make parametric assumptions (e.g., linear or quadratic dependence on x) regarding the shape of f, or one may model it nonparametrically using splines, wavelets, Gaussian processes, etc. Scatter plots or goodness of fit tests can be used to ascertain the validity of a linear or quadratic model in (4.1), but such procedures are only feasible in relatively simple settings. In complex and/or high dimensional problems, there is clearly a need for an automatic data-driven procedure to adapt between models of varying complexity. With this motivation, the fHS priors encourage shrinkage towards a parametric class of models embedded inside a larger semiparametric model, as long as a suitable projection operator can be defined. For example, in (4.1), f will be shrunk towards a linear or quadratic function if such parametric assumptions are supported by the data, and will remain unshrunk otherwise. As

noted already, my approach is not limited to the univariate regression context and can be extended to the varying coefficient model (Hastie and Tibshirani, 1993), density estimation via log-spline models (Kooperberg and Stone, 1991) and additive models (Hastie and Tibshirani, 1986), among others. Further details are provided in Section 4.4. In the additive regression context, the proposed approach performs well compared to state-of-the-art procedures like *Sparse Additive Model* (SpAM) of Ravikumar et al. (2009) and *High-dimensional Generalized Additive Model* (HGAM) by Meier et al. (2009).

I provide theoretical justification for the method by showing an adaptive property of the approach in the context of (4.1). Specifically, I show that the posterior contracts at the parametric rate if the true function belongs to the pre-designated subspace, and contracts at the optimal rate for α-smooth functions otherwise. In other words, my approach adapts to the parametric shape of the unknown function while allowing deviations from the parametric shape in a nonparametric fashion.

4.2 Preliminaries

I begin by introducing some notation. For $\alpha > 0$, let $\lfloor \alpha \rfloor$ denote the largest integer smaller than or equal to α and $\lceil \alpha \rceil$ denote the smallest integer larger than or equal to α. Let $C^{\alpha}[0,1]$ denote the Hölder class of α smooth functions on $[0,1]$ that have continuously differentiable derivatives up to order $\lfloor \alpha \rfloor$, with the $\lfloor \alpha \rfloor$th order derivative being Lipschitz continuous of order $\alpha - \lfloor \alpha \rfloor$. For a vector $x \in \mathbb{R}^d$, let $\|x\|$ denote its Euclidean norm. For a function $g : [0,1] \to \mathbb{R}$ and points $x_1, \ldots, x_n \in [0,1]$, let $\|g\|_{2,n}^2 = n^{-1} \sum_{i=1}^n g^2(x_i)$; I shall refer to $\|\cdot\|_{2,n}$ as the empirical L_2 norm. For an $m \times d$ matrix A with $m > d$ and $\mathrm{rk}(A) = d$, let $\mathfrak{L}(A) = \{A\beta : \beta \in \mathbb{R}^d\}$ denote the column space of A, which is a d-dimensional subspace of $\mathbb{R}^m$. Let $\mathrm{Q}_A = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ denote the projection matrix on $\mathfrak{L}(A)$.

4.3 Functional Horseshoe Prior

In the nonparametric regression model in (4.1), I model the unknown function f as spanned by a set of pre-specified basis functions $\{\phi_j\}_{1\leq j\leq K_n}$ as follows:

$$f(x) = \sum_{j=1}^{K_n} \beta_j \phi_j(x). \tag{4.2}$$

I work with B-spline basis functions (De Boor, 1978) for illustrative purpose here. However, the methodology generalizes to a larger class of basis functions. The details about B-spline basis functions were described in Section 1.1.2. Letting $\beta = \{\beta_1, \ldots, \beta_{K_n}\}^{\mathrm{T}}$ denote the vector of basis coefficients and $\Phi = \{\phi_j(X_i)\}_{1\leq i\leq n, 1\leq j\leq K_n}$ denote the $n \times K_n$ matrix of B-spline basis functions evaluated at the observed covariates. Model (4.1) can then be expressed as

$$Y \mid \beta \sim \mathrm{N}(\Phi\beta, \sigma^2 \mathrm{I}_n). \tag{4.3}$$

A standard choice for a prior on β is a g-prior $\beta \sim \mathrm{N}(0, g(\Phi^{\mathrm{T}}\Phi)^{-1})$ (Zellner, 1986). The g-priors have been commonly used in linear models since they incorporate the correlation structure of the covariates inside the prior variance. The posterior mean of β with a g-prior can be expressed as $\{1 - 1/(1+g)\}\widehat{\beta}$, where $\widehat{\beta} = \mathrm{Q}_\Phi Y$ is the maximum likelihood estimate of β. Thus, the posterior mean shrinks the maximum likelihood estimator towards zero, with the amount of shrinkage controlled by the parameter g. Bontemps (2011) studied asymptotic properties of the resulting posterior by providing bounds on the total variation distance between the posterior distribution and a Gaussian distribution centered at the maximum likelihood estimator with the inverse Fisher information matrix as covariance. In his work, the g parameter was fixed *a priori* depending on the sample size n and the error variance σ^2. The results in particular imply minimax optimal posterior convergence for α-smooth functions. Among related work, Ghosal and van der Vaart (2007) established minimax optimality with isotropic Gaussian

priors on β.

My goal is to define a broader class of shrinkage priors on β that facilitate shrinkage towards a *null subspace* that is fixed in advance, rather than shrinkage towards the origin or any other fixed a priori guess β_0. For example, if I have *a priori* belief that the function is likely to attain a linear shape, then I would like to impose shrinkage towards the class of linear functions. In general, my methodology allows shrinkage towards any null subspace spanned by the columns of a null regressor matrix Φ_0, with $d_0 = \text{rank}(\Phi_0)$ equal to the dimension of the null space. For example in the linear case, I define the null space as $\mathfrak{L}(\Phi_0)$ with $\Phi_0 = \{\mathbf{1}, \mathbf{x}\} \in \mathbb{R}^{n\times 2}$, where $\mathbf{1}$ is a $n \times 1$ vector of ones and $d_0 = 2$. Shrinkage towards quadratic, or more generally polynomial, regression models are achieved similarly.

With the above ingredients, I propose the fHS prior through the following conditional specification:

$$\pi(\beta \mid \tau) \propto (\tau^2)^{-(K_n - d_0)/2} \exp\left\{-\frac{1}{2\sigma^2\tau^2}\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta\right\}, \tag{4.4}$$

$$\pi(\tau) \propto \frac{(\tau^2)^{b-1/2}}{(1+\tau^2)^{(a+b)}}\mathbf{1}_{(0,\infty)}(\tau), \tag{4.5}$$

where $a, b > 0$. Recall that $\mathrm{Q}_0 = \Phi_0(\Phi_0^{\mathrm{T}}\Phi_0)^{-1}\Phi_0^{\mathrm{T}}$ denotes the projection matrix of Φ_0.

When $\Phi_0 = 0$, (4.4) is equivalent to a g-prior with $g = \tau^2$. The key additional feature in my proposed prior is he introduction of the quantity $(\mathrm{I} - \mathrm{Q}_0)$ in the exponent, which enables shrinkage towards subspaces rather than a single point. Although the proposed prior may be singular, it follows from subsequent results that the joint posterior of (β, τ^2) is proper. Note that the prior on the scale parameter τ follows a half-Cauchy distribution when $a = b = 1/2$. Half-Cauchy priors have been recommended as a default prior choice for global scale parameters in the linear regression framework (Polson and Scott, 2012). Using the reparameterization $\omega = 1/(1+\tau^2)$, the prior in (4.5) can be interpreted as the prior induced on τ through a Beta(a, b) prior on ω. I work in the ω parameterization for reasons to be evident shortly.

Exploiting the conditional Gaussian specification, the conditional posterior of β is also Gaussian, and can be expressed as

$$\beta \mid Y, \omega \sim \mathrm{N}(\widetilde{\beta}_\omega, \widetilde{\Sigma}_\omega), \tag{4.6}$$

where

$$\widetilde{\beta}_\omega = \left(\Phi^{\mathrm{T}}\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi \right)^{-1} \Phi^{\mathrm{T}}Y, \quad \widetilde{\Sigma}_\omega = \sigma^2 \left(\Phi^{\mathrm{T}}\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi \right)^{-1}. \tag{4.7}$$

I now state a lemma which delineates the role of ω as the parameter controlling the shrinkage.

Lemma 1. *Suppose that* $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$. *Then,*

$$\mathbb{E}\left[\Phi\beta \mid Y, \omega\right] = \Phi\widetilde{\beta}_\omega = (1-\omega)\mathrm{Q}_\Phi Y + \omega \mathrm{Q}_0 Y,$$

where Q_Φ *is the projection matrix of* Φ.

The above lemma shows that the conditional posterior mean of the regression function given ω is a convex combination of the classical B-spline estimator $\mathrm{Q}_\Phi Y$ and the parametric estimator $\mathrm{Q}_0 Y$. The parameter $\omega \in (0, 1)$ controls the shrinkage effect; the closer ω is to 1, the greater the shrinkage towards the parametric estimator. I learn the parameter ω from the data with a Beta(a, b) prior on ω. The hyperparameter $b < 1$ controls the amount of prior mass near one.

Figure 4.1 illustrates the connection between the choice of the hyperparameters a and b and the shrinkage behavior of the prior. The first and the second column in Figure 4.1, with a fixed at $1/2$ shows that the prior density of ω increasingly concentrates near 1 as b decreases from $1/2$ to 10^{-1}. The third column in Figure 4.1 depicts the prior probability that $\omega > 0.95$ and

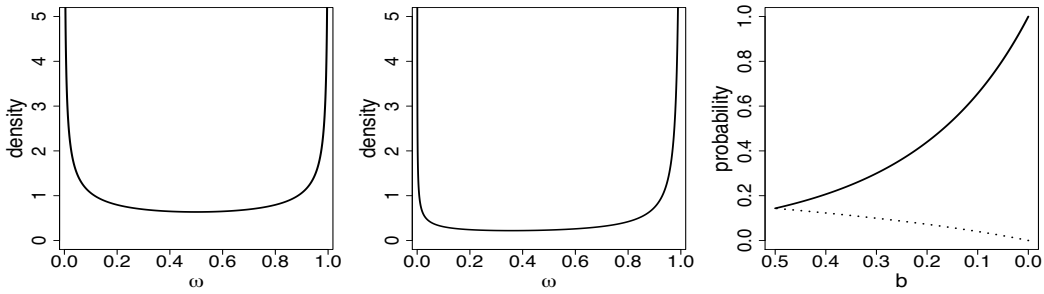


Figure 4.1: The first two columns illustrate the prior density function of ω with different hyper-parameters (a, b): $(1/2, 1/2)$ for the first column and $(1/2, 10^{-1})$ for the second column. The third column shows the prior probability that $\omega > 0.95$ (solid line) and $\omega < 0.05$ (dotted line) for varying b and a fixed $a = 1/2$.

$\omega < 0.05$. Clearly, as b decreases, the amount of prior mass around one increases, which results in stronger shrinkage towards the parametric estimator. In particular, when $a = b = 1/2$, the resulting "horseshoe" prior density derives its name from the shape of the prior on ω (Carvalho et al., 2010).

When $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$, one can orthogonally decompose $\mathrm{Q}_\Phi = \mathrm{Q}_1 + \mathrm{Q}_0$, where the columns of Q_1 are orthogonal to Q_0, i.e., $\mathrm{Q}_1^{\mathrm{T}}\mathrm{Q}_0 = 0$. For $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$, this follows because we can use Gram-Schmidt orthogonalization to create $\widetilde{\Phi} = [\Phi_0; \Phi_1]$ of the same dimension as Φ with $\Phi_1^{\mathrm{T}}\Phi_0 = 0$ and $\mathfrak{L}(\Phi) = \mathfrak{L}(\widetilde{\Phi})$. Let Q_1 denote the projection matrix on $\mathfrak{L}(\Phi_1)$. Simple algebra shows that

$$\begin{aligned}\pi(\omega \mid Y) &= \int \pi(\omega, \beta \mid Y) d\beta = \frac{\pi(\omega)}{m(Y)} \int f(Y \mid \beta, \omega) \pi(\beta \mid \omega) d\beta \\ &= \omega^{a+(K_n - d_0)/2-1} (1-\omega)^{b-1} \exp\{-H_n \omega\} / m(Y), \end{aligned} \tag{4.8}$$

where $H_n = Y^{\mathrm{T}}\mathrm{Q}_1 Y/(2\sigma^2)$ and $m(Y) = \int_0^1 \omega^{a+(K_n-d_0)/2-1}(1-\omega)^{b-1} \exp\{-H_n\omega\}\, d\omega$.

To investigate the asymptotic behavior of the resulting posterior, it is crucial to find tight two-sided bounds on $m(Y)$. Such bounds are specified in Lemma 2.

Lemma 2. *(Bounds on the normalizing constant) Let A_n and B_n be arbitrary sequences satisfying $A_n \to \infty$ as $n \to \infty$ and $B_n = O(1)$.*

Let $t_n = \int_0^1 \omega^{A_n - 1}(1-\omega)^{B_n - 1} \exp\{-H_n \omega\} d\omega$. Then,

$$\frac{\Gamma(A_n)\Gamma(B_n)}{\Gamma(A_n + B_n)} \exp\{-H_n\}(1 + Q_n^L) \leq t_n \leq \frac{\Gamma(A_n)\Gamma(B_n)}{\Gamma(A_n + B_n)} \exp\{-H_n\}(1 + Q_n^U),$$

where,

$$\begin{aligned} Q_n^U &= \frac{B_n}{A_n + B_n} \exp(H_n), \\ Q_n^L &= \frac{B_n H_n}{A_n + B_n} + \frac{D B_n (B_n + T_n)^{-A_n}}{(A_n + B_n)^{3/2}} \left(\exp\{H_n\} - 1 - H_n - (T_n + 2)^{-1/2}\right)_+, \end{aligned}$$

$T_n = \max\{A_n^2, 3\lceil H_n \rceil\}$ *and D is some positive constant.*

By setting $A_n = a + K_n/2$ and $B_n = b$, Lemma 2 shows that the magnitude of the normalizing constant $m(Y)$ in (4.8) is determined by an interplay between the relative sizes of b and $\exp(H_n)$. When b is small enough to dominate $\exp(H_n)$, $m(Y) \approx \text{Be}(a + K_n/2, b) \exp(-H_n)$, where $\text{Be}(\cdot, \cdot)$ denotes the beta function. Otherwise, ignoring polynomial terms, $m(Y) \approx Be(a + K_n/2, b)b$. This asymptotic behavior of $m(Y)$ is the key ingredient to identify the posterior contraction rate of the fHS prior. I also note that the magnitude of a does not affect the strength of shrinkage for large n as long as a is a fixed constant, since the prior contribution ω^{a-1} is dominated by the likelihood contribution $\omega^{K_n/2}$.

4.3.1 Posterior Concentration Rate

I first state a set of assumptions that have been used by others (Zhou et al., 1998; Claeskens et al., 2009) to prove minimax optimality of B-spline estimators. Assume that the following conditions hold:

(A1). Let $u = \max_{1 \leq j \leq (K_n - 1)}(t_{j+1} - t_j)$. There exists a constant $C > 0$, such that $u / \min_{1 \leq j \leq (K_n - 1)}(t_{j+1} - t_j) \leq C$ and $u = o(K_n^{-1})$.

(A2). There exists some distribution function G with a positive continuous density such that

$$\sup_{x \in [0,1]} |G_n(x) - G(x)| = o(K_n^{-1}),$$

where G_n is the empirical distribution of the covariates $\{x_i\}_{1 \leq i \leq n}$, which are assumed to be fixed by design.

Under **(A1)** and **(A2)**, Zhou et al. (1998) showed that the mean square error of the B-spline estimator $\mathrm{Q}_\Phi Y$ achieves the minimax optimal rate. If the true function $f_0 \in C^\alpha[0,1]$ is α-smooth and the number of basis functions $K_n \asymp n^{1/(2\alpha+1)}$, then Zhou et al. (1998) shows that

$$\mathbb{E}_0 \left[\left\| \mathrm{Q}_\Phi Y - F_0 \right\|_{2,n}^2 \right] = O\left(n^{-2\alpha/(1+2\alpha)} \right), \tag{4.9}$$

where $\mathbb{E}_0(\cdot)$ represents an expectation with respect to the true data generating distribution of Y.

I now state main results on the posterior contraction rate of the functional horseshoe prior.

Theorem 6. *Consider the model* (4.1) *equipped with the functional horseshoe prior* (4.4)-(4.5). *Assume* **(A1)** *and* **(A2)** *hold and* $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$. *Further assume that for some integer* $\alpha \geq 1$, *the true regression function* $f_0 \in C^\alpha[0,1]$ *and the B-spline basis functions* Φ *are constructed with* $K_n - \lfloor \alpha \rfloor$ *knots and* $\lfloor \alpha \rfloor - 1$ *degree, where* $K_n \asymp n^{1/(1+2\alpha)}$. *Suppose that the prior hyperparameters* a *and* b *in* (4.5) *satisfy* $a \in (\delta, 1-\delta)$ *for some constant* $\delta \in (0, 1/2)$, *and*

$K_n \log K_n \prec -\log b \prec (nK_n)^{1/2}$. *Then,*

$$\mathbb{E}_0 \left[P \left\{ \left\| \Phi\beta - F_0 \right\|_{2,n} > M_n (f_0)^{1/2} \mid Y \right\} \right] = o(1), \tag{4.10}$$

where

$$M_n(f_0) = \begin{cases} \zeta_n n^{-1}, \ \textit{if } F_0 \in \mathfrak{L}(\Phi_0), \\ \zeta_n n^{-2\alpha/(1+2\alpha)} \log n, \ \textit{if } F_0^{\mathrm{T}} (\mathrm{I} - \mathrm{Q}_0) F_0 \asymp n, \end{cases}$$

and ζ_n can be any arbitrary sequence that diverges to infinity as n tends to ∞.

Theorem 6 exhibits an adaptive property of the fHS prior. If the true function is α-smooth, then the posterior contracts around the true function at the near minimax rate of $n^{-\alpha/(2\alpha+1)} \log n$. However, if the true function F_0 belongs to the finite dimensional subspace $\mathfrak{L}(\Phi_0)$, then the posterior contracts around F_0 in the empirical L_2 norm at the parametric $1/\sqrt{n}$ rate. I note that the bound $K_n \log K_n \prec -\log b \prec (nK_n)^{1/2}$ is key to the adaptivity of the posterior, since the strength of the shrinkage towards $\mathfrak{L}(\Phi_0)$ is controlled by b. If $-\log b \prec K_n \log K_n$, then the shrinkage towards $\mathfrak{L}(\Phi_0)$ is too weak to achieve the parametric rate when $F_0 \in \mathfrak{L}(\Phi_0)$. On the other hand, if $-\log b \succ (nK_n)^{1/2}$, the resulting posterior distribution would strongly concentrate around $\mathfrak{L}(\Phi_0)$, and it would fail to attain the optimal nonparametric rate of posterior contraction when $F_0 \notin \mathfrak{L}(\Phi_0)$.

I ignore the subspace of functions such that $\{F \in \mathbb{R}^n : F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F = o(n), \ F \notin \mathfrak{L}(\Phi_0)\}$. I only focus on the function space that can be strictly separated from the null space $\mathfrak{L}(\Phi_0)$. However, I acknowledge that it would be meaningful to illustrate the shrinkage behavior when the regression function f approaches the null space in the sense that $F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F/n \to 0$ as $n \to \infty$.

4.4 Simulation Studies for Univariate Examples

In this section, I examine the performance of the functional horseshoe prior on various simulated data sets. I consider three models as follows:

$$\text{(i) simple regression model:} \qquad Y_i = f(x_i) + \epsilon_i \tag{4.11}$$

$$\text{(ii) varying coefficient model:} \qquad Y_i = w_i f(x_i) + \epsilon_i \tag{4.12}$$

$$\text{(iii) density function estimation:} \qquad p(Y_i) = \frac{\exp\{f(Y_i)\}}{\int \exp\{f(t)\} dt}. \tag{4.13}$$

In scenario (i) and (ii), $\epsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, n$. In (iii), $p(\cdot)$ is the density function of Y. The varying coefficient model (Hastie and Tibshirani, 1993) in (4.12) reduces to a linear model when the coefficient function f is constant, and the density function p is Gaussian when the log-density function f is quadratic in the log-spline model in (4.13); (Kooperberg and Stone, 1991). These facts motivate the use of the fHS prior in these examples to shrink towards the respective parametric alternatives. For each setting, I considered the case corresponding to the relevant parametric model.

For (i) and (ii), I generated the covariates independently from a uniform distribution between $-\pi$ and π and set the error variance $\sigma^2 = 1$. For each scenario (i) - (iii), I considered three parametric choices for f. For scenario (i), I considered f to be linear, quadratic, and sinusoidal. For (ii), I considered constant, quadratic and sinusoidal functions. For (iii), I considered normal, log-normal and mixture of normal distributions. For the first two cases, I standardized the true function so as to obtain a signal-to-noise ratio of 1.0.

I used the B-spline basis with $K_n = 8$ in (4.2) to model the function f in each setting. To shrink the regression function in (4.11) towards linear subspaces, I set $\Phi_0 = \{\mathbf{1}, \mathbf{x}\}$ in the fHS prior (4.4). For the varying coefficient model (4.12), I set $\Phi_0 = \{\mathbf{1}\}$ to shrink f towards constant functions, whence the resulting model reduces to a linear regression model.

Table 4.1: Results of univariate examples

True function	Method	$n = 200$	$n = 500$	$n = 1000$
Linear	fHS	**0.93** (0.81)	**0.44** (0.45)	**0.17** (0.17)
	B-spline	3.57 (1.60)	1.54 (0.74)	0.76 (0.38)
Quadratic	fHS	3.63 (1.73)	1.55 (0.74)	0.77 (0.37)
	B-spline	3.59 (1.60)	1.56 (0.74)	0.78 (0.38)
Sine	fHS	3.64 (1.58)	1.50 (0.74)	0.75 (0.36)
	B-spline	3.57 (1.60)	1.53 (0.74)	0.76 (0.38)
Constant	fHS	**0.13** (0.15)	**0.06** (0.08)	**0.03** (0.04)
	B-spline	1.33 (0.63)	0.48 (0.26)	0.25 (0.13)
Quadratic	fHS	1.35 (0.62)	0.51 (0.27)	0.27 (0.13)
	B-spline	1.36 (0.64)	0.51 (0.26)	0.27 (0.13)
Sine	fHS	1.35 (0.63)	0.48 (0.26)	0.25 (0.13)
	B-spline	1.33 (0.63)	0.48 (0.26)	0.25 (0.13)
Normal	fHS	**1.34** (1.35)	**0.59** (0.52)	**0.35** (0.31)
	B-spline	10.30 (5.00)	3.68 (1.42)	1.96 (0.77)
Log-normal	fHS	5.15 (2.70)	3.35 (1.14)	2.91 (0.98)
	B-spline	6.37 (4.21)	3.27 (1.86)	2.83 (1.14)
Mixture	fHS	4.42 (2.18)	1.79 (0.85)	1.04 (0.39)
	B-spline	5.31 (3.61)	1.85 (0.93)	1.04 (0.39)

Finally, I set $\Phi_0 = \{\mathbf{1}, Y, Y^2\}$ to shrink f towards the space of quadratic functions in (4.13), which results in the density p being shrunk towards the class of Gaussian distributions. I note that the prior for p in (4.13) is data-dependent. An inverse-gamma prior with parameters $(1/100, 1/100)$ was imposed on σ^2 for the fHS prior in (i) and (ii). In all three examples, I set $b = \exp\{-K_n \log n/2\}$ to satisfy the conditions of Theorem 6 and arbitrarily set $a = 1/2$. Although Theorem 6 only applies to the regression model (4.11), the empirical results for these hyperparameter choices are promising for the varying coefficient model and the log-density model as well.

I imposed Jeffrey's prior, $\pi(\beta, \sigma^2) \propto 1/\sigma^2$, on the B-spline coefficients for the simple regression model and the varying coefficient model as a competitor to the fHS prior. Following Ghosal et al. (2008), I assigned independent $U(-\pi, \pi)$ priors on the B-spline coefficients,

which are known to guarantee the minimax rate of posterior convergence rate for the log-density model. For each prior, I used the posterior mean $\hat{f}$ as a point estimate for f, and report the empirical *Mean Square Error* (MSE), i.e. $\left\|\widehat{f} - f\right\|_{n,2}^2$.

In Table 4.1, I report 100 times MSE of the posterior mean estimator and its standard deviation over 100 replicates in estimating the unknown function f for all three models, for sample sizes $n = 200$, 500, and 1000. The first top three rows are for the simple regression model; the second three rows for the varying coefficient model; the last three rows for the density estimation. "Mixture" in the last row indicates a mixture of Gaussian densities as $0.3\mathrm{N}(2, 1) + 0.7\mathrm{N}(-1, 0.5)$. In all three settings, when the true function f belongs to the nominal parametric class, the posterior mean function resulting from the functional horseshoe prior clearly outperforms the B-spline prior. When the true function does not belong to the parametric model, the functional horseshoe prior performs comparably to the B-spline prior.

Figure 4.2 depicts the point estimate (posterior mean) and pointwise 95% credible bands for the unknown function f for a single data set for each of the three examples when the true function belongs to the parametric class; that is, a linear function in (4.11), a constant function in (4.12), and a quadratic function in (4.13). Figure 4.3 depicts the corresponding estimates when the data generating function does not fall into the assumed parametric class. It is evident from Figure 4.2 that when the parametric assumptions are met, the fHS prior performs similarly to the parametric model. THis fact empirically corroborates my findings in Theorem 6 that the posterior contracts at a near parametric rate when the parametric assumptions are met. It is also evident that the fHS procedure automatically adapts to deviations from the parametric assumptions in Figure 4.3, again confirming the conclusion of Theorem 6 that when the true function is well-separated from the parametric class, the posterior concentrates at a near optimal minimax rate. I reiterate that the same hyperparameters $a = 1/2$ and $b = \exp\{-K_n \log n/2\}$ for the fHS prior were used in the examples in Figure 4.2 and Figure 4.3.

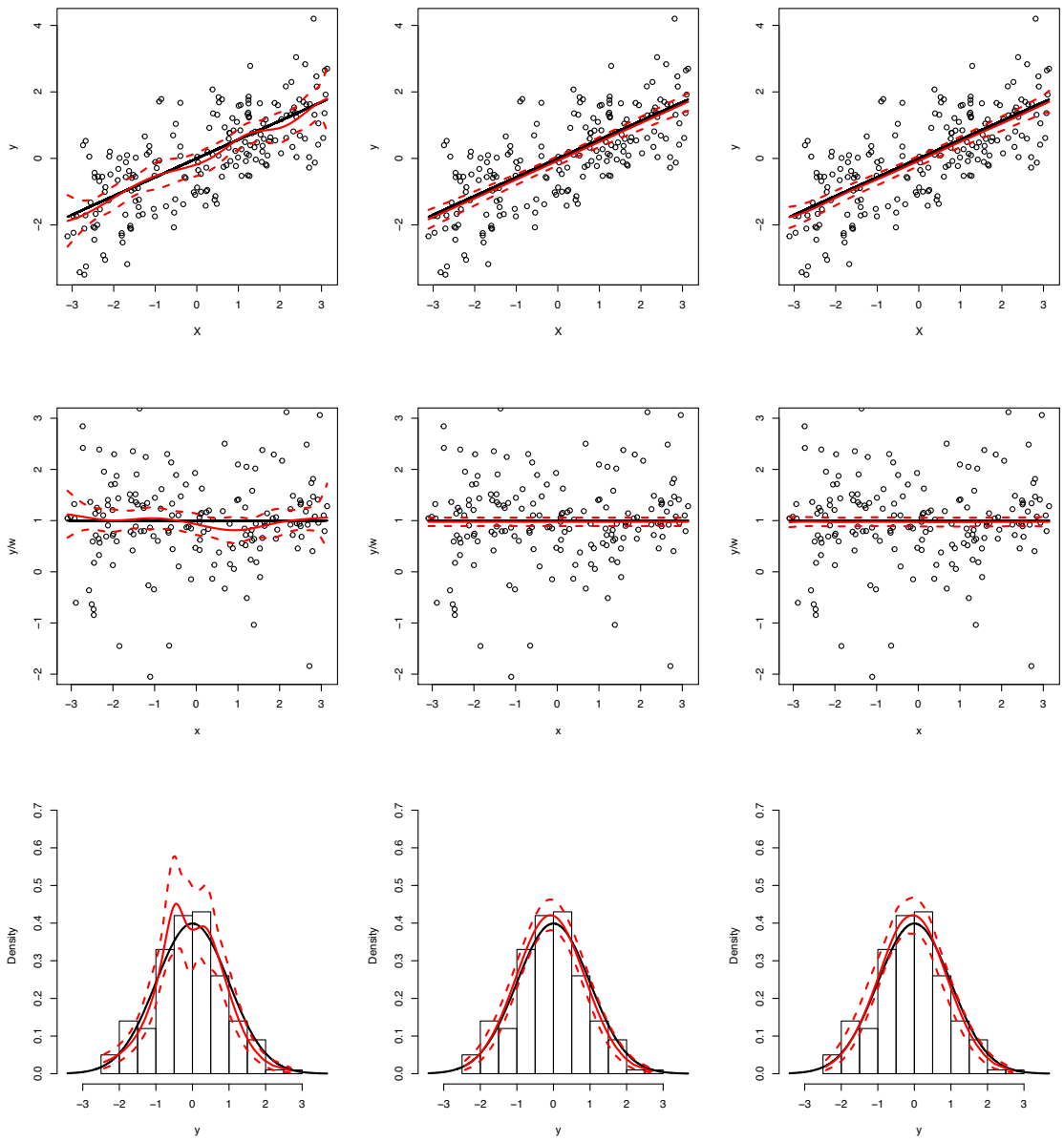


Figure 4.2: Examples when the underlying true functions are parametric. Posterior mean of each procedure (red solid), its 95% pointwise credible bands (red dashed), and the true function (black solid) from a single example with $n = 200$ for each model. The top row is for the simple regression model; the second row is for the varying coefficient model; the last row is for the density estimation. The Bayesian B-spline procedure, the Bayesian parametric model procedure, and functional horseshoe priors are illustrated in the first, second, and third columns, respectively.

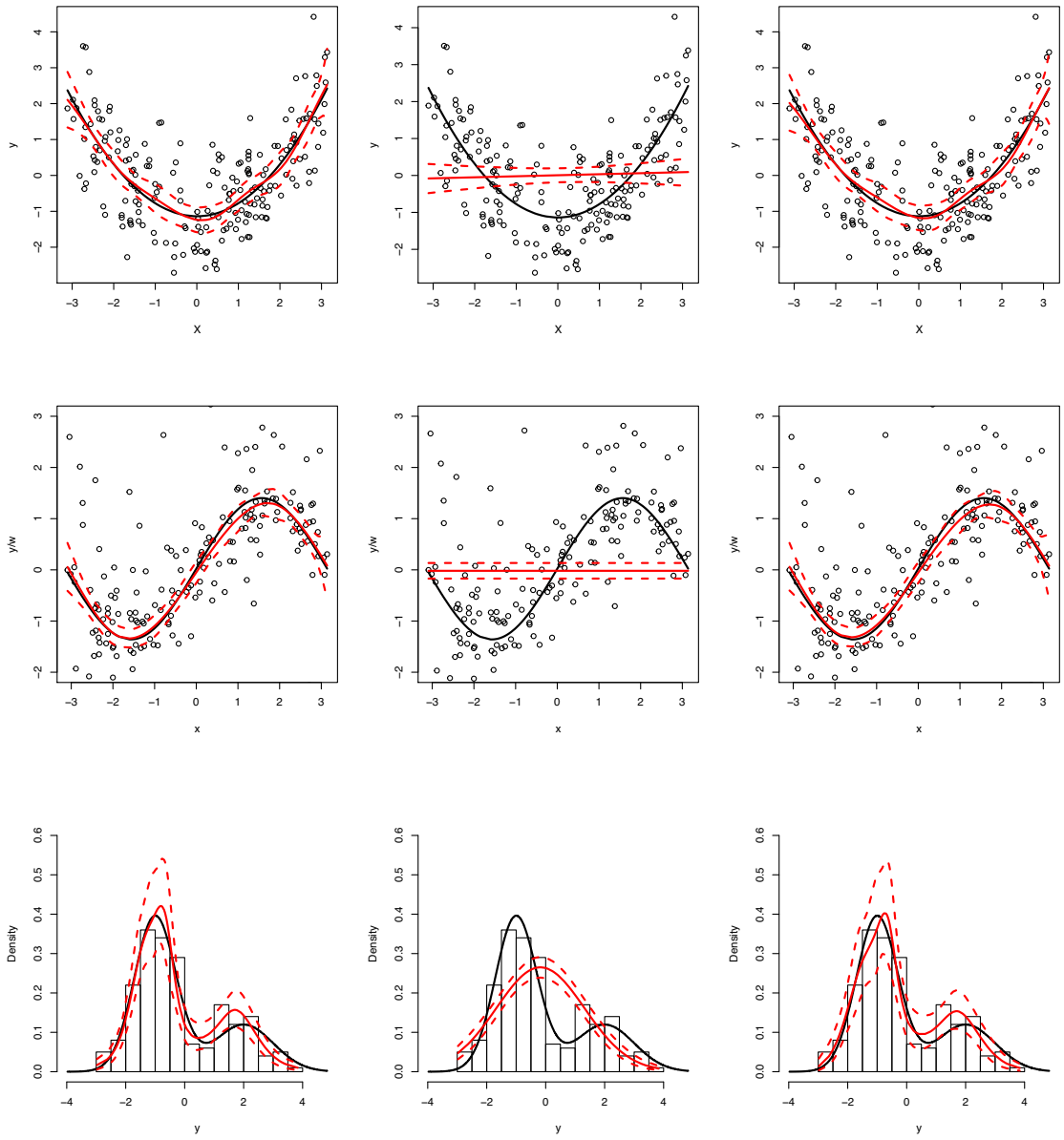


Figure 4.3: Examples when the underlying true functions are nonparametric. Posterior mean of each procedure (red solid), its 95% pointwise credible bands (red dashed), and the true function (black solid) from a single example with $n = 200$ for each model. The top row is for the simple regression model; the second row is for the varying coefficient model; the last row is for the density estimation. The Bayesian B-spline procedure, the Bayesian parametric model procedure, and functional horseshoe priors are illustrated in the first, second, and third columns, respectively.

4.5 Applications to Additive Models

My regression examples in the previous subsection involved one predictor variable. In the case of multiple predictors, a popular modeling framework is the class of additive models (Hastie and Tibshirani, 1986), where the unknown function relating p candidate predictors to a univariate response is modeled as the sum of p univariate functions, with the jth function only dependent on the jth predictor $X_j = \{x_{1j}, \ldots, x_{nj}\}$. In this section, I apply the fHS prior to additive models and compare results obtained under this prior to several alternative methods. To be consistent with my previous notation, I express additive models as

$$Y = \sum_{j=1}^{p} F_j + \epsilon, \tag{4.14}$$

where $F_j = \{f_j(x_{1j}), \ldots, f_j(x_{nj})\}$ for $j = 1, \ldots, p$ and $\epsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n)$. I let Φ_j denote the spline basis matrix for X_j and let $\beta_j = \{\beta_{j1}, \ldots, \beta_{jK_n}\}$ denote the corresponding coefficient. In general, each component function can be modeled nonparametrically, for example, using the B-spline basis functions as described in the previous section, $f_j(x) = \sum_{l=1}^{K_n} \beta_{jl}\phi_l(x) = \Phi_j\beta_j$ for $j = 1, \ldots, p$. However, if there are many candidate predictors, then nonparametrically estimating p functions may be statistically difficult and in addition, may result in a loss of precision if only a small subset of the variables are significant. With this motivation, I extend the fHS framework to additive models, where I assign independent fHS priors to the f_j's with $Q_0 = 0$ in (4.4) to facilitate shrinkage of each of these functions towards the null function. Therefore, the resulting prior specification can be expressed as

$$\begin{aligned} \pi(\beta \mid \tau^2, \sigma^2) &\propto \exp\left\{ -\frac{1}{\sigma^2} \sum_{j=1}^{p} \frac{\beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_j \beta_j}{\tau_j^2} \right\} \\ \pi(\tau_j) &\propto \frac{(\tau_j^2)^{b-1/2}}{(1+\tau_j^2)^{(a+b)}} \mathbf{1}_{(0,\infty)}(\tau_j), \end{aligned}$$

for $j = 1, \ldots, p$. This prior imposes a shrinkage effect on each $\left\| f_j \right\|_{2,n}^2 = \beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_j \beta_j$ towards the null function. Thus, the resulting posterior distribution of F_j concentrates on the zero function when the marginal effect of F_j is negligible.

4.5.1 A Comparison to the Standard Horseshoe Prior

For the additive model in (4.14), one can impose a product of standard horseshoe (HS) priors (Carvalho et al., 2010) on the spline coefficients as

$$\begin{aligned} \pi(\beta \mid \lambda, w, \sigma^2) &\propto \exp\left\{ -\frac{1}{\sigma^2 \lambda^2} \sum_{j=1}^{p} \sum_{l=1}^{K_n} \frac{\beta_{jl}^2}{\psi_{jl}^2} \right\} \\ \lambda &\sim C^+(0,1) \\ \psi_{jl} &\sim C^+(0,1), \end{aligned} \tag{4.15}$$

where $C^+(0,1)$ is the half-Cauchy distribution and β_{jl} is the l-th spline coefficient for the component function of the j-th covariate for $j = 1, \ldots, p$ and $l = 1, \ldots, K_n$. Polson and Scott (2010b) states that this prior imposes global-local shrinkage rules. The parameter λ serves a global shrinkage parameter controlling the concentration near zero, while the ψ_{jl}'s are local shrinkage parameters that control the tail heaviness of the individual coefficients. The use of the standard horseshoe prior would impose strong shrinkage effects towards zero on each coefficient, but it does not take into account the grouping structure in the spline expansions of the components. From a frequentist perspective, this issue was addressed in Huang et al. (2010). There, the authors stated that the standard Lasso (Tibshirani, 1996) with sparsity constraints on individual marginal coefficients is not appropriate for additive models. The group Lasso penalties (Yuan and Lin, 2006) on the spline coefficients achieve much better performance in prediction and model selection compared to the standard Lasso in such settings. The same illustration can be applied to the Bayesian additive model. The use of the standard horseshoe prior might degrade the estimation performance due to the ignorance of the grouping structure

in the spline coefficients. In the following sections, I provide simulated and real examples where the standard horseshoe prior does not perform well for additive models, but the procedure based on the fHS prior shows excellent performances compared to other state-of-the-art methods.

4.5.2 Simulation Studies

For additive models, Ravikumar et al. (2009) proposed penalized likelihood procedures called *Sparse Additive Models* (SpAM) that combine ideas from model selection and additive nonparametric regression. The penalty term of SpAM can be described as a weighted group Lasso penalty (Yuan and Lin, 2006) in which the coefficients for each component function f_j for $j = 1, \ldots, p$ are forced to simultaneously shrink towards zero. Meier et al. (2009) proposed *High-dimensional Generalized Additive Model* (HGAM) that differs from SpAM because its penalty term imposes both shrinkage towards zero and regularization on the smoothness of the function. Huang et al. (2010) introduced a two step procedure of adaptive group Lasso (AdapGL) for additive models. The first step estimates the weight of the group penalty, and the second step applies it to the adaptive group Lasso penalty. Since the performance of penalized likelihood methods is sensitive to the choice of the tuning parameter, in the simulation studies that follow I considered two criterion for tuning parameter selection: AIC and BIC. `R` packages `SAM`, `hgam`, and `grpLasso` were used to implement SpAM, HGAM, and AdapGL, respectively. I also considered the standard HS prior. Its computation was implemented by the `R` package `monomv`. For the fHS prior and the HS prior, I imposed a prior on $\pi(\sigma^2)$ proportional to $1/\sigma^2$. I used $20,000$ samples from the MCMC algorithms after $10,000$ burn-in iterations to estimate the posterior mean estimator.

I define the signal-to-noise ratio as SNR $= \mathrm{Var}(f(X))/Var(\epsilon)$, where f is the true underlying regression function, and I examine the same simulation scenarios that were considered in Meier et al. (2009) as follows:

Scenario 1: ($p = 200$, SNR ≈ 15). This is the same as Example 1 in Meier et al. (2009). A similar scenario was also considered in Härdle et al. (2012) and Ravikumar et al. (2009). The true model is

$$Y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + \epsilon_i,$$

where $\epsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, 1)$ for $i = 1, \ldots, n$, with

$$\begin{aligned} f_1(x) &= -\sin(2x), \;\; f_2(x) = x^2 - 25/12, \;\; f_3(x) = x, \\ f_4(x) &= \exp\{-x\} - 2/5 \cdot \sinh(5/2). \end{aligned}$$

The covariates are independently generated from a uniform distribution between -2.5 to 2.5.

Scenario 2: ($p = 80$, SNR ≈ 7.9). This is equivalent to Example 3 in Meier et al. (2009) and similar to an example in Lin and Zhang (2006). The true model is

$$Y_i = 5f_1(x_{i1}) + 3f_2(x_{i2}) + 4f_3(x_{i3}) + 6f_4(x_{i4}) + \epsilon_i,$$

where $\epsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, 1.74)$ for $i = 1, \ldots, n$, with

$$\begin{aligned} f_1(x) &= x, \;\; f_2(x) = (2x-1)^2, \;\; f_3(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)}, \\ f_4(x) &= 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) \\ &\quad +0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x). \end{aligned}$$

The covariate $\mathbf{x}_j = \{x_{1j}, \ldots, x_{nj}\}^{\mathrm{T}}$ for $j = 1, \ldots, p$ is generated by $\mathbf{x}_j = (W_j + U)/2$, where $W_1, \ldots, W_p$ and U are independently simulated from $\mathrm{U}(0, 1)$ distributions.

Scenario 3 ($p = 60$, SNR ≈ 11.25). This scenario is equivalent to Example 4 in Meier et al. (2009), and a similar example was also considered in Lin and Zhang (2006). The same functions and the same process to generate the covariates used as in *Setting 2* were used in this scenario. The true model is

$$\begin{aligned} Y_i &= f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) \\ &\quad + 1.5f_1(x_{i5}) + 1.5f_2(x_{i6}) + 1.5f_3(x_{i7}) + 1.5f_4(x_{i8}) \\ &\quad + 2.5f_1(x_{i9}) + 2.5f_2(x_{i10}) + 2.5f_3(x_{i11}) + 2.5f_4(x_{i12}) + \epsilon_i, \end{aligned}$$

where $\epsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, 0.5184)$ for $i = 1, \ldots, n$.

To evaluate the estimation performance of the fHS prior, I report the MSE for each method. To measure the performance of variable selection, I examined the proportion of times the true model was selected, as well as *Matthews correlation coefficient* (MCC; Matthews (1975)), defined as,

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})},$$

where TP, TN, FP, and FN denote the number of true positive, true negatives, false positives, false negatives, respectively. MCC is generally regarded as a balanced measure of the performance of classification methods, which simultaneously takes into account TP, TN, FP, and FN. I note that MCC is bounded by 1, and the closer MCC is to 1, the better the model selection performance is.

For variable selection using the fHS prior and the HS prior, I used 95% pointwise credible bands for each component function to exclude component functions whose credible bands uniformly contained the zero function on the entire support of the corresponding covariate. To

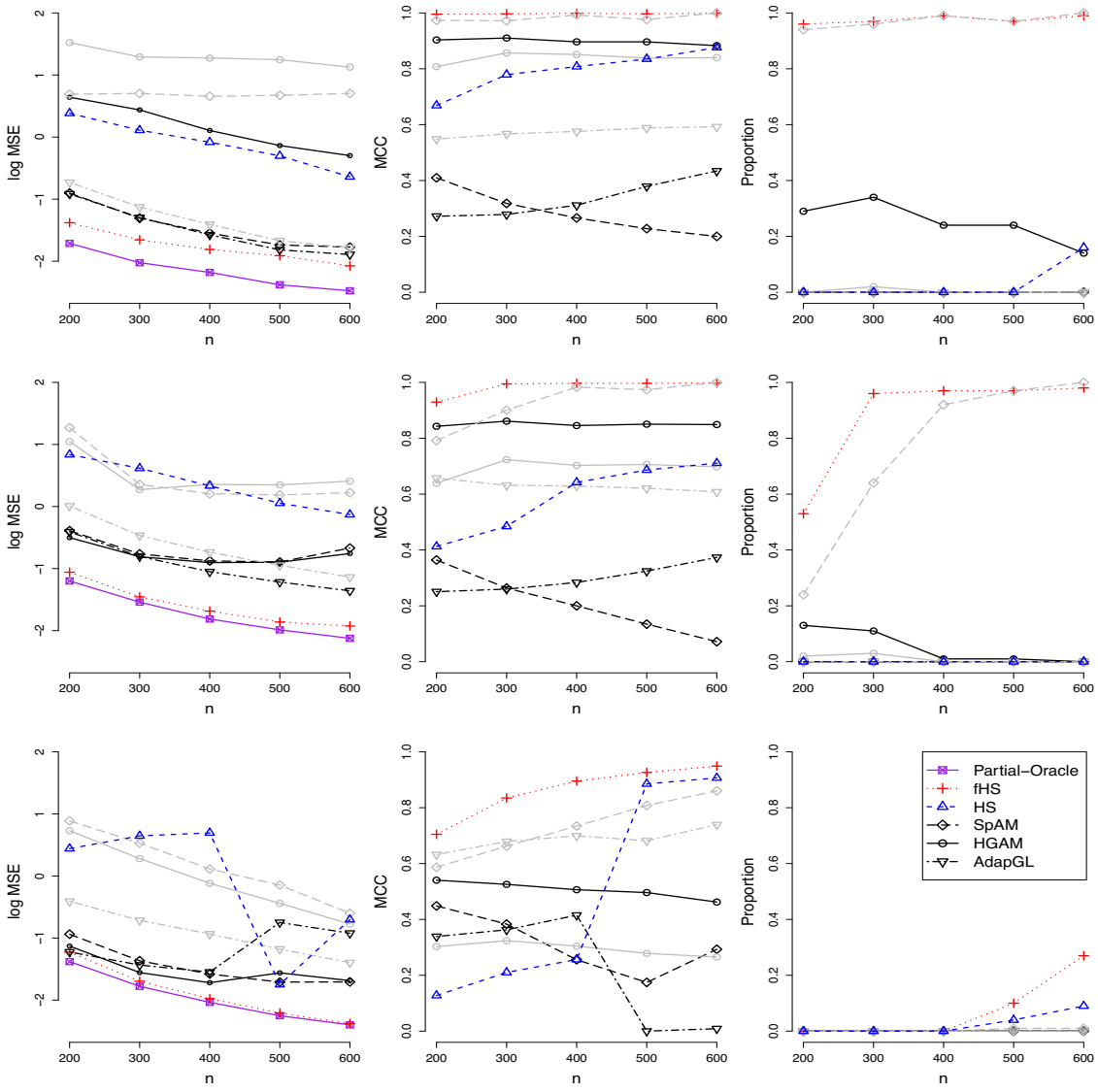


Figure 4.4: The first column illustrates the logarithm of the MSE of each method; the second column displays the MCC; the third column is the proportion of times the each procedure selected the true model. The top row, the middle row, and the bottom row represent the *Scenario 1*, *Scenario 2*, and *Scenario 3*, respectively. For penalized likelihood methods, AIC (black) and BIC (grey) were used to choose the tuning parameter.

investigate the performance achieved by the proposed method, I compared it to a “partial oracle estimator". The partial oracle estimator refers to the B-spline least squares estimator when the variables in the true model are given, but the true component functions in the additive model are not provided.

Results from a simulation study to compare these methods are depicted in Figure 4.4. In all three settings,the procedure based on the fHS prior has smaller MSE than the estimator based on the horsehoe prior and the penalized likelihood estimators. The proposed procedure also provides comparable or better variable selection than the other methods. I note that the SpAM procedure with tuning parameter selected by BIC provides comparable variable selection performance to the fHS prior in *Scenario 1*, yet its MSE is at least 8 times larger than that of the procedure based on the fHS prior (note that the reported results are on the logarithmic scale). The results suggest that the fHS prior provides improvement over the penalized likelihood methods in terms of both MSE and model selection performance in these simulation scenarios.

4.5.3 Real Data Analysis: Boston Housing Data and Ozone Data

In this section, I apply the functional horseshoe prior to two well known data sets: the first concerns ozone levels and the second considers housing prices in Boston. Both data sets are available in the R package `mlbench`. These two data sets have been previously analyzed by various researchers, including Buja et al. (1989), Breiman (1995), Lin and Zhang (2006) and Xue (2009). Following the pre-processing step in Xue (2009), I standardized both the response and independent variables prior to my analyses.

I first consider the Boston housing data set that contains the median value of 506 owner-occupied homes in the Boston area, together with several variables that might be associated with the median value. To examine the performance of my method in eliminating extraneous predictors, I added 40 spurious variables generated as i.i.d. standard Gaussian deviates. Using

the standard notation for the variable in this data set, I then assumed a model of the following form:

$$\begin{aligned} \texttt{medv} &= \beta_0 + f_1(\texttt{crim}) + f_2(\texttt{indus}) + f_3(\texttt{nox}) + f_4(\texttt{rm}) + f_5(\texttt{age}) + f_6(\texttt{dis}) \\ &\quad + f_7(\texttt{tax}) + f_8(\texttt{ptratio}) + f_9(\texttt{b}) + f_{10}(\texttt{lstat}) + \epsilon, \end{aligned}$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n)$. Each component function was modeled by the B-spline bases with $K_n = 8$. Fifty test data points were randomly selected to estimate the out-of-sample prediction error. Five hundreds simulations of each procedure were used to generate the plots in Table 4.2.

I also modeled the ozone data set using each of the procedures that were applied the housing data. The ozone data consists of the daily maximum one-hour-average ozone readings and nine meteorological variables for 330 days in the Los Angeles basin in 1976. The model applied to these data can be expressed as follows:

$$\begin{aligned} \texttt{ozone} &= \beta_0 + f_1(\texttt{height}) + f_2(\texttt{wind}) + f_3(\texttt{humidity}) + f_4(\texttt{temp1}) \\ &\quad + f_5(\texttt{temp2}) + f_6(\texttt{inv height}) + f_7(\texttt{gradient}) + f_8(\texttt{inv temp}) \\ &\quad + f_9(\texttt{visibility}) + \epsilon. \end{aligned}$$

Like the Boston Housing data case, I added 40 spurious variables generated as i.i.d. standard Gaussian deviates. I used B-spline bases with $K_n = 5$ to model the component functions. I performed a cross-validation experiment to assess the predictive performance of the competing methods. In each of 500 simulated data sets, I held out 30 data values as the test set and used the remaining observations to estimate the model. The parameter settings described in Section 4.5.2 were again used for the functional horseshoe prior. Also, for each training data set I generated $30,000$ posterior samples by following the MCMC algorithm described in the

Table 4.2: Results of real data examples

Boston Housing Data			
Method	Test Error	NN	Selected Model
Original	0.156(0.065)		
fHS	**0.154**(0.067)	**0.00**	crim, nox, rm, dis, ptratio, lstat
HS	0.180(0.081)	**0.00**	crim, nox, rm, dis, ptratio, lstat
SpAM(AIC)	0.224(0.072)	21.06	All
SpAM(BIC)	0.344(0.093)	2.00	crim, nox, rm, dis, ptratio, lstat
HGAM(AIC)	0.212(0.095)	37.49	All
HGAM(BIC)	0.222(0.115)	1.06	indus, nox, age, dis, tax, ptratio
AdaptGL(AIC)	0.579(0.214)	40.00	All
AdaptGL(BIC)	0.218(0.144)	4.17	nox, rm, dis, tax, ptratio, lstat
Ozone Data			
Original	0.311(0.085)		
fHS	**0.278**(0.092)	0.02	temp2, gradient
HS	0.294(0.296)	**0.00**	temp2
SpAM(AIC)	0.427(0.156)	20.67	All but height and inv temp
SpAM(BIC)	0.624(0.213)	0.07	temp1, temp2, gradient
HGAM(AIC)	0.298(0.109)	23.12	All but gradient
HGAM(BIC)	0.631(0.260)	0.208	humidity, temp1
AdaptGL(AIC)	0.359(0.131)	21.91	All but height and inv temp
AdaptGL(BIC)	0.341(0.142)	2.252	humidity, temp1, temp2, inv height, gradient, visibility

Appendix, and only the last $20,000$ samples were used in the analysis. I compared the performance of the procedure based on the proposed priors with that of SpAM, HGAM, AdapGL and the classical B-spline estimator. The classical B-spline estimator was fit without the spurious noise variables. For the penalized likelihood methods, AIC and BIC were used to choose tuning parameters. Table 4.2 displays the average of test set errors, the average number of selected noise variables, and the most frequently selected model for each method.

In Table 2, “Test Error" refers to the average of empirical L_2 test errors, and “NN" represents the averaged number of selected spurious variables. “Original" indicates the B-spline least square estimator from the original model without spurious variables. Table 2 shows that

for both data sets the procedure based on the fHS prior achieved the smallest test errors, and it also selected the minimum number of spurious variables. Moreover, its test error was smaller than that of the original estimator that was estimated without the spurious variables. The HS prior also selects parsimonious models in the sense that the average number of selected spurious variables is close to zero, but its prediction error is much larger than the fHS prior. For both data sets, the model selected by the fHS prior was similar to that chosen by SpAM with BIC. However, the test error of the SpAM procedure was roughly twice that of fHS. More generally, the fHS procedure outperformed all of the other procedures in these examples.

4.6 Conclusion

I have proposed a class of shrinkage priors which I call the functional horseshoe priors. When appropriate, these priors imposes strong shrinkage towards a pre-specified class of functions. The shrinkage term in this prior is new. It directly allows the nonparametric function to shrink towards parametric functions. By so doing, it preserves the minimax optimal parametric rate of posterior convergence $n^{-1/2}$ when the true underlying function is parametric, and it also comes within $O(\log n)$ of achieving the minimax nonparametric rate when the true function is strictly separated from the class of parametric functions.

The novel shrinkage term contained in the proposed prior, $F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F$ (i.e., (4.4)), can be naturally applied to a new class of penalized likelihood methods having a general form expressible as

$$-l(Y \mid F) + p_\lambda\big(F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F\big),$$

where $l(Y \mid F)$ is the logarithm of a nonparametric likelihood function and p_λ is the penalty term. In contrast to other penalized likelihood methods, this form of penalty allows shrinkage towards the space spanned by a projection matrix Q_0, rather than simply a zero function.

5. NONLOCAL FUNCTIONAL PRIORS FOR NONPARAMETRIC HYPOTHESIS TESTING AND HIGH-DIMENSIONAL MODEL SELECTION

5.1 Introduction

Consider the following nonparametric additive model (Hastie and Tibshirani, 1986) that was discussed in Chapter 4; i.e., for a response variable $\mathbf{y} = \{y_1, \ldots, y_n\}$ and the covariates $X = \{X_1, \ldots, X_p\}$,

$$\mathbf{y} = \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2 \mathrm{I}_n)$, and f_j is the j-th marginal regression function. Also, X_j is the j-th covariate for $j = 1, \ldots, p$. I assume that some of the functions f_j are nonzero and the rest are zero functions.

For additive models, significant progress in selecting a subset of variables has been made over the past decades under high-dimensional settings. From a frequentist perspective, this problem has been examined in Ravikumar et al. (2009), Meier et al. (2009), and Huang et al. (2010). Theoretical properties of associated estimation properties have been investigated in Raskutti et al. (2012) and Yuan and Zhou (2016). In a Bayesian framework, Shang and Li (2014) investigated asymptotic properties of high-dimensional model selection procedures defined by Gaussian priors.

From a Bayesian perspective, Choi et al. (2009) investigated the asymptotic property of nonparametric Bayesian testing procedure. More recently, Choi and Rousseau (2015) studied a Bayesian hypothesis test on the regression function in partially linear models using a Gaussian process prior and showed its consistency. However, neither Choi et al. (2009) or Choi and Rousseau (2015) provided the exact convergence rate of the evidence when the null hypothesis

is true.

In this dissertation, I propose new classes of prior densities called *nonlocal functional prior densities* for nonparametric Bayesian hypothesis testing problems, and I apply the proposed prior to the model selection procedure for additive models under high-dimensional settings. I investigate theoretical properties of the resulting Bayesian model selection procedure and show its model selection consistency under high-dimensional settings when the number of covariates p increases at an sub-exponential rate of n.

The proposed prior densities are a novel extension of nonlocal priors (Johnson and Rossell, 2010) to nonparametric settings. For parametric models, a family of nonlocal prior densities assigns negligible density around the null value of the parameter, and Johnson and Rossell (2010) showed that the Bayes factor based on the nonlocal priors penalizes the alternative hypothesis at a faster rate than that of local prior density functions that are strictly positive at the null value, when the data-generating process is consistent with the null hypothesis. For high-dimensional linear model selection problems, Johnson and Rossell (2012) and Shin et al. (2017) provided desirable theoretical properties of model selection procedures based on nonlocal priors. However, their extension to nonparametric models has been hindered because it is not clear how to define the null space to construct nonlocal prior densities on the space of functions due to the fact that the null hypothesis is composite for nonparametric hypothesis tests.

I first introduce a novel discrepancy quantity between the null space of functions and the function objective to be inferred, and construct nonlocal functional prior densities based on the null space defined by the discrepancy. I then show the theoretical properties of the Bayesian hypothesis test (model selection) based on the proposed priors for univariate nonparametric regression models. I derive the asymptotic rate of the Bayes factor in favor of the alternative based on the local prior densities, e.g. g-priors Zellner (1986), and show that it diminishes only at a polynomial rate under a true null, but increases at an exponential rate of the sample size under a true alternative. On the other hand, when data sets are generated from the null

hypothesis, the Bayes factor based on the proposed prior densities not only achieve a faster rate than that from existing local priors, but also can attain a sub-exponential rate, rather than a polynomial rate, under some conditions. I provide description of these properties detailed in Section 5.3. I also discuss some applications of the proposed priors and test their finite sample behavior by simulation studies in Section 5.4.

I apply the proposed nonlocal functional prior density to additive model selection problems under high-dimensional settings. In Section 5.5.1 I show that the resulting model selection procedure is consistent in the sense that the posterior model probability of the true data-generating model converges to one in probability under mild regularity conditions. In Section 5.5.2, I also provide a convergence rate of the logarithm of posterior model probabilities defined by the nonlocal functional priors, and I show that this rate can be decomposed as a sum of the logarithm of posterior model probabilities from local priors (e.g. Gaussian priors) and an addtional penalty term on the model. It is shown that the additional penalty term is adaptively determined by the marginal effect of the B-spline estimator, and this property explains why the model selection procedure based on the proposed priors outperforms the other methods in simulation studies and real data examples.

Choosing an appropriate hyperparameter for the prior densities is important when implementing a Bayesian models selection (hypothesis testing). In Section 5.5.5, I propose a practical procedure to choose the hyperparameter of the nonlocal functional priors by comparing the null distribution and the prior density of the discrepancy measure. For computation, I describe a scalable algorithm that is a modified version of the Simplified Shotgun Stochastic Search with Screening (S5) (Shin et al., 2017). Originally, S5 was designed to efficiently explore the space of linear models. Here, I modify the S5 to be suitable for nonparametric additive model selection. All computational functions used in this dissertation are available in the `R` package `BayesS5`.

5.2 Bayesian Nonparametric Hypothesis Testing Procedures

To illustrate the idea of nonlocal functional priors, I assume the nonparametric univariate regression model with the regression function f can be expressed as

$$y_i = f(x_i) + \epsilon_i, \tag{5.1}$$

where $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$. I suppose that the predictor is compactly supported, and assume without loss of generality that $x_i \in [0, 1]$ for each i. I denote $\mathbf{y} = \{y_i\}_{i=1,\ldots,n}$, $\mathbf{x} = \{x_i\}^{\mathrm{T}}_{i=1,\ldots,n}$ and $F = f(\mathbf{x}) = \{f(x_1), \ldots, f(x_n)\}^{\mathrm{T}}$. For simplicity, I assume that σ^2 is known. By using B-spline basis functions, I model $F = \sum_{j=1}^{K_n} \beta_j \phi_j(\mathbf{x}) = \Phi\beta$, where ϕ_j denote the j-th B-spline basis for $j = 1, \cdots, K_n$, $\beta = \{\beta_1, \ldots, \beta_{K_n}\}$, and Φ is the $n \times K_n$ matrix of the B-spline bases.

I aim to test if the regression function F belongs to a certain class of parametric functions that can be linearly spanned by a design matrix Φ_0 with a dimension d_0. So, the null space of the corresponding hypothesis test can be defined as $\mathfrak{L}(\Phi_0) = \{F \in \mathbb{R}^n : F = \Phi_0\alpha \text{ for some } \alpha \in \mathbb{R}^{d_0}\}$. For example, any linear function of $\mathbf{x}$ can be expressed as $a\mathbf{1} + b\mathbf{x}$ for some $a, b \in \mathbb{R}$, where $\mathbf{1}$ indicates the vector with all entries of 1, which means that the null space of a linearity test of F is defined by $\mathfrak{L}(\Phi_0) = \{F : F = \Phi_0\alpha \text{ for some } \alpha \in \mathbb{R}^2\}$ with $\Phi_0 = \{\mathbf{1}, \mathbf{x}\}$.

A general class of hypothesis tests on F to examine if F belongs to a pre-specified class of parametric functions can be defined as

$$H_0 : F \in \mathfrak{L}(\Phi_0) \quad vs. \quad H_1 : F \notin \mathfrak{L}(\Phi_0). \tag{5.2}$$

A common Bayesian hypothesis testing procedure is based on Bayes factor (Jeffreys, 1961) which measures the evidence in favor of the alternative hypothesis. Given the hypotheses in (5.2), the Bayes factor in favor of H_1, denoted by $B_{10}(\mathbf{y})$, is defined by the ratio of the marginal

likelihoods of the null and the alternative, $m_0(\mathbf{y})$ and $m_1(\mathbf{y})$, respectively. This is expressible as

$$B_{10}(\mathbf{y}) = \frac{m_1(\mathbf{y})}{m_0(\mathbf{y})} = \frac{\int L(\mathbf{y} \mid \beta, H_1)\pi_1(\beta)d\beta}{\int L(\mathbf{y} \mid \alpha, H_0)\pi_0(\alpha)d\alpha},$$

where $L(\mathbf{y}|\cdot)$ is the likelihood function resulting from each hypothesis, and π_0 and π_1 denote the prior densities under the null and the alternative hypothesis. Note that the large value of $B_{10}(\mathbf{y})$ indicates a strong evidence in support of H_1 while values closer to zero indicate evidence in favor of the null hypothesis.

I note that $F^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F = 0$ if and only if $F \in \mathfrak{L}(\Phi_0)$, where Q_0 is the projection matrix of the null regressor Φ_0, i.e., $Q_0 = \Phi_0(\Phi_0^{\mathrm{T}}\Phi_0)^{-1}\Phi_0^{\mathrm{T}}$. By using this fact, I redefine the hypothesis tests in (5.2) as

$$H_0 : F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F = 0 \quad vs. \quad H_1 : F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F \neq 0.$$

Under H_0, the resulting model is thus the parametric regression model with $F = \Phi_0\alpha$ for some $\alpha \in \mathbb{R}^{d_0}$. By using the semi-norm $F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F$, I consider the null space of the hypothesis as $\{F \in \mathbb{R}^n : F^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F = 0\}$.

In testing a point null hypothesis of a scalar-valued parameter, i.e., $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ for some parameter $\theta \in \mathbb{R}$, Johnson and Rossell (2010) pointed out that the Bayes factor based on local prior densities in favor of H_0 is $O_p(n^{-1/2})$ when the true parameter is the null value. However, when data are generated from alternative hypotheses, the Bayes factor in favor of H_1 increases at an exponential rate of n as discussed in Walker (2004). Johnson and Rossell (2010) showed that the convergence rate of the Bayes factor based on the nonlocal prior densities can be more equitably balanced under the true null and true alternative hypotheses. In Section 5.3, I observed the similar imbalance of the asymptotic rate of the Bayes factor from

the local prior even in nonparametric Bayesian testing. To ameliorate the asymmetry of the convergence rate, I extend the original idea of nonlocal prior densities to functional spaces and introduce *nonlocal functional prior densities* on the coefficient β defined as follows:.

Definition 3. *Let Q_0 be the projection matrix of the null regressor Φ_0. If for any $\epsilon > 0$, there exists $\delta > 0$ such that $\pi(\beta) < \epsilon$ for any $\boldsymbol{\beta}$ with $\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - Q_0)\Phi\beta < \delta$, then I define $\pi(\beta)$ to be a nonlocal functional prior density.*

In contrast, *local prior densities* have strictly positive values of $\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta$ even on the null space $\{\beta \in \mathbb{R}^{K_n} : \beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta = 0\}$. I propose a nonlocal functional prior density $\pi^{NL}(\beta)$ as the product of a nonlocal kernel $h(\beta)$ and a local prior density $\pi_L(\beta)$ as

$$\pi^{NL}(\beta) = E_{\pi_L}\{h(\beta)\}^{-1}h(\beta)\pi_L(\beta), \tag{5.3}$$

where $E_{\pi_L}(\cdot)$ denotes the expectation with respect to the local prior density π_L. Also, the nonlocal kernel $h(\beta)$ satisfies the condition that for any $\epsilon > 0$, there exists $\delta > 0$ such that $h(\beta) < \epsilon$ for any β with $\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - Q_0)\Phi\beta < \delta$, so the resulting prior π^{NL} attains the nonlocal property.

In this dissertation, I consider Gaussian priors as the local base priors to define nonlocal prior densities as in (5.3), which can be expressed as

$$\pi_L(\beta) \sim N\left(\mu, \sigma^2\Sigma_n\right), \tag{5.4}$$

where $\mu \in \mathbb{R}^{K_n}$ and Σ_n are the mean and the covariance of the Gaussian distribution, respectively.

Under the alternative hypothesis, a natural choice of the local prior is Zellner's g-prior Zellner (1986) that is a special case of (5.4) with $\mu = \mathbf{0}$ and $\Sigma_n = g_n(\Phi^{\mathrm{T}}\Phi)^{-1}$ with a hyperparameter g_n. For the null hypothesis, I also impose a g-prior, $N\{\mathbf{0}, \sigma^2 g_n(\Phi_0^{\mathrm{T}}\Phi_0)^{-1}\}$ on

the regression parameters in (5.1). The resulting marginal likelihood is $m_0(\mathbf{y}) = \int L(\mathbf{y} \mid \alpha, \sigma^2, H_0)\pi_0(\alpha)d\alpha$. In the same way, the marginal likelihood of the alternative hypothesis is defined as $m_1(\mathbf{y}) = \int L(\mathbf{y} \mid \beta, \sigma^2, H_1)\pi_1(\beta)d\beta$, where π_1 is the prior on the B-spline coefficients. I denote the marginal density as $m_1^L(\mathbf{y})$ under the alternative hypothesis if the prior on coefficients is local, and $m_1^{NL}(\mathbf{y})$ if the prior is nonlocal.

I propose two classes of nonlocal functional prior densities defined from the following nonlocal kernels in (5.3).

First, *r-th moment functional prior densities* can be defined by the nonlocal kernel

$$h_M^r(\beta \mid r) = \{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta\}^r, \tag{5.5}$$

so the resulting nonlocal functional prior density is

$$\pi_{M^r}(\beta \mid \sigma^2, \mu, \Sigma_n) \propto \{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta\}^r \exp\left\{-\frac{1}{2\sigma^2}(\beta - \mu)^{\mathrm{T}}\Sigma_n^{-1}(\beta - \mu)\right\}. \tag{5.6}$$

Second, I introduce *inverse moment functional prior densities* by applying the nonlocal kernel

$$h_I(\beta \mid \tau_n) = \exp\left[-\sigma^2\tau_n\{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta\}^{-1}\right], \tag{5.7}$$

to obtain the nonlocal density function

$$\pi_I(\beta \mid \sigma^2, \tau_n, \mu, \Sigma_n) \propto \exp\left\{-\frac{1}{2\sigma^2}(\beta - \mu)^{\mathrm{T}}\Sigma_n^{-1}(\beta - \mu) - \frac{\sigma^2\tau_n}{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta}\right\}. \tag{5.8}$$

5.3 Convergence Rates of Bayes Factor

5.3.1 Preliminaries

I define some notation that will be used in the following sections. For sequences a_n and b_n, $a_n \preceq b_n$ and $a_n \prec b_n$ indicate $b_n = O(a_n)$ and $b_n = o(a_n)$, respectively, and $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$.

I define the functional space $C^{\alpha}[0,1]$ to be the space of α_0 times continuously differentiable functions f with $||f||_{\alpha} < \infty$, where α_0 is the greatest integer less than α and the semi-norm $||\cdot||_{\alpha}$ is defined by

$$\|f\|_{\alpha} = \sup_{\{(x,w):x \neq w\}} \frac{|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(w)|}{|x-w|^{\alpha-\alpha_0}}.$$

I also define the empirical L_2 norm as $||f||_{n,2}^2 = \sum_{i=1}^{n} f^2(x_i)/n$ for some function f.

Let P_0 denote the probability measure that generates data $\mathbf{y}$ under the null hypothesis, having true regression function f_0 and $F_0 = \{f_0(x_1), \ldots, f_0(x_n)\}^{\mathrm{T}}$. Let $E_{\beta|y}(\cdot)$ denote the expectation operator with respect to the posterior distribution of β induced by the local prior π^L. Let $E_{\pi^L}(\cdot)$ indicate the expectation operator with respect to the local prior π^L.

5.3.2 Local Priors

I now state a theorem that demonstrates the convergence rate of Bayes factor based on local priors (5.4).

Theorem 7. *Consider the nonparametric regression model* (5.1) *and a hypothesis test on the regression function in* (5.2)*. Suppose that the prior on the B-spline coefficients is the local prior in* (5.4) *with* $\mu = \mathbf{0}$ *and* $\Sigma_n = g_n(\Phi^{\mathrm{T}}\Phi)^{-1}$ *under the alternative hypothesis, and consider a g-prior on the coefficients with a hyper parameter* g_n *for the null hypothesis. Assume that* σ^2 *is known and* $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$*. Then for any diverging sequence* $v_n \to \infty$*,*

$$P_0\left[\left|\log\left\{\frac{m_1^L(\mathbf{y})}{m_0(\mathbf{y})}\right\} - T_n\right| > \{2F_0^{\mathrm{T}}(Q_{\Phi} - Q_0)F_0/\sigma^2 + K_n - d_0\}^{1/2} v_n\right] = o(1),$$

where $m_1^L(\mathbf{y})$ is the marginal likelihood based on the local prior under the alternative hypothesis and

$$T_n = -\frac{K_n - d_0}{2}\log(1+g_n) + \frac{g_n}{2(1+g_n)}\left\{F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0/\sigma^2 + K_n - d_0\right\},$$

where $d_0 = rank(Q_0)$.

Theorem 7 states that asymptotic behavior of the Bayes factor that is derived from the local prior densities is determined by the interplay between $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0$, K_n and g_n. Under H_0, the fact that $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 = 0$ implies that the logarithm of the Bayes factor approximately concentrates on $-(K_n - d_0)\log(1+g_n)/2 + g_n K_n/\{2(1+g_n)\}$, and this quantity would be dominated by the first term $-(K_n - d_0)\log(1+g_n)/2$ when $g_n \succ K_n$. On the other hand, under H_1 there exists a constant δ such that $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0/n > \delta$ for any n, so the convergence rate of the Bayes factor is dominated by $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0/(2\sigma^2) \asymp n$ when $K_n \log g_n \prec n$. When $g_n = O(n)$ as recommended in Zellner (1986) and George and Foster (2000), this means that the asymptotic behavior of the Bayes factor in favor of alternative hypotheses can be summarized as follows.

- For a true null hypothesis, the Bayes factor in favor of the alternative hypothesis decreases only at rate $O_p(n^{-(K_n - d_0)/2})$
- For a true alternative hypothesis, the Bayes factor in favor of the alternative hypothesis increases at rate $O_p(\exp\{cn\})$ for some constant c.

Because K_n should be chosen to be much smaller than n, these Bayes factor rates imply that the resulting hypothesis testing procedure highly tends to provide stronger evidence in favor of a true alternative hypothesis than it does for a true null hypothesis.

These asymptotic results are similar to those obtained from the Bayesian parametric hypothesis tests using local prior densities on a scalar-valued parameter as described in Bahadur

and Bickel (1967), Walker (1969), and Johnson and Rossell (2010). The Bayes factor in favor of alternative hypotheses, when data are generated under an alternative hypothesis, increases exponentially fast. The Bayes factor, when the null hypothesis is true, decreases only at a polynomial rate of n. In nonparametric hypothesis testing, Choi et al. (2009) and Choi and Rousseau (2015) provided similar results of consistency of Bayes factor for semiparametric regression model and partially linear models. Yet neither article discussed the convergence rate of Bayes factor under the true null. Scott and Walker (2015) also derived a similar rate of Bayes factor for a monotonicity test for regression function.

As discussed in Rossell and Telesca (2017), the Bayes factor based on nonlocal priors can be decomposed into a product between the Bayes factor defined by a local prior and the ratio between the posterior expectation and the prior expectation of a nonlocal kernel h. In other words, the Bayes factor $BF_{10}^{NL}(\mathbf{y})$ based on the nonlocal prior derived from (5.3) can be expressed as

$$BF_{10}^{NL}(\mathbf{y}) = \frac{m_1^{NL}(y)}{m_0(y)} = \frac{m_1^{L}(y) D_n(h;\mathbf{y})}{m_0(y)} = BF_{10}^{L}(\mathbf{y}) D_n(h;\mathbf{y}), \tag{5.9}$$

where $BF_{10}^{L}(\mathbf{y})$ be the Bayes factor resulting from the local prior in (5.4), i.e., $BF_{10}^{L}(\mathbf{y}) = m_1^{L}(\mathbf{y})/m_0(\mathbf{y})$, and $D_n(h;\mathbf{y}) = E_{\beta|y}\{h(\beta)\}/E_{\pi_L}\{h(\beta)\}$. Recall that two classes of the nonlocal kernel h are introduced in (5.5) and (5.7). The decomposition in (5.9) means that the Bayes factor $BF_{10}^{NL}(\mathbf{y})$ based on the nonlocal prior with the nonlocal kernel h only differs from the Bayes factor based on the local prior by a product of $D_n(h;\mathbf{y})$. Thus, the asymptotic properties of the Bayes factors derived from the nonlocal priors can be identified by the asymptotic behavior of $D_n(h;\mathbf{y})$ that will be discussed in the following subsections.

5.3.3 Moment Functional Prior Densities

Theorem 8. *Assume that the conditions of Proposition 7 apply. Consider the moment nonlocal function prior* π_{M^r} *in* (5.6) *with the nonlocal kernel* h_M^r *in* (5.5)*, with* $\mu = \mathbf{0}$ *and* $\Sigma_n =$

$g_n(\Phi^{\mathrm{T}}\Phi)^{-1}$. *Suppose* $g_n \to \infty$ *as* $n \to \infty$. *Then, for any diverging sequence* $v_n \to \infty$,

$$P_0\left[\left|D_n(h_M^r;\mathbf{y}) - T_{n,M}^r\right| > \{g_n(K_n - d_0)\}^{-r/2}(T_{n,M}^r)^{1/2}v_n\right] = o(1),$$

where

$$T_{n,M}^r = \left\{\frac{F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 + K_n - d_0}{g_n(K_n - d_0)}\right\}^r,$$

for $r = 1, 2$.

Theorem 8 states that the rate of $D_n(h_M^r;\mathbf{y})$ is determined by the interplay between $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0$ and g_n. Under H_0, $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 = 0$ so that the rate of $T_{n,M}^r$ is g_n^{-r}. On the other hand, under H_1, if $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 \succ K_n$, the rate of Bayes factor is governed by $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0$. The condition $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 \succ K_n$ is reasonable in the sense that the scale of $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0$ is in the order of $n(\succ K_n)$, when the true function f_0 is fixed.

Corollary 9. *Assume that the conditions of Theorem 8 apply. Under* H_1, *assume that* F_0 *and the B-spline basis function satisfy that* $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 \asymp n$. *Suppose that* $g_n \asymp n$ *and* $K_n \prec n$. *Then, under* H_1, $D_n(h_M^r;\mathbf{y}) = O_p(1)$, *and under* H_0, $D_n(h_M^r;\mathbf{y}) = O_p(n^{-r})$ *for* $r = 1, 2$.

Corollary 9 considers a simple setting that $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 \asymp n$ under the alternative hypothesis with the choice of the hyperparameter $g_n \asymp n$. For this setting, the hypothesis test procedures based on moment functional prior densities enjoys the extra penalty $O_p(n^{-r})$ on the Bayes factor in favor of the alternative compared to that from the local prior densities. On the other hand, Corollary 9 states that the convergence rate of Bayes factor is asymptotically invariant compared to that based on the local priors, under a true alternative hypothesis. This indicates that using nonlocal functional prior densities in nonparametric hypothesis tests not only improves the convergence rate of the Bayes factor by a polynomial rate when the null is true, but also does not attenuate the rate when the alternative is true, at least in an asymptotic sense.

5.3.4 Inverse Moment Functional Prior Densities

Theorem 10. *Assume that the conditions of Theorem 8 apply, but now consider the inverse moment functional prior densities in* (5.8) *with the nonlocal kernel* h_I *in* (5.7)*, with* $\mu = \mathbf{0}$ *and* $\Sigma_n = \tau_n(\Phi^{\mathrm{T}}\Phi)^{-1}$*. Then, for any diverging sequence* v_n*,*

$$P_0\left[\frac{M_{n,I}}{-\log D_n(h_I;\mathbf{y})} > v_n\right] = o(1),$$

where $M_{n,I} = \tau_n(d_n + \sigma^2(nd_n)^{1/2})^{-1}$ *for* $d_n = F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 + \sigma^2(K_n - d_0)$*.*

This theorem shows that the under H_0 with $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 = 0$, the rate of $\{-\log D_n(h_I;\mathbf{y})\}^{-1}$ is asymptotically bounded by $(nK_n)^{1/2}\tau_n^{-1}$ in probability, while under H_0 with $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 = 0 \succ K_n$, it is asymptotically bounded by $(nF_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0)^{1/2}\tau_n^{-1}$. The following corollary provides a simpler setting to evaluate the convergence rate of $D_n(h_I;\mathbf{y})$.

Corollary 11. *Assume that the conditions of Theorem 10 apply. Suppose that under* H_1*,* $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 \asymp n$*. Assume that* $\tau_n \asymp n$ *and* $K_n \prec n$*. Then, under* H_1*,* $D_n(h_I;\mathbf{y}) = O_p(1)$*, and under* H_0*,* $D_n(h_I;\mathbf{y}) = O_p(\exp\{-cn^{1/2}K_n^{-1/2}\})$ *for some positive constant* c*.*

Corollary 11 shows that under H_0, the Bayes factor based on the inverse moment functional prior achieves an exponentially faster convergence rate $O_p(\exp\{-cn^{1/2}K_n^{-1/2}\})$ than does the Bayes factor based on the local priors. Moreover, the use of the inverse moment functional prior does not degrade the convergence rate of the Bayes factor in an asymptotic sense when the alternative is true because $D_n(h_I;\mathbf{y}) = O_p(1)$. Therefore, we can expect significant improvement in the convergence rate of the Bayes factor under true null hypotheses and asymptotically the same convergence rate as when the local prior is deployed under true alternative hypotheses.

5.3.5 The Choice of K_n

The asymptotic behavior of Bayesian nonparametric inference based on B-spline basis functions was well-studied in Bontemps (2011) and Ghosal and van der Vaart (2007). In particular, Ghosal and van der Vaart (2007) showed that the minimax rate of $n^{-\alpha/(1+2\alpha)}$ for posterior concentration under the L_2 norm can be achieved by setting $K_n \asymp n^{1/(1+2\alpha)}$ when $f_0 \in C^\alpha[0,1]$. Similar results were obtained in a frequentist perspective in Zhou et al. (1998); Claeskens et al. (2009). However, the asymptotic results regarding the Bayes factors that are discussed in this section do not require the optimal condition on K_n ($\asymp n^{1/(1+2\alpha)}$). The following proposition illustrates why this is so.

Proposition 12. *Suppose that $f_0 \in C^\alpha[0,1]$ and $K_n \to \infty$ as n tends to ∞. Then, $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 \asymp F_0^{\mathrm{T}}(I - Q_0)F_0$.*

This proposition shows that the asymptotic behavior of $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0$ is solely determined by $F_0^{\mathrm{T}}(I - Q_0)F_0$, without any dependence on the rate of K_n. Even though the asymptotic estimation performance would be sub-optimal when the rate of K_n is misspecified, the convergence rates of the Bayes factors discussed in the previous theorems are still valid. When the data are generated under true alternative hypotheses (i.e. $F_0^{\mathrm{T}}(I - Q_0)F_0 \asymp n$), Proposition 12 guarantees that the condition $F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0 \asymp n$ in Corollary 9 and Corollary 11 is satisfied. Therefore, the results of these corollaries hold with any diverging K_n.

5.4 Examples of Bayesian Hypothesis Tests Using Nonlocal Functional Priors

In this section, I examine the behavior of Bayes factors based on different priors in finite samples. I first consider a simple hypothesis setting to test the sparsity of the regression function in (5.1).

$$H_0 : F = \mathbf{0} \quad vs. \quad H_1 : F \neq \mathbf{0}. \tag{5.10}$$

Here, the projection matrix on the null space Q_0 is equivalent to a matrix with all zero entries, resulting in $\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta = \beta^{\mathrm{T}}\Phi^{\mathrm{T}}\Phi\beta$. Another example is a hypothesis test for linearity specified as

$$H_0 : F \text{ is linear} \quad vs. \quad H_1 : F \text{ is not linear.} \tag{5.11}$$

In this case, the null regressors Φ_0 can be defined as $\{\mathbf{1}, \mathbf{x}\}$ and Q_0 is the projection matrix derived from Φ_0.

I also consider a hypothesis test on the coefficient function in a varying coefficient model as introduced in Hastie and Tibshirani (1993). This model can be defined as $y_i = t_i f(x_i) + \epsilon_i$, where ϵ_i is i.i.d $N(0, \sigma^2)$. I also assume that $\mathbf{x} = \{x_i\}_{1\le i\le n}$ and $\mathbf{t} = \{t_i\}_{1\le i\le n}$ are independent variables. Some practitioners might be interested in testing if F is constant so that the resulting model is equivalent to a simple linear regression model. To test this hypothesis, I construct the contrasting hypotheses on the coefficient function F.

$$H_0 : F \text{ is constant} \quad vs. \quad H_1 : F \text{ is not constant.} \tag{5.12}$$

The resulting nonlocal prior densities can be generated by setting $\Phi_0 = \{\mathbf{1}\}$.

To compare the performance of the local and nonlocal functional prior densities, I considered several functions for the hypothesis tests as follows:

$$\begin{aligned} f_{Sp}(u) &= 0, & f_Q(u) &= u^2 \\ f_C(u) &= 1, & f_s(u) &= \sin(u) \\ f_L(u) &= u, & f_{pL}(u) &= (u-\pi/3)_+ - (-\pi/3-u)_+, \end{aligned} \tag{5.13}$$

where $u \in [-\pi, \pi]$ and $(\cdot)_+$ is the truncation function on negative values by zero. The null

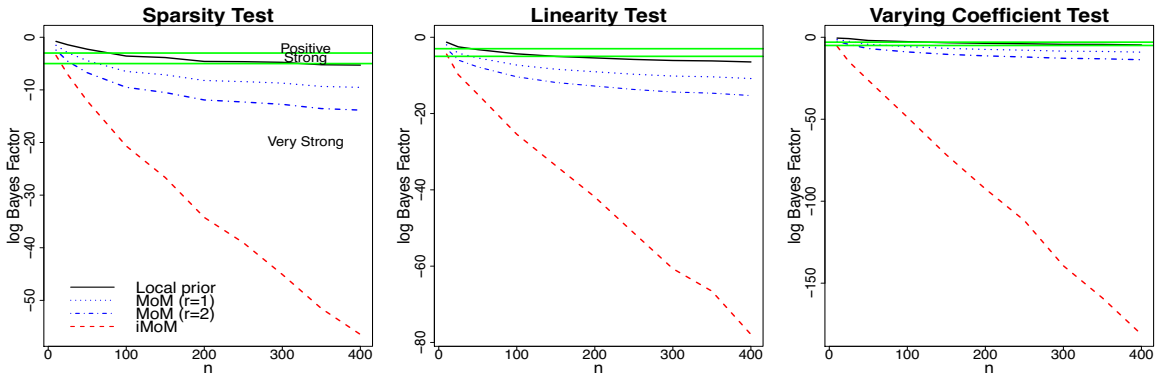


Figure 5.1: When the null hypothesis is true, the averaged logarithm of Bayes factor in favor of the alternative hypothesis based on the nonlocal prior and the local prior densities for sparsity, linearity, and varying coefficient test with varying sample size n. "MOM" and "iMOM" indicate the moment functional prior and the inverse moment functional prior. The two horizontal green lines are on -3 and -5 of Bayes factor.

functions of the hypothesis tests for sparsity, linearity, and varying coefficient are f_{Sp}, f_L, and f_C, respectively. I generated independent variables **x** and **t** from a uniform distribution on $(-\pi, \pi)$. Given a regression function f, the corresponding dependent variable is generated from $y_i = Af(x_i) + \epsilon_i$, where ϵ_i follows a $N(0, \sigma^2)$ with $\sigma = 1/2$ for $i = 1, \cdots, n$ and A is a constant that models $Var\{Af(x)\} = 1$. This specification controls the signal-to-noise ratio for different regression functions. For varying coefficient models, I simulated the dependent variable by setting $y_i = Rx_i f(t_i) + \epsilon_i$, where R is chosen by solving $Var\{Rxf(t)\} = 1$. One hundreds replicated data sets were used for each simulation setting with sample sizes varying from 10 to 400. I also set $K_n = 4$, $g_n = n/K_n$ and $\tau_n = n$. I use a simple Monte Carlo simulation to evaluate the Bayes factors. Because the posterior distribution of β based on the local prior has a closed form (Gaussian), the posterior expectation of the nonlocal kernels can be easily evaluated from the Gaussian samples of the posterior distribution.

The performance comparison between the nonlocal functional priors and the local priors is illustrated in Figure 5.1 for data generated under the null hypothesis. As Figure 5.1 shows, evidence in favor of the alternative hypothesis resulting from the local prior density in (5.4)

	Test	Sparsity		Linearity			VC	
	True function	f_L	f_s	f_{pL}	f_Q	f_s	f_L	f_s
$(n = 50)$	$\log BF^L$	48.27	43.02	0.49	35.87	1.40	38.22	37.10
	$MOM^{r=1}$	1.18	1.15	-0.60	1.64	0.79	0.21	0.16
	$MOM^{r=2}$	1.99	1.92	-1.62	2.62	0.96	-0.05	-0.16
	$iMOM$	1.54	1.41	-3.67	2.28	0.55	-0.36	-0.53
$(n = 100)$	$\log BF^L$	193.18	187.94	11.18	176.85	66.80	181.67	177.99
	$MOM^{r=1}$	1.12	1.10	-0.68	1.72	0.80	0.17	0.18
	$MOM^{r=2}$	1.85	1.80	-1.96	2.76	0.93	-0.16	-0.16
	$iMOM$	1.59	1.56	-5.62	2.66	0.95	-0.33	0.31
$(n = 400)$	$\log BF^L$	789.47	780.95	55.07	778.07	293.74	775.11	733.18
	$MOM^{r=1}$	1.10	1.10	-0.75	1.78	0.82	0.18	0.14
	$MOM^{r=2}$	1.80	1.78	-2.16	2.87	0.95	-0.15	-0.24
	$iMOM$	1.62	1.61	-7.63	2.77	1.12	-0.24	-0.38

Table 5.1: Under alternative hypotheses, the expectation of logarithm of Bayes factor and the $D_n(h; \mathbf{y})$ for nonparametric Bayesian hypothesis tests: Sparsity test in (5.10), Linearity test in (5.11), and Varying Coefficient test (VC) in (5.12). $\log BF_{\pi^L}$ denotes the averaged logarithm of Bayes factor in favor of the alternative hypothesis based on the local prior. $MOM^{r=1}$, $MOM^{r=2}$, and $iMOM$ indicates the average of the logarithm of $D_n(h_M^{r=1}; \mathbf{y})$, $D_n(h_M^{r=2}; \mathbf{y})$, and $D_n(h_I; \mathbf{y})$, respectively, for each alternative function over 100 replicates, and the data-generating true function f_L, f_s, f_{pL} and f_Q are defined in (5.13).

decreases much slower than that of the moment functional prior or the inverse moment functional prior densities when the data were generated from a model that is consistent with the null hypothesis. Even with small size of samples the moment or inverse moment functional prior provided “very strong" support in favor of the null hypothesis. The logarithm of the Bayes factor in favor of the alternative was less than -5. On the other hand, the local prior (5.4) requires a relatively large sample size to attain the same strength of evidence of the null hypothesis as the nonlocal functional priors do. Moreover, as discussed by Johnson and Rossell (2012) in parametric model selection, the local prior densities provide evidence in favor of the null which is not strong enough so that it fails to achieve desirable model selection consistency when a diverging number of models or hypotheses is considered.

I note that nonlocal functional prior densities often provide stronger evidence in favor of the null hypothesis, especially when the discrepancy of the true regression function and the null space, $F_0^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F_0/n$, is expected to be small. In Table 1, most considered alternative models showed negligible differences between the Bayes factors based on the nonlocal prior and the local prior densities. However, when the piece-wise linear function f_{pL} was adopted as an alternative hypothesis for the linearity test, the Bayes factor in favor of alternative hypotheses based on the inverse moment nonlocal priors is significantly attenuated compared to that of the local prior densities. This result stems from the fact that the shape of the piece-wise linear function is quite similar to that of a linear function. So, the discrepancy measure $f_{pL}(\mathbf{x})^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)f_{pL}(\mathbf{x})/n$ is expected to be much smaller than that from the other alternative functions considered, resulting in non-negligible $\log D_n(h_I;\mathbf{y})$.

5.5 Nonparametric Additive Model Selection Using Nonlocal Functional Priors

In this section, I apply the proposed nonlocal functional priors to model selection problems for high-dimensional nonparametric additive models as in (4.14). I denote the true model, which is the index set of variables involved in the data-generating process, by $\mathbf{t}$. I denote the

true marginal regression function of X_j for $j \in \mathbf{t}$ by $f_{0,j}$ and the true regression function by $f_0 = \sum_{j \in \mathbf{t}} f_{0,j}$. I also denote the empirical realization of the true regression function and the true marginal regression functions by F_0 and $F_{0,j}$, respectively.

I model each marginal regression function F_j by a linear combination of B-spline basis functions of the term $F_j = \Phi_j \beta_j$ for $\beta_j \in \mathbb{R}^{K_n}$. Here, $\Phi_j = \{\phi_l(x_j)\}_{l=1,\ldots,K_n}$, and ϕ_l is the l-th B-spline basis function for $l = 1, \ldots, K_n$ and $j = 1, \ldots, p$. I define $\Phi_{\mathbf{k}}$ to be a set of basis function for the covariates in model $\mathbf{k}$, i.e., $\Phi_{\mathbf{k}} = \{\Phi_j\}_{j \in \mathbf{k}}$.

For a given model $\mathbf{k}$, let

$$\widehat{\beta}_{\mathbf{k}} = (\Phi_{\mathbf{k}}^{\mathrm{T}} \Phi_{\mathbf{k}})^{-1} \Phi_{\mathbf{k}}^{\mathrm{T}} \mathbf{y}, \quad \widehat{F}_{\mathbf{k}} = \Phi_{\mathbf{k}} \widehat{\beta}_{\mathbf{k}}, \quad \text{and} \quad P_{\mathbf{k}} = \Phi_{\mathbf{k}} (\Phi_{\mathbf{k}}^{\mathrm{T}} \Phi_{\mathbf{k}})^{-1} \Phi_{\mathbf{k}}^{\mathrm{T}}. \tag{5.14}$$

For $1 \leq j \leq p$, define

$$\widehat{\beta}_j = (\Phi_j^{\mathrm{T}} \Phi_j)^{-1} \Phi_j^{\mathrm{T}} \mathbf{y}, \quad \widehat{F}_j = \Phi_j \widehat{\beta}_j, \quad \text{and} \quad P_j = \Phi_j (\Phi_j^{\mathrm{T}} \Phi_j)^{-1} \Phi_j^{\mathrm{T}}. \tag{5.15}$$

Given a model $\mathbf{k}$, I consider a nonlocal functional prior for the additive model that is a product of independent inverse moment functional priors (5.8) as

$$\pi^{NL}(\beta_{\mathbf{k}} \mid \mathbf{k}, \sigma^2, \tau_n) \propto \prod_{j \in \mathbf{k}} \exp\left\{ -\frac{\beta_j^{\mathrm{T}} \beta_j}{2\sigma^2 \tau_n} - \frac{\sigma^2 \tau_n}{\beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j \beta_j} \right\}. \tag{5.16}$$

Here, β_j is a K_n-dimensional coefficient vector for the B-splines basis functions corresponding to x_j for $j = 1, \ldots, p$ and Q_0 is the projection matrix of an n-dimensional one vector. This prior assigns zero density to the space of constant functions since $F^{\mathrm{T}}(I - Q_0)F = 0$ when F is constant. A constant marginal regression function implies that the corresponding covariate is not associated with the response. Thus, the proposed prior for the additive model induces a nonlocal functional prior for model selection. For high-dimensional additive model selection, I only focus on the inverse moment prior. Even though model selection procedures based on

the moment functional priors are computationally efficient to implement, the convergence rates resulting Bayes factor are not strong enough to control for multiplicity. The resulting model selection procedures thus fail to achieve model selection consistency in high-dimensional settings.

In addition to imposing priors on the B-spline coefficients given a model, I place a prior on the model space to complete the prior specification. I consider a uniform prior on the model space restricted to models having size less than or equal to q_n, with $q_n < n$. That is, the prior on the model space can be written as

$$\pi(\mathbf{k}) \propto I(|\mathbf{k}| \leq q_n), \tag{5.17}$$

where $I(\cdot)$ denote the indicator function. With a slight abuse of notation, I denote the prior on the space of models by π as well.

One might also consider nonuniform model priors on the model space. For instance, the following model prior is introduced by Castillo et al. (2015) in high-dimensional model selection for linear models:

$$\pi(\mathbf{k}) \propto \binom{p}{|\mathbf{k}|}^{-1} a_1^{-|\mathbf{k}|} p^{-a_2|\mathbf{k}|}, \;\; a_1, a_2 > 0. \tag{5.18}$$

This prior strongly penalizes large-sized models when p is large. Castillo et al. (2015) derived the posterior contraction rate and model selection consistency for linear model selection problems based on this model prior. However, for nonparameteric additive model selection in high-dimensional settings, the asymptotic properties of the procedure have not been investigated. In contrast, I show that the model selection procedure based on the nonlocal functional priors can achieve model selection consistency without the stronger prior on the model space.

Based on the defined priors, the posterior distribution is defined by

$$\pi(\mathbf{k} \mid \mathbf{y}) = \frac{m_{\mathbf{k}}(\mathbf{y})\pi(\mathbf{k})}{\sum_{\mathbf{l}} m_{\mathbf{l}}(\mathbf{y})\pi(\mathbf{l})},$$

where $m_{\mathbf{k}}(\mathbf{y}) = \int L(\mathbf{y} \mid \mathbf{k}, \beta_{\mathbf{k}}, \sigma^2)\pi(\beta_{\mathbf{k}} \mid \mathbf{k}, \sigma^2)d\beta_{\mathbf{k}}$.

In the following sections, I illustrate some desirable theoretical properties of the model selection procedure based on the proposed prior in (5.16).

5.5.1 Additive Model Selection Consistency for High-dimensional Settings

I first state the regularity conditions that are assumed.

(A1) The true model $\mathbf{t}$ is fixed regardless of n and p.

(A2) For any $\mathbf{k}$ with $|\mathbf{k}| \leq q_n$, where q_n is defined in (5.17), there exist positive sequences ζ_{n*} and ζ_n^* such that

$$\zeta_{n*} \sum_{j=1}^{|\mathbf{k}|} \beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_j \beta_j \leq \beta_{\mathbf{k}}^{\mathrm{T}} \Phi_{\mathbf{k}}^{\mathrm{T}} \Phi_{\mathbf{k}} \beta_{\mathbf{k}} \leq \zeta_n^* \sum_{j=1}^{|\mathbf{k}|} \beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_j \beta_j.$$

for any $\beta_{\mathbf{k}} = \{\beta_j\}_{j \in \mathbf{k}}$.

(A3) There exist positive constants λ_* and λ^* such that,

$$\frac{n\lambda_*}{K_n}\beta^{\mathrm{T}}\beta \leq \min_{j=1,\ldots,p} \beta^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_j \beta \leq \max_{j=1,\ldots,p} \beta^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_j \beta \leq \frac{n\lambda^*}{K_n}\beta^{\mathrm{T}}\beta,$$

for any $\beta \in \mathbb{R}^{K_n}$.

(A4) For any $\mathbf{k}$ with $|\mathbf{k}| \leq q_n$ and $j \notin \mathbf{k}$, $\beta^{\mathrm{T}} \Phi_j^{\mathrm{T}} P_{\mathbf{k}} \Phi_j \beta \leq \beta^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - P_{\mathbf{k}}) \Phi_j \beta$ for any $\beta \in \mathbb{R}^{K_n}$.

(A5) For $j \in \mathbf{t}$, $F_0^{\mathrm{T}} P_j F_0 / n$ converges to some constant c_j as n tends to ∞, and for $j \notin \mathbf{t}$, $F_0^{\mathrm{T}} P_j F_0 \prec \log p$. Also, $\min_{\mathbf{k}:\mathbf{t} \not\subset \mathbf{k}, |\mathbf{k}| \leq q_n} \{F_0^{\mathrm{T}} (P_{\mathbf{k} \cup \mathbf{t}} - P_{\mathbf{k}}) F_0\} \succ n / \log n$ and

$\max_{\mathbf{k}:|\mathbf{k}| \leq q_n} \{F_0^{\mathrm{T}} (P_{\mathbf{k} \cup \mathbf{t}} - P_{\mathbf{t}}) F_0\} \prec q_n \log p$.

The condition **(A2)** is essential for model identifiability. When some basis functions evalu-

ated at the observed covariates are extremely correlated, any model selection procedure for the additive models would fail to distinguish the corresponding variables with the highly correlated basis functions. This results in identifiability issues between marginal functions. The condition **(A3)** uniformly controls the maximum and minimum eigenvalues of marginal basis matrices Φ_j, for $j = 1, \ldots, p$. As stressed in Ghosal and van der Vaart (2007), the B-spline basis matrix for a single covariate is asymptotically isotropic, i.e., there exist constants C_1 and C_2 such that, as n increases for any $\beta \in \mathbb{R}^{K_n}$,

$$C_1 \frac{n}{K_n} \beta^{\mathrm{T}} \beta \leq \beta^{\mathrm{T}} \Phi^{\mathrm{T}} \Phi \beta \leq C_2 \frac{n}{K_n} \beta^{\mathrm{T}} \beta,$$

where Φ is the B-spline basis matrix of a single variable. However, this result does not assure that the isotropic property holds uniformly over all basis matrices under high-dimensional settings. Since the marginal likelihoods contains the determinant of the basis matrix corresponding to a model, it is necessary to set **(A3)** as a regularity condition on the basis matrices to evaluate the convergence rate of the marginal likelihoods.

Let $\pi(\mathbf{t} \mid \mathbf{y})$ denote the posterior model probability of the true model obtained under the product of inverse moment functional prior densities in (5.16) on the B-spline coefficients and a truncated uniform prior on all models of size less than or equal to q_n in (5.17). I now illustrate that model selection procedures based on the product of inverse moment functional priors is consistent in the sense that the posterior true model probability converges to one in probability as n increases, even when the number of predictors increases at sub-exponential rate.

Theorem 13. *Suppose (A2)–(A5) hold with* $K_n \prec \min\{n^{1/2}, \log p\}$ *and* $q_n \prec \log n$. *Assume* σ^2 *is known. Suppose that there exists* $\eta \in (0, 1)$ *such that* $\log p = O(n^{\eta})$. *Assume that* $\zeta_{n*} \succ \log p / \{n(\log n)^2\}$ *and* $\zeta_n^* = O(1)$. *If* $\zeta_{n*}^{-1/2} u_n^{1/2} (\zeta_{n*}^{-1} + K_n + \log p)^{1/2} n^{1/2} \log p \prec \tau_n \prec$

$\zeta_{n*}^{-1/2}(\zeta_{n*}^{-1} + K_n + \log p)^{1/2} n^{3/2} / \log n$, *where* $u_n = q_n^2 (\log n)^2$,

$$\pi(\mathbf{t} \mid \mathbf{y}) \xrightarrow{p} 1.$$

Theorem 13 imposes constraints on the hyperparameter τ_n that are determined by the dimension p, K_n and ζ_{n*}. In particular, when some basis functions evaluated at the observed covariates are highly correlated in the sense that ζ_{n*} decreases at a faster rate than those of K_n and $\log p$, the rate of τ_n resulting model selection consistency is asymptotically determined by the rate of ζ_{n*}.

5.5.2 Asymptotic Rates of Marginal Likelihood for Additive Models

In this section, I discuss a unique property of nonlocal functional priors that distinguishes it from the local priors for additive models. I shall show that the marginal likelihood from the nonlocal functional prior attains a very different form from that of local priors (5.21).

For a given model $\mathbf{k}$, the convergence rate of the logarithm of the marginal likelihood $m_{\mathbf{k}}^L(\mathbf{y})$ based on the local prior with a hyper parameter g_n in (5.21) can be expressed as

$$\log m_{\mathbf{k}}^L(\mathbf{y}) \approx \log L(\mathbf{y} \mid \widehat{\beta}_{\mathbf{k}}, \mathbf{k}) - \frac{|\mathbf{k}| K_n}{2} \log g_n, \tag{5.19}$$

where $\beta_{\mathbf{k}}$ is defined in (5.14) and $\pi(\mathbf{k})$ is a model prior. An important point is that the penalty on the model $\mathbf{k}$ depends solely on the model size $|\mathbf{k}|$. In other words, adding one extra variable to a model penalizes the marginal likelihood by $(K_n \log g_n)/2$, regardless of the strength of the signal from the estimated marginal regression function.

On the other hand, the the asymptotic rate of the logarithm of the marginal likelihood based on the nonlocal function prior involves a totally different penalty on model $\mathbf{k}$. This penalty is adaptively determined by the estimated marginal function $\widehat{F}_j$ for $j \in \mathbf{k}$. Under the regularity conditions considered in Section 5.5.1, the asymptotic marginal likelihood $m_{\mathbf{k}}^{NL}(\mathbf{y})$ from the

nonlocal functional priors can be written as

$$\log m_{\mathbf{k}}^{NL}(\mathbf{y}) \approx \overbrace{\log L(\mathbf{y} \mid \widehat{\beta}_{\mathbf{k}}, \mathbf{k}) - \frac{|\mathbf{k}|K_n}{2}\log \tau_n}^{(A)} \overbrace{- \sum_{j \in \mathbf{k}} \frac{c\sigma^2 \tau_n}{\widehat{F}_j^{\mathrm{T}}(I - Q_0)\widehat{F}_j + \epsilon_n}}^{(B)}, \tag{5.20}$$

where $\widehat{F}_j$ is defined in (5.15) and $\epsilon_n = \{n\widehat{F}_j^{\mathrm{T}}(I - Q_0)\widehat{F}_j\}^{1/2}$. This can be shown by using Lemma A.3.1 in Appendix. The term (A) in (5.20) is exactly the same as the rate of marginal likelihoods defined by local priors. Additional to (A), the rate of the marginal likelihood based on the nonlocal prior attains an extra penalty term (B). This penalty term adapts to the semi-norm $\widehat{F}_j^{\mathrm{T}}(I - Q_0)\widehat{F}_j$ for eah $j \in \mathbf{k}$. When $j \in \mathbf{t}$, the assumption **(A5)** implies that the marginal penalty term for the j-th variable is $O_p(\tau_n/n)$. On the other hand, when $j \notin \mathbf{t}$, the marginal term is $O_p(\tau_n/(n \min\{K_n, \log p, \zeta_{n*}^{-1}\})^{1/2})$ since $\widehat{F}_j^{\mathrm{T}}(I - Q_0)\widehat{F}_j$ follows a noncentral chi square distribution $\chi^2_{K_n}(R_n)$, where the noncentral parameter is $R_n = F_0^{\mathrm{T}} P_j (I - Q_0) P_j F_0 \prec \log p$. The latter property again follows from the assumption **(A5**. This property of nonlocal functional priors shows that a model containing variables in the true model will be weakly penalized, and a model with any spurious variables will be strongly penalized, since $\min\{K_n, \log p, \zeta_{n*}^{-1}\} \prec n$. This property results in a promising performance of model selection by nonlocal functional priors with finite samples. Simulated and real data sets are provided to examine the model selection performance of my procedure in Section 5.5.4 and Section 5.6.

The posterior model probability $\pi(\mathbf{k} \mid \mathbf{y})$ of a model $\mathbf{k}$ is proportional to the product of the marginal likelihood and the model prior. In the above cases, under a uniform model prior, i.e. $\pi(\mathbf{k}) \propto 1$, the logarithm of the posterior model prior is asymptotically equivalent to the logarithm of the marginal likelihood. One can also consider a nonuniform model prior in (5.18) on the model space. By using Stirling's approximation, one can show that that model prior is asymptotically equivalent to $a_1^{-|\mathbf{k}|} p^{-(1+a_2)|\mathbf{k}|}$. The rate of marginal likelihood can be matched

to the rate of the model prior. For example, the logarithm of the model prior for $a_1 = 1$ and $a_2 = 1$, plus the log-marginal likelihood with the local prior for $g_n = n$, is asymptotically equivalent to the logarithm of the marginal likelihood of the local prior for $g_n = np^{4/K_n}$. Since the logarithm of the marginal likelihood of the nonlocal functional prior can be expressed as the sum of the logarithm of the marginal likelihood of the local prior and the extra penalty term, the model prior in (5.18) is embedded in the marginal likelihood of the nonlocal functional prior.

5.5.3 Computational Strategy Using S5

In the previous sections, I have discussed desirable theoretical properties of nonlocal functional priors. From a computational viewpoint, implementing Bayesian variable selection procedures for high-dimensional additive models is a challenging problem. For high-dimensional model selection, full posterior sampling using Markov Chain Monte Carlo (MCMC) algorithms, such as the reversible jump MCMC (Green, 1995), is highly inefficient and often not feasible from a practical perspective. Recently, Shin et al. (2017) proposed a scalable algorithm called the Simplified Shogun Stochastic Search with Screening (S5) that is optimized to efficiently explore high-dimensional model spaces in Bayesian variable selection for linear models. The S5 algorithm is a simplified version of an existing search algorithm, Shogun Stochastic Search (SSS) by Hans et al. (2007) and utilizes a screening step. The screening step is embedded in the algorithm to reduce the model space to be searched. Shin et al. (2017) empirically showed that S5 efficiently searches the model space and dramatically reduces the computational burden in exploring linear models.

Here, I modify S5 to be suitable for high-dimensional additive models by adding a nonparametric screening step called the Iterative Nonparametric Independence Screening (INIS; Fan et al. (2011)). For a given model $\mathbf{k}$, the first step in INIS is to calculate the residual $r_{\mathbf{k}}$ of the additive model using its B-spline least square estimator. Second, the residuals are used to

evaluate the nonparametric screened set $\mathbf{S}_{\mathbf{k}}^{INIS}(M)$ with a screening size M defined as

$$\mathbf{S}_{\mathbf{k}}^{INIS}(M) = \{j \in \{1, \ldots, p\} : rank(\|\widehat{F}_{\mathbf{k},j}\|_{n,2}^2) \leq M\},$$

where $\widehat{F}_{\mathbf{k},j} = P_j r_{\mathbf{k}}$ for $j = 1, \ldots, p$. The function $rank(a_j)$ for some a_j in $\{a_l\}_{1 \leq l \leq p}$ is a rank function that evaluates the decent order of a_j in $\{a_l\}_{1 \leq l \leq p}$. This screened set $S_{\mathbf{k}}^{INIS}(M)$ includes the M top variables that have the large empirical L_2 norm of teh estimated marginal function. I then restrict the S5 algorithm to the screened set of variables $S_{\mathbf{k}}^{INIS}(M)$ whose cardinality is M ($\ll p$), so that the target model space can be significantly reduced and the computation can be highly accelerated. The screening step is performed and the screened set is updated in every iteration in S5, so even when some significant variables are ignored in early iterations, they can re-enter the model in subsequent iterative screening. The formal statement of the algorithm is provided in Appendix. For more details and discussions about S5, see Section 3.

To evaluate the marginal likelihood for each model, I used the Laplace approximation. The derivation of this procedure is described in Appendix. All computational functions used in this dissertation are provided in the `R` package `BayesS5`.

5.5.4 Simulation Studies

To examine the performance of the model selection procedure based on the nonlocal functional priors, I considered several simulations settings that were perviously proposed by others. However, I used a different sample size n and dimension p for each setting to examine high-dimensional settings. Let SNR $= \mathrm{Var}(f_0(\mathbf{x}))/Var(\epsilon)$ denote the signal-to-noise ratio, where f_0 is the true underlying function; i.e. $f_0 = \sum_{j \in \mathbf{t}} f_j(X_j)$.

Scenario 1: ($n = 150$, $p = 3000$, SNR ≈ 15) The true model is

$$y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + \epsilon_i,$$

where $\epsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, 1)$ for $i = 1, \ldots, n$, with

$$\begin{aligned} f_1(x) &= -\sin(2x), \;\; f_2(x) = x^2 - 25/12, \;\; f_3(x) = x, \\ f_4(x) &= \exp\{-x\} - 2/5 \cdot \sinh(5/2). \end{aligned}$$

The covariates are independently generated from a uniform distribution between -2.5 to 2.5. Similar settings with this were considered in multiple articles such as Härdle et al. (2012), Meier et al. (2009) and Ravikumar et al. (2009).

Scenario 2: ($n = 150$, $p = 3000$, SNR ≈ 6.7) The same scenario was considered in Meier et al. (2009), but the dimension and the sample size in that paper are different from this scenario ($n = 100$ and $p = 1000$ in Meier et al. (2009)). This scenario is similar to *Scenario 1*, but the covariates are instead generated from a Gaussian distribution with a covariance matrix Σ, where $\Sigma_{k,j} = 0.5^{|k-j|}$ for $1 \leq k, j \leq p$; i.e., $x_i \sim \mathrm{N}(0, \Sigma)$ for $i = 1, \ldots, n$.

Scenario 3: ($n = 200$, $p = 3000$, SNR ≈ 3.11). I consider a scenario that was examined in Huang et al. (2010). The true model is

$$y_i = 5f_5(x_{i1}) + 3f_6(x_{i2}) + 4f_7(x_{i3}) + 6f_8(x_{i4}) + \epsilon_i,$$

where

$$
\begin{aligned}
f_5(x) &= x,\ \ f_6(x) = (2x-1)^2,\ \ f_7(x) = \sin(2\pi x)/(2-\sin(2\pi x)), \\
f_8(x) &= 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3.
\end{aligned}
$$

The covariates are generated as follows. First, I generate W_{ij}, V_i, and U_i independently from s $\mathrm{N}(0,1)$ truncated to the interval, $[0,1]$ for $i = 1,\ldots,n$ and $p = 1,\ldots,p$. Then I set $x_{ij} = (W_{ij} + U_i)/2$ for $j = 1,\ldots,4$ and $x_{ij} = (W_{ij} + V_i)/2$ for $j = 5,\ldots,p$. This guarantees that the variables in the true model and the spurious variables are independent.

Scenario 4: ($n = 400$, $p = 1000$, SNR ≈ 11.25) This scenario was considered in Meier et al. (2009), and a similar example was also considered in Lin and Zhang (2006). The same functions are used as in *Scenario 3* and the covariates are independently and uniformly generated on the interval (0,1). The model is

$$
\begin{aligned}
y_i &= f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) \\
&\quad + 1.5f_1(x_{i5}) + 1.5f_2(x_{i6}) + 1.5f_3(x_{i7}) + 1.5f_4(x_{i8}) \\
&\quad + 2.5f_1(x_{i9}) + 2.5f_2(x_{i10}) + 2.5f_3(x_{i11}) + 2.5f_4(x_{i12}) + \epsilon_i,
\end{aligned}
$$

where $\epsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, 0.5184)$ for $i = 1,\ldots,n$.

For comparisons with a classical local prior, I consider a simple local prior defined by a product of g-priors, expressible as

$$
\pi^L(\beta_{\mathbf{k}} \mid \mathbf{k}, \sigma^2) \propto \prod_{j \in \mathbf{k}} \exp\left\{ -\frac{\beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_j \beta_j}{2\sigma^2 g_n} \right\}. \tag{5.21}
$$

To my best knowledge there do not exist references regarding theoretical properties about the model selection procedure using the prior in (5.21) for high-dimensional additive mod-

	Scenario 1		*Scenario 2*		*Scenario 3*		*Scenario 4*	
Methods	MSE	MSPE	MSE	MSPE	MSE	MSPE	MSE	MSPE
NLfP	0.147	1.270	0.162	1.649	0.891	3.127	0.271	0.927
g-prior	0.148	1.270	0.157	1.649	0.916	3.052	0.268	0.924
HGAM	2.128	6.501	1.706	5.905	1.236	3.544	0.656	1.452
SpAM	0.496	2.189	0.575	3.347	0.740	3.294	0.271	1.161
AdapGL	0.492	1.989	0.563	2.873	0.902	3.939	0.305	1.293

Table 5.2: Optimal MSE and MSPE of each method for the considered settings.

els. However, the use of the prior is natural for the variable selection in additive models since each component function is modeled by a linear combination of B-spline basis functions. For the Bayesian procedures, I used a noninformative prior on σ^2 proportional to $1/\sigma^2$. I set the number of the basis functions $K_n = 5$ for all simulation scenarios.

I compared results from the Bayesian procedures to several penalized likelihood approaches. These approaches included the following. Ravikumar et al. (2009) introduced a penalized likelihood method called *Sparse Additive Model* (SpAM) for model selection for additive models. The penalty term of SpAM can be expressed as a weighted group LASSO penalty (Yuan and Lin, 2006). Meier et al. (2009) proposed *High-dimensional Generalized Additive Model* (HGAM) combining sparsity on the coefficients of basis functions and regularization on the smoothness of the marginal regression function. Huang et al. (2010) introduced a two-step procedure using adaptive group LASSO (AdapGL) for variable selection in additive models, which first derives the weights of the group penalty, then applies it to the adaptive group LASSO penalty. In the simulations, I used the `R` packages `SAM`, `hgam` and `grpLasso` to implement SpAM, HGAM, AdapGL, respectively.

To compare the performances of the procedures independently of the choice of tuning parameters, I used Precision-Recall (PR) curves. PR curves plot the precision=TP/(TP+FP) versus the recall (or senstivity)=TP/(TP+FN), where TP, FP and FN respectively denote the number of true positives, false positives and false negatives, as the tuning parameter varies

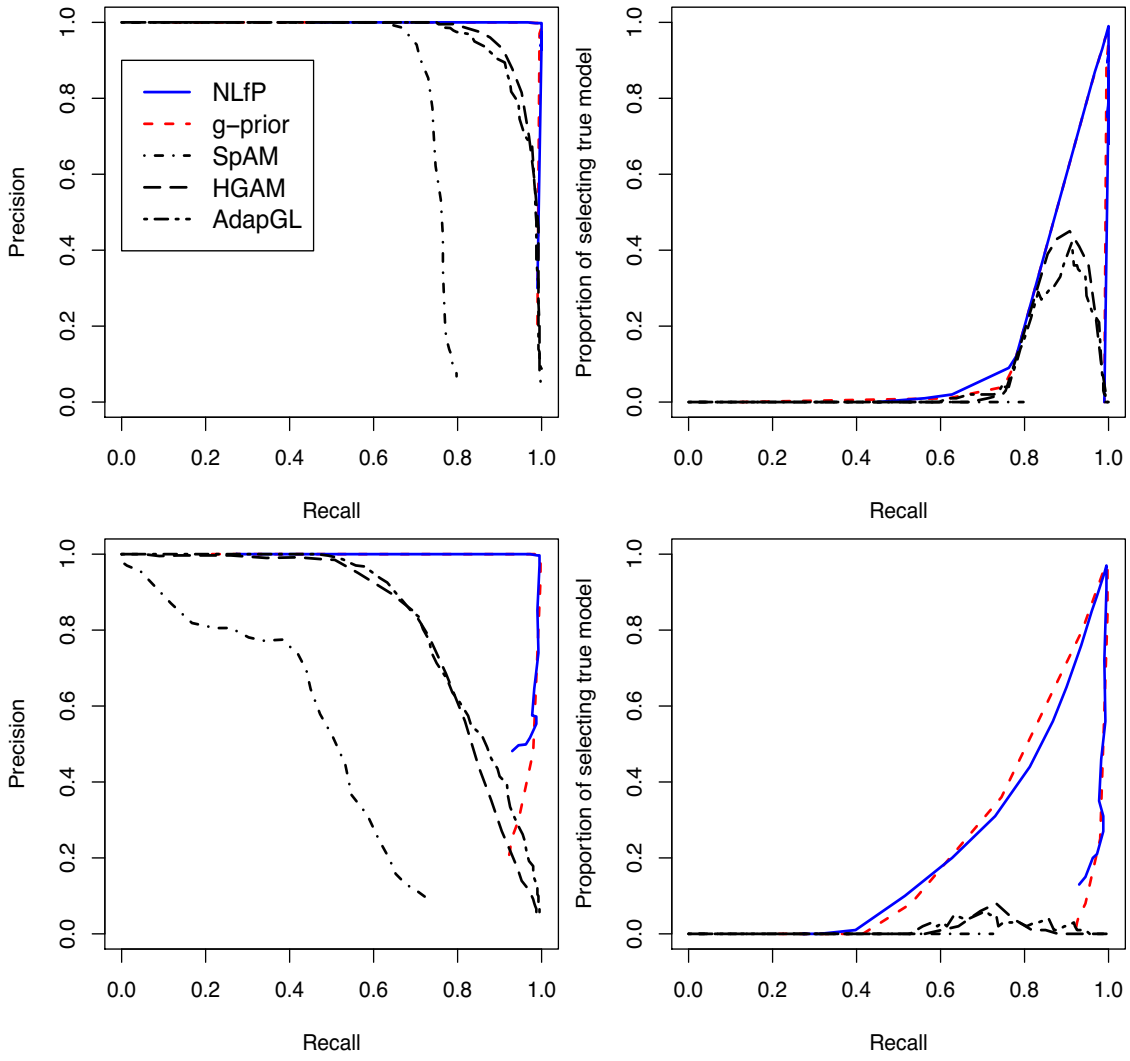


Figure 5.2: The first column illustrates the PR curve of each method; the second column displays proportion of selecting the true model versus recall. The results of *Scenario 1* and *Scenario 2* are represented in the first row and the second row, respectively. "NLfP" denotes the model selection procedure based on the product of inverse moment functional priors, and "g-prior" is the model selection procedure defined by the product of g-priors as in (5.21).

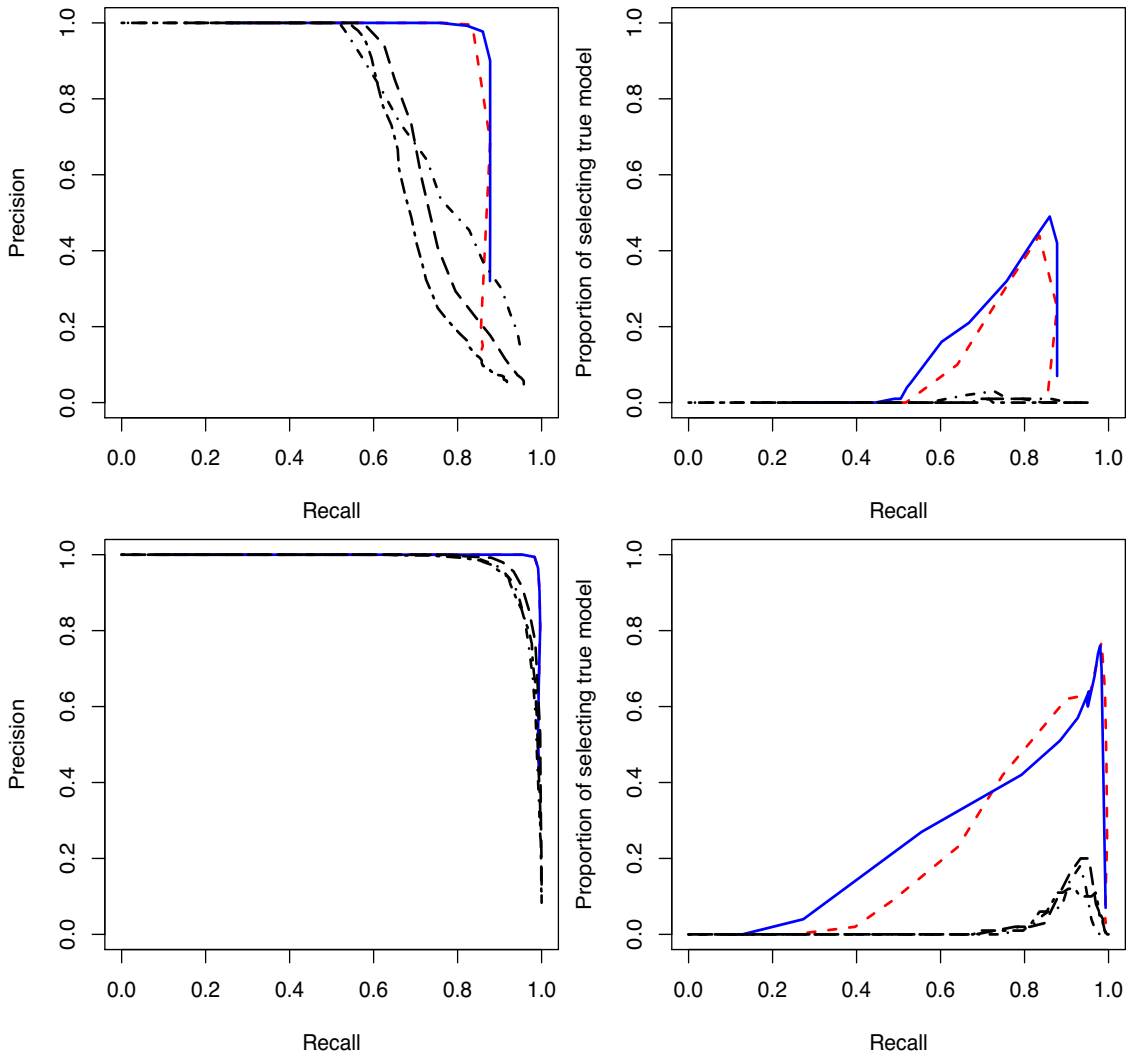


Figure 5.3: The results of *Scenario 3* and *Scenario 4* are represented in the first row and the second row, respectively. The detailed description of this figure is given in the caption of Figure 5.2.

from a large value to a small value. Since precision is 1 - False Discovery Rate (FDR) and recall is 1-Type II error rate, a PR curve illustrates the trade-off between the False Discovery Rate (FDR) and the Type-I error rate. The performance of a procedure can be measured by the area under the PR curve. The greater the area, the more accurate the method in model selection. In Davis and Goadrich (2006), PR curves were proposed as alternatives to the Receiver Operating Characteristic (ROC) curves in high-dimensional settings. Shin et al. (2017) also considered this measure in their simulation studies of model selection for high-dimensional linear regression models. I plotted the proportion times each procedure selected the true model versus recall as its tuning parameter varies. For each method, the curve is drawn by varying different tuning parameters, so the comparison of the model selection procedures is free from the choice of tuning parameters. I report the averaged results over 100 independent replicates for each scenario.

Figure 5.2 and Figure 5.3 demonstrate the PR curves and plots the proportion times the true model is selected versus recall. The performance of the model selection based on the nonlocal functional prior was better than the penalized likelihood estimators in the sense that they achieve a larger area under the PR curve. In addition, they more frequently selected the true model at any recall level. This follows from the the plot of the proportion times the true model selected that showed that the curve of the nonlocal procedure fully covered these curves of the penalized likelihood methods on any recall value. Compared to the Bayesian procedure based on the g-prior, the nonlocal procedure performed similiarly in all scenarios except the third, where the nonlocal prior performed better.

In Table 5.2, I report the Mean Square Error (MSE) and the Mean Square Predictive Error (MSPE) for each procedure. The tuning parameter used in Table 5.2 were chosen to minimize the MSE, i.e.,

$$\widehat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, E_0(\|F_0 - \widetilde{F}_\lambda\|_{n,2}^2),$$

where $\widetilde{F}_\lambda$ is the estimated additive regression function with a tuning parameter λ. For the Bayesian procedures, maximum a posteriori (MAP) estimators were used to evaluate the MSE and the MSPE. These results suggest that the model selection procedure based on the nonlocal functional prior shows promising performance compared to the penalized likelihoods procedures, and its estimation and prediction performances are almost same as those based on the product of g-priors.

5.5.5 Practical Selection of Hyperparameter Values

In the simulation studies, it was not necessary to choose a specific hyperparameter τ_n because PR-curves were free from the choice of the hyperparameter. However, in practice one must choose a value for the hyperparameter.

To choose an appropriate hyperparameter τ_n, I used a procedure in which I compared the null density of the maximum likelihood estimator (MLE) of the discrepancy measure $F^{\mathrm{T}}(I - Q_0)F/\sigma^2$ to the prior density of $\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I - Q_0)\Phi\beta/\sigma^2$ under the alternative hypothesis where $\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I - Q_0)\Phi\beta/\sigma^2 > 0$. Both densities were evaluated from randomly selected covariates. This idea stems from Nikooienejad et al. (2016) who proposed a general idea for selecting a hyperparameters in linear models. I extend their idea to nonparametric settings by using the semi-norm $F^{\mathrm{T}}(I - Q_0)F$. Under the additive regression model in (4.14), I first evaluated $\widehat{\sigma}^2$ by using a few variables chosen by INIS. I then defined the null distribution of $\widehat{F}^{\mathrm{T}}(I - Q_0)\widehat{F}/\widehat{\sigma}^2$, where $\widehat{F}$ denotes the B-spline MLE of F under the null hypothesis where $\mathbf{y} = \epsilon$ for $\epsilon \sim N(0, \widehat{\sigma}^2 I)$, so that the null distribution follows $\chi^2_{K_n-1}$. For a given τ_n and estimated $\widehat{\sigma}^2$, I randomly sampled a $j \in \{1, \ldots, p\}$ and a β_j from the prior density proportional to

$$\exp\left\{-\frac{\beta_j^{\mathrm{T}}\beta_j}{2\widehat{\sigma}^2\tau_n} - \frac{\widehat{\sigma}^2\tau_n}{\beta_j^{\mathrm{T}}\Phi_j^{\mathrm{T}}(I - Q_0)\Phi_j\beta_j}\right\}.$$

I evaluated $\beta_j^{\mathrm{T}}\Phi_j^{\mathrm{T}}(I - Q_0)\Phi_j\beta_j/\widehat{\sigma}^2$ by plugging the sampled β_j in the discrepancy measure, and repeated this procedure many times to approximate the prior density of the discrepancy

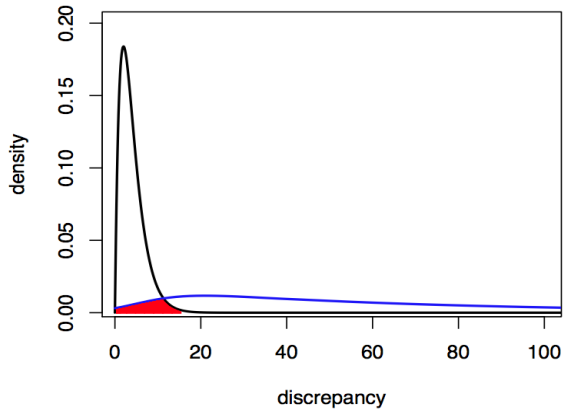


Figure 5.4: The black line and the blue line are the density functions of the null and the prior distribution of $F^{\mathrm{T}}(I - Q_0)F/\widehat{\sigma}^2$, respectively, for a given τ_n and $\widehat{\sigma}^2$

measure.

Figure 5.4 illustrates the resulting null density and the prior density on the discrepancy measure $F^{\mathrm{T}}(I - Q_0)F/\widehat{\sigma}^2$ for some τ_n and $\widehat{\sigma}^2$. I numerically determined the value of the hyperparameter so that the overlap of the null density and the prior density of the discrepancy (red-colored in Figure 5.4) falls below a certain threshold t. For example, I took $t = p^{-1}$. By choosing τ_n to be large enough so that the intersection of these two densities is smaller than the specified threshold, I was able to approximately bound the probability of false positives in the model. As the threshold t decreases to zero, the prior density deviates more from the null distribution. This results in a model selection procedure that strongly penalizes models with covariates that have a smaller discrepancy measure than the intersection point between the null distribution and the prior distribution on the discrepancy.

I applied this procedure to choose the hyperparameter for the nonlocal functional priors in the following the section for real data analyses.

5.6 Applications to Real Data Sets

5.6.1 Bardet-Biedl Syndrome Gene Expression Data

I considered again the Bardet-Biedl syndrome data set that was used in Section 3.4.1. The detailed description of this data set is provided in Section 3.4.1.

5.6.2 Near Infrared Spectroscopy Data

I also evaluated Near Infrared (NIR) Spectroscopy data set. This data set has previously analysed in Liebmann et al. (2009) and Curtis et al. (2014), and is available from the `R` package `chemometrics`. The NIR data includes glucose and ethanol concentration (in g/L) for 166 alcoholic fermentation mashes of different feedstock (rye, wheat and corn). The data set is modeled by 235 NIR spectroscopy absorbance values acquired in the wavelength range of 115-2285 nanometer (nm) by a transflectance probe (Liebmann et al., 2009). I implement my model selection procedure on the data values with response variables defined by ethanol concentrations ($n = 166$ and $p = 235$). I set the training and test set size to be 146 and 20, respectively.

5.6.3 Technical Details and Results

For all data sets, the response variable was centered so that its sample mean was zero, and each covariate was standardized so that its sample mean and standard deviation are zero and one. Each model selection procedure was conducted on the training data set and the performance of the procedures was examined on the test samples, and I repeated this process for 200 replicates. I report the out-of-sample prediction error (PE), which is the sum of square prediction errors divided by the test sample size, and the average of model size (MS) in Table 5.3. I set $K_n = 5$, and assumed $\pi(\sigma^2) \propto 1/\sigma^2$.

Following procedures described in Huang et al. (2008), I choose the tuning parameters for the penalized likelihood approaches using Bayesian Information Criterion (BIC) (Schwarz,

	Bardet-Biedl Data		NIR Data	
Method	PE	MS	PE	MS
NLfP	**1.667** (3.77)	6.00 (1.51)	**1.220** (0.61)	5.80 (0.78)
g-prior($g_n = n$)	93.636 (459.82)	16.97 (0.31)	2.634 (2.01)	9.23 (1.26)
g-prior($g_n = p$)	82.186 (378.09)	16.92 (1.04)	2.302 (1.71)	8.70 (1.24)
g-prior($g_n = p^{5/4}$)	32.829 (200.49)	5.24 (1.12)	1.718 (1.31)	6.86 (1.12)
g-prior($g_n = p^{3/2}$)	3.705 (34.03)	2.05 (0.57)	1.428 (1.08)	5.83 (1.03)
g-prior($g_n = p^2$)	1.804 (3.13)	1.01 (0.07)	1.395 (0.81)	4.14 (0.61)
HGAM(EBIC)	1.882 (1.76)	2.74 (3.66)	1.941 (0.97)	47.14 (3.41)
HGAM(BIC)	1.846 (1.72)	3.38 (4.70)	1.925 (0.96)	47.61 (3.17)
SpAM(EBIC)	2.904 (26.17)	5.86 (2.01)	54.52 (47.91)	9.05 (9.42)
SpAM(BIC)	2.931 (27.35)	5.89 (1.99)	17.101 (21.21)	16.00 (11.02)
AdapGL(EBIC)	2.301 (8.86)	6.80 (8.59)	23.299 (18.57)	5.11 (0.56)
AdapGL(BIC)	16.404 (92.11)	15.73 (7.58)	8.093 (7.19)	7.35 (1.57)

Table 5.3: Real data examples. PE and MS indicate the out-of-sample prediction error and the average of model size; the PE for the Bardet-Biedl Data is scaled by 10^{-2}. The standard deviation of each quantity over 200 replicates is noted in parentheses.

1978) and Extended BIC (EBIC) (Chen and Chen, 2008) for the penalized likelihood methods. These criteria can be expressed as

$$\begin{aligned} \text{BIC}(\lambda) &= \log(\text{RSS}_\lambda) + |\widehat{\mathbf{k}}_\lambda| K_n(\log n)/n \\ \text{EBIC}(\lambda) &= \log(\text{RSS}_\lambda) + |\widehat{\mathbf{k}}_\lambda| K_n(\log n)/n + \nu |\widehat{\mathbf{k}}_\lambda| K_n(\log p)/n, \end{aligned}$$

where $0 \leq \nu \leq 1$ is a constant and λ is a tuning parameter. RSS_λ and $\widehat{\mathbf{k}}_\lambda$ are the residual sum of square and the selected model for a given λ, respectively. As in Huang et al. (2010), I use $\nu = 0.5$ for the EBIC. Also, I have considered multiple hyperparameters ($g_n = \{n, p, p^{5/4}, p^{3/2}, p^2\}$) for the model selection procedure based on g-priors. The hyperparameter τ_n for my approach was chosen by the procedure described in Section 5.5.5. For Bayesian procedures, I used the MAP estimators to summarize PE.

Table 4.2 summarizes the results from these studies. In both data sets, the model selection

procedure based on the nonlocal functional prior led to the smallest values of PE. In particular, my procedure showed better prediction performance compared to the considered penalized likelihood procedures. Even though multiple hyperparameters were considered for the g-priors, these procedures had larger prediction errors than the nonlocal procedure. The g-prior results were very sensitive to the choice of the hyperparameter g_n.

5.7 Conclusion

This dissertation has proposed new classes of nonlocal functional prior densities that have favorable asymptotic properties for nonparametric Bayesian testing problems. I have discussed their advantages over local alternative priors with respect to the convergence rate of Bayes factors. I have focused on B-spline based nonparametric models. However, my methodology can be applied to general classes of nonparametric functional models including Gaussian process regression models. I suggested three natural examples for the usage of the proposed priors. These included sparsity and linearity tests for the nonparametric simple regression models, and a constant function test for the varying coefficient model.

In Section 5.5, I applied one of the proposed priors (the inverse moment functional prior) to additive model selection problems for high-dimensional settings. I showed that the resulting model selection procedure achieved consistency in the sense that the posterior true model probability converged to one in probability under certain regularity conditions. I provided a procedure to select an appropriate hyperparameter τ_n. I also have examined its finite sample performance in model selection using simulated data sets and real data sets. The model selection procedure based on my inverse moment functional priors performed better according to several measures that several alternative procedures.

Finally, I have proposed a scalable computation algorithm that is a modified version of S5. The computational functions for the additive model selection procedure described in this dissertation are available in the `R` package `BayesS5`.

REFERENCES

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–3430.

Bahadur, R. R. and Bickel, P. J. (1967). Asymptotic optimality of Bayes test statistics. *Technical Report. University of Chicago*.

Barber, R. F., Drton, M., and Tan, K. M. (2016). Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data*, pages 15–36. Springer.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897.

Bernardo, J. M. and Smith, A. F. (1994). Bayesian theory. 1994. *John Willey and Sons. Valencia (España)*.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, 36:113–133.

Bondell, H. and Reich, B. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624.

Bontemps, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Annals of Statistics*, 39(5):2557–2584.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *Annals of Statistics*, 17(2):453–510.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

Caron, F. and Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning*, pages 88–95. Association for Computing Machinery.

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

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

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics*, 40(4):2069–2101.

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

Choi, T., Lee, J., and Roy, A. (2009). A note on the Bayes factor in a semiparametric regression model. *Journal of Multivariate Analysis*, 100(6):1316–1327.

Choi, T. and Rousseau, J. (2015). A note on Bayes factor consistency in partial linear models. *Journal of Statistical Planning and Inference*, 166:158–170.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544.

Curtis, S. M., Banerjee, S., and Ghosal, S. (2014). Fast Bayesian model assessment for non-parametric additive regression. *Computational Statistics & Data Analysis*, 71:347–358.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors: an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Royal Statistical Society: Series B*, 70(5):849–911.

George, E. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

Ghosal, S., Lember, J., and van der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, pages 711–732.

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

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

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for large p regression. *Journal of the American Statistical Association*, 102(478):507–516.

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. (2012). *Nonparametric and semiparametric models*. Springer Science & Business Media.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Royal Statistical Society: Series B*, 55(4):757–796.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38(4):2282–2313.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, 1(1961):361–379.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford.

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

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

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673.

Kim, Y., Kwon, S., and Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13:1037–1057.

Kirkpatrick, S. and Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation. *Computational Statistics and Data Analysis*, 12(3):327–347.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., and Manly, K. F. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2(1):51–61.

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

Liang, F., Song, Q., and Yu, K. (2013). Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association*, 108(502):589–606.

Liebmann, B., Friedl, A., and Varmuza, K. (2009). Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Analytica Chimica Acta*, 642(1):171–178.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297.

Liu, Q. and Ihler, A. (2013). Variational algorithms for marginal MAP. *Journal of Machine Learning Research*, 14(1):3165–3200.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Meier, L., Van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, 37(6B):3779–3821.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics*, 42(2):789–817.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–767.

Nikooienejad, A., Wang, W., and Johnson, V. E. (2016). Bayesian variable selection for binary outcomes in high dimensional genomic studies using non-local priors. *Bioinformatics*, 32:1338–1345.

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

Polson, N. and Scott, J. (2010a). Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*, volume 9, pages 501–538. Oxford University Press.

Polson, N. and Scott, J. (2010b). Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*, volume 9, pages 501–538. Oxford University Press.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Polson, N. G., Scott, J. G., and Windle, J. (2014). The Bayesian bridge. *Royal Statistical Society: Series B*, 76(4):713–733.

Radchenko, P. and James, G. M. (2011). Improved variable selection with forward-lasso adaptive shrinkage. *Annals of Applied Statistics*, 5(1):427–448.

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

Raskutti, G., Wainwright, M. J., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13:389–427.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Royal Statistical Society: Series B*, 71(5):1009–1030.

Rockova, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Rossell, D. and Telesca, D. (2017+). Non-local priors for high-dimensional estimation. *Journal of the American Statistical Association*, (to appear).

Rossell, D., Telesca, D., and Johnson, V. E. (2013). High-dimensional Bayesian classifiers using non-local priors. In *Statistical Models for Data Analysis*, pages 305–313. Springer.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., and Casavant, T. L. (2006). 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.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Scott, J. and Berger, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619.

Scott, J. G. and Walker, S. G. (2015). Nonparametric Bayesian testing for monotonicity. *Biometrika*, 102(3):617–630.

Shang, Z. and Li, P. (2014). High-dimensional Bayesian inference in nonparametric additive models. *Electronic Journal of Statistics*, 8(2):2804–2847.

Shin, M., Bhattacharya, A., and Johnson, E. J. (2017+). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, (to appear).

Shin, M. and Tian, R. (2017). *BayesS5: Bayesian Variable Selection Using Simplified Shotgun Stochastic Search with Screening (S5)*. R package version 1.30.

Song, Q. and Liang, F. (2015). High-dimensional variable selection with reciprocal l 1-regularization. *Journal of the American Statistical Association*, 110(512):1607–1620.

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

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244.

Wahba, G. (1990). Spline models for observational data. In *CBMS-NSF regional conference series in applied mathematics (59)*. Philadelphia: Society for Industrial and Applied Mathematics.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Royal Statistical Society: Series B*, 31(1):80–88.

Walker, S. G. (2004). Modern Bayesian asymptotics. *Statistical Science*, 19(1):111–117.

Xue, L. (2009). Consistent variable selection in additive models. *Statistica Sinica*, 19:1281–1296.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional bayesian variable selection. *Annals of Statistics*, 44(6):2497–2532.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Royal Statistical Society: Series B*, 68(1):49–67.

Yuan, M. and Zhou, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *Annals of Statistics*, 44(6):2564–2593.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North Holland, Amsterdam.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942.

Zhang, D., Lin, Y., and Zhang, M. (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3:781–796.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, 106(495):1099–1112.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics*, 26(5):1760–1782.

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

APPENDIX A

PROOFS OF THEORETICAL RESULTS

A.1 Nonlocal Prior Densities for High-dimensional Linear Model Selection

Preliminary Results

Lemma A.1.1. *For $Q_{\mathbf{k}}$ defined in* (2.6)*, $\prod_{j=1}^{k} Q_{\mathbf{k},j}^{L} \leq Q_{\mathbf{k}} \leq \prod_{j=1}^{k} Q_{\mathbf{k},j}^{U}$,*

where

$$\begin{aligned} Q_{\mathbf{k},j}^{L} &= c_1(\sigma^2)^{1/2}(n\nu_{\mathbf{k}}^{*} + 1/\tau_{n,p})^{-1/2}\exp\{-\tau_{n,p}/\widetilde{\beta}_{\mathbf{k},j}^{*2}\}, \\ Q_{\mathbf{k},j}^{U} &= c_2(\sigma^2)^{1/2}(n\nu_{\mathbf{k}*} + 1/\tau_{n,p})^{-1/2}\exp\{-\tau_{n,p}/(|\widetilde{\beta}_{\mathbf{k},j}| + \widetilde{\epsilon}_n)^2\}, \end{aligned}$$

and $\widetilde{\epsilon}_n \asymp (n\nu_{\mathbf{k}*}/\tau_{n,p})^{-1/4}$, with $\widetilde{\beta}_{\mathbf{k},j}^{*} \in [\widetilde{\beta}_{\mathbf{k},j} - \widetilde{\epsilon}_n, \widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n] \setminus (-\widetilde{\epsilon}_n, \widetilde{\epsilon}_n)^c$ for some positive constants c_1 and c_2.

Proof. Recall $\widetilde{\Sigma}_{\mathbf{k}} = (X_{\mathbf{k}}^T X_{\mathbf{k}} + 1/\tau_{n,p}\mathrm{I}_{\mathbf{k}})^{-1}$. From (2.8), all eigenvalues of $(\widetilde{\Sigma}_{\mathbf{k}})^{-1}$ are bounded between $n\nu_{\mathbf{k}*} + 1/\tau_{n,p}$ and $n\nu_{\mathbf{k}}^{*} + 1/\tau_{n,p}$, which implies for all $x \in \mathbb{R}^{|\mathbf{k}|}$, $(n\nu_{\mathbf{k}*} + 1/\tau_{n,p})x^T x \leq x^T(\widetilde{\Sigma}_{\mathbf{k}})^{-1}x \leq (n\nu_{\mathbf{k}}^{*} + 1/\tau_{n,p})x^T x$. Let $T_{1n} = \{(n\nu_{\mathbf{k}}^{*} + 1/\tau_{n,p})/\sigma^2\}^{1/2}$ and $T_{2n} = \{(n\nu_{\mathbf{k}*} + 1/\tau_{n,p})/\sigma^2\}^{1/2}$. Substituting the above inequality in the expression for $Q_{\mathbf{k}}$, we have

$$\prod_{j=1}^{|\mathbf{k}|} g_1(\widetilde{\beta}_{\mathbf{k},j}) \leq Q_{\mathbf{k}} \leq \prod_{j=1}^{|\mathbf{k}|} g_2(\widetilde{\beta}_{\mathbf{k},j}), \tag{A.1}$$

where

$$g_i(\widetilde{\beta}_{\mathbf{k},j}) = \int_{-\infty}^{\infty} \exp\{-T_{in}^2(\beta_{\mathbf{k},j} - \widetilde{\beta}_{\mathbf{k},j})^2/2 - \tau_{n,p}/\beta_{\mathbf{k},j}^2\}d\beta_{\mathbf{k},j}, \tag{A.2}$$

for $i = 1, 2$. We establish the lower bound first by showing that $g_1(\widetilde{\beta}_{\mathbf{k},j}) \geq Q_{\mathbf{k},j}^{L}$ for all $j =$

$1, \ldots, |\mathbf{k}|$. Recall $\widetilde{\epsilon}_n \asymp (n\nu_{\mathbf{k}*}/\tau_{n,p})^{-1/4}$ from the statement of the Lemma. We have

$$\begin{aligned} g_1(\widetilde{\beta}_{\mathbf{k},j}) &\geq \int_{[\widetilde{\beta}_{\mathbf{k},j}-\widetilde{\epsilon}_n, \widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n]\setminus(-\widetilde{\epsilon}_n,\widetilde{\epsilon}_n)^c} \exp\{-T_{1n}^2(\beta_{\mathbf{k},j}-\widetilde{\beta}_{\mathbf{k},j})^2/2 - \tau_{n,p}/\beta_{\mathbf{k},j}^2\} d\beta_{\mathbf{k},j} \\ &\geq \exp\{-\tau_{n,p}/\widetilde{\beta}_{\mathbf{k},j}^{*2}\} \int_{[\widetilde{\beta}_{\mathbf{k},j}-\widetilde{\epsilon}_n, \widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n]\setminus(-\widetilde{\epsilon}_n,\widetilde{\epsilon}_n)^c} \exp\{-T_{1n}^2(\beta_{\mathbf{k},j}-\widetilde{\beta}_{\mathbf{k},j})^2/2\} d\beta_{\mathbf{k},j}, \end{aligned}$$

for some $\widetilde{\beta}_{\mathbf{k},j}^* \in [\widetilde{\beta}_{\mathbf{k},j}-\widetilde{\epsilon}_n, \widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n] \setminus (-\widetilde{\epsilon}_n, \widetilde{\epsilon}_n)^c$. Then, the integral in the last line of the above display is equivalent to

$$\int_{[-\widetilde{\epsilon}_n,\widetilde{\epsilon}_n]\setminus(-\widetilde{\beta}_{\mathbf{k},j}-\widetilde{\epsilon}_n, -\widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n)^c} e^{-T_{1n}^2 t^2/2} dt \geq c_1 T_{1n}^{-1} \int_0^{T_{1n}\widetilde{\epsilon}_n} e^{-z^2/2} dz \geq c_2 T_{1n}^{-1},$$

where c_1 and c_2 are some positive constants and the last inequality in the above display follows since $T_{1n}\widetilde{\epsilon}_n \geq 1$ for large n. Substituting back in the previous display, $g_1(\widetilde{\beta}_{\mathbf{k},\mathbf{j}}) \geq c_1 T_{1n}^{-1} \exp\{-\tau_{n,p}/\widetilde{\beta}_{\mathbf{k},\mathbf{j}}^{*2}\}$ for some constant $c_1 > 0$, completing the proof of the lower bound.

We now establish the upper bound by showing that $g_2(\widetilde{\beta}_{\mathbf{k},j}) \leq Q_{\mathbf{k},j}^U$ for all $j = 1, \ldots, |\mathbf{k}|$. It is straightforward to see that g_2 is a symmetric function (i.e, $g_2(\widetilde{\beta}_{\mathbf{k},j}) = g_2(|\widetilde{\beta}_{\mathbf{k},j}|)$), so that it is enough to establish the bound for $\widetilde{\beta}_{\mathbf{k},j} > 0$; without loss of generality we assume that $\widetilde{\beta}_{\mathbf{k},j} > 0$. We have

$$\begin{aligned} & \int_{-\infty}^{\infty} \exp\{-T_{2n}^2(\beta_{\mathbf{k},j}-\widetilde{\beta}_{\mathbf{k},j})^2/2 - \tau_{n,p}\beta_{\mathbf{k},j}^2\} d\beta_{\mathbf{k},j} \\ = & \int_{-\infty}^{0} \exp\{-T_{2n}^2(\beta_{\mathbf{k},j}-\widetilde{\beta}_{\mathbf{k},j})^2/2 - \tau_{n,p}/\beta_{\mathbf{k},j}^2\} d\beta_{\mathbf{k},j} \\ + & \int_{0}^{\widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n} \exp\{-T_{2n}^2(\beta_{\mathbf{k},j}-\widetilde{\beta}_{\mathbf{k},j})^2/2 - \tau_{n,p}/\beta_{\mathbf{k},j}^2\} d\beta_{\mathbf{k},j} \\ + & \int_{\widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n}^{\infty} \exp\{-T_{2n}^2(\beta_{\mathbf{k},j}-\widetilde{\beta}_{\mathbf{k},j})^2/2 - \tau_{n,p}/\beta_{\mathbf{k},j}^2\} d\beta_{\mathbf{k},j}. \end{aligned}$$

Define the first term of the above as W_1, the second as W_2, and the third term as W_3. First, we shall show that $W_1 \leq cT_{2n}^{-1} \exp\{-T_{2n}(2\tau_{n,p})^{1/2}\}$ for some positive constant c. By transforming

the variable $t = \beta_{\mathbf{k},j} - \widetilde{\beta}_{\mathbf{k},j}$,

$$\begin{aligned} W_1 &= \int_{-\infty}^{0} \exp\{-T_{2n}^2 t^2/2 + T_{2n}^2 t\widetilde{\beta}_{\mathbf{k},j} - T_{2n}^2 \widetilde{\beta}_{\mathbf{k},j}^2/2 - \tau_{n,p}/t^2\}dt \\ &\leq \int_{-\infty}^{0} \exp\{-T_{2n}^2 t^2/2 - \tau_{n,p}/t^2\}dt \\ &\leq c_3 T_{2n}^{-1} \exp\{-T_{2n}(2\tau_{n,p})^{1/2}\}, \end{aligned}$$

for some constant c_3, since $\int \exp\{-\mu/t^2 - \zeta t^2\}dt = (\pi/\zeta)^{-1/2}\exp\{-2(\mu\zeta)^{1/2}\}$ for $\mu > 0$ and $\zeta > 0$.

Second, by changing the variable $z = t - \widetilde{\epsilon}$,

$$\begin{aligned} W_2 &= \int_{-\widetilde{\epsilon}_n}^{\widetilde{\beta}_{\mathbf{k},j}} \exp\{-T_{2n}^2(z - \widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2/2 - \tau_{n,p}/(z + \widetilde{\epsilon}_n)^2\}dz \\ &\leq \exp\{-\tau_{n,p}/(\widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2\} \int_{-\infty}^{\infty} \exp\{-T_{2n}^2(z - \widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2/2\} \\ &\leq c_4 T_{2n}^{-1} \exp\{-\tau_{n,p}/(\widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2\}, \end{aligned}$$

for some positive constant c_4.

Third, by changing the variable $z = t - \widetilde{\beta}_{\mathbf{k},j}$, there exists some positive constant c such that

$$\begin{aligned} W_3 &= \int_{\widetilde{\epsilon}_n}^{\infty} \exp\{-T_{2n}^2 z^2/2 - \tau_{n,p}/(z + \widetilde{\beta}_{\mathbf{k},j})^2\}dz \\ &\leq \exp\{-T_{2n}^2 \widetilde{\epsilon}_n^2/4\} \int_{-\infty}^{\infty} \exp\{-T_{2n}^2 z^2/4\}dz \\ &\leq c_5 T_{2n}^{-1} \exp\{-c_6 T_{2n} \tau_{n,p}^{1/2}\}, \end{aligned}$$

for some constants $c_5$and c_6, since $\widetilde{\epsilon}_n \asymp (n\nu_{\mathbf{k}*}/\tau_{n,p})^{-1/4}$. Then,

$$\begin{aligned} g_2(\widetilde{\beta}_{\mathbf{k},j}) \leq c_3 T_{2n}^{-1} \exp\{-T_{2n}(2\tau_{n,p})^{1/2}\} + c_4 T_{2n}^{-1} \exp\{-\tau_{n,p}/(\widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2\} \\ + c_5 T_{2n}^{-1} \exp\{-c_6 T_{2n} \tau_{n,p}^{1/2}\}. \end{aligned}$$

Since $\widetilde{\epsilon}_n \asymp (n\nu_{\mathbf{k}*}/\tau_{n,p})^{-1/4}$, when $\widetilde{\beta}_{\mathbf{k},j} < \widetilde{\epsilon}_n$, $\tau_{n,p}/(\widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2 < \tau_{n,p}/(4\widetilde{\epsilon}_n^2) \asymp T_{2n}\tau_{n,p}^{1/2}$, and when $\widetilde{\beta}_{\mathbf{k},j} \geq \widetilde{\epsilon}_n$, $\tau_{n,p}/(\widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2 \leq \tau_{n,p}/(4\widetilde{\beta}_{\mathbf{k},j}^2) < T_{2n}\tau_{n,p}^{1/2}$. In overall, the right-hand side of the above display would be dominated by the second term, which shows that $g_2(\widetilde{\beta}_{\mathbf{k},j}) \leq cT_{2n}^{-1}\exp\{-\tau_{n,p}/(\widetilde{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n)^2\}$ for some constant c. When $\widetilde{\beta}_{\mathbf{k},j} < 0$, we can show the same result by following exactly the same steps explained above. □

We now present some auxiliary results that are used to prove Theorems 1 and 2. We make use of the following simple union bound multiple times: for non-negative random variables $V_1, \ldots, V_m$ and $a > 0$,

$$P(\sum_{l=1}^{m} V_l > a) \leq \sum_{l=1}^{m} P(V_l > a/m) \leq m \max_{1\leq l\leq m} P(V_l > a/m). \tag{A.3}$$

We define some notations that are used in the subsequent proofs. Let $\mathbf{t}$ denote the true data generating model, and let $\beta_{\mathbf{t}}^0$ denote the true regression coefficient corresponding to $\mathbf{t}$. Let $\mathbf{c_t} = \mathbf{t} \setminus \mathbf{k}$, $\mathbf{c_k} = \mathbf{k} \setminus \mathbf{t}$, and $\mathbf{u} = \mathbf{k} \cup \mathbf{t}$. Also, we define the cardinality of a model $\mathbf{k}$ as k and in the same spirit, denote $c_k = |\mathbf{c_k}|$, $c_t = |\mathbf{c_t}|$, and $t = |\mathbf{t}|$. $\{x\}_j$ denotes the *j*-th element of the vector x, and $diag\{A\}_j$ refers to the *j*-th diagonal element in the square matrix A. We denote $\chi_m^2(\lambda)$ a non-central chi-square distribution with the degrees of freedom m and non-centrality parameter λ; a central chi-square distribution is simply denoted by χ_m^2.

An important property that is used in the subsequent proofs concerns the distribution of the marginal ridge estimator. Let $\widetilde{\beta}_{\mathbf{k}} = (X_{\mathbf{k}}^{\mathrm{T}}X + 1/\tau_{n,p}I_{\mathbf{k}})^{-1}X_{\mathbf{k}}^{\mathrm{T}}y$ and $\widetilde{\beta}_{\mathbf{k},j} = \{\widetilde{\beta}_{\mathbf{k}}\}_j$. Then,

$$\widetilde{\beta}_{\mathbf{k},j} \sim N(\beta_{\mathbf{k},j}^*, \sigma_{\mathbf{k},j}^{2*}), \tag{A.4}$$

where $\beta_{\mathbf{k},j}^* = \{(X_{\mathbf{k}}^{\mathrm{T}}X + 1/\tau_{n,p}I_{\mathbf{k}})^{-1}X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{t}}\beta_{\mathbf{t}}^*\}_j$ and $\sigma_{\mathbf{k},j}^{2*} = \sigma^2 diag\{(X_{\mathbf{k}}^{\mathrm{T}}X_{\mathbf{k}} + 1/\tau_{n,p}I_{\mathbf{k}})^{-1}\}_j$. It is also evident that $(\widetilde{\beta}_{\mathbf{k},j} - \beta_{\mathbf{k},j}^*)^2/\sigma_{\mathbf{k},j}^{2*} \sim \chi_1^2$.

A set of technical results follow that are used in the proof of the main results. Define

$$H_{1n} = \sum_{\substack{\mathbf{k}:\mathbf{t}\subsetneq\mathbf{k},\\ |\mathbf{k}|\le q_n}} \frac{m_{\mathbf{k}}(y)\pi(\mathbf{k})}{m_{\mathbf{t}}(y)\pi(\mathbf{t})} = \sum_{\substack{\mathbf{k}:\mathbf{t}\subsetneq\mathbf{k},\\ |\mathbf{k}|\le q_n}} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)}, \quad H_{2n} = \sum_{\substack{\mathbf{k}:\mathbf{t}\not\subseteq\mathbf{k},\\ |\mathbf{k}|\le q_n}} \frac{m_{\mathbf{k}}(y)\pi(\mathbf{k})}{m_{\mathbf{t}}(y)\pi(\mathbf{t})} = \sum_{\substack{\mathbf{k}:\mathbf{t}\not\subseteq\mathbf{k},\\ |\mathbf{k}|\le q_n}} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)}. \tag{A.5}$$

Lemma A.1.2. *Fix $\epsilon > 0$. Let $\Gamma_d = \{\mathbf{k} : |\mathbf{k}| \le q_n, \mathbf{t} \subsetneq \mathbf{k}, |\mathbf{k}| - |\mathbf{t}| = d\}$ for $d = 1, \ldots, q_n - |\mathbf{t}|$. Suppose there exist constants $c, \delta > 0$ such that $\max_{\mathbf{k}\in\Gamma_d} P\{\pi(\mathbf{k} \mid y)/\pi(\mathbf{t} \mid y) > \epsilon p^{-d}/q_n\} \le cp^{-d(1+\delta)}$ for $d = 1, \ldots, q_n - |\mathbf{t}|$. Then, H_{1n} converges to zero in probability as n tends to ∞, where H_{1n} is as in* (A.5).

Proof. Clearly, $|\Gamma_d| = \binom{p-|\mathbf{t}|}{d}$. Using (A.3), we bound

$$\begin{aligned} P\Big\{\sum_{\mathbf{k}:\mathbf{t}\subsetneq\mathbf{k}} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon\Big\} &= P\Big\{\sum_{d=1}^{q_n-|\mathbf{t}|}\sum_{\mathbf{k}\in\Gamma_d} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon\Big\} \\ &\le \sum_{d=1}^{q_n-|\mathbf{t}|} P\Big\{\sum_{\mathbf{k}\in\Gamma_d} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon/q_n\Big\} \\ &\le \sum_{d=1}^{q_n-|\mathbf{t}|} \binom{p-|\mathbf{t}|}{d} \max_{\mathbf{k}\in\Gamma_d} P\Big\{\frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon p^{-d}/q_n\Big\} \le \sum_{d=1}^{q_n-|\mathbf{t}|} cp^{-d\delta}. \end{aligned}$$

Finally, $\sum_{d=1}^{q_n-|\mathbf{t}|} cp^{-d\delta} \le cq_np^{-\delta} \to 0$ as $n \to \infty$. □

Lemma A.1.3. *Fix $\epsilon > 0$ and let $t = |\mathbf{t}|$. Define $\Gamma_{k,c_k,c_t} = \{\mathbf{k} : |\mathbf{k}| \le q_n, |\mathbf{k}| = k,\ |\mathbf{k}\backslash\mathbf{t}| = c_k,\ |\mathbf{t}\backslash\mathbf{k}| = c_t\}$ for $k = 0, \ldots, q_n;\ c_k = 0, \ldots, k;\ c_t = 1, \ldots, t$. Suppose*

$$\max_{\mathbf{k}\in\Gamma_{k,c_k,c_t}} P\Big[\frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon n^{-3}p^{-k}n^{-c_k}t^{-t}\Big] \le cp^{-k(1+\delta)},$$

with some postive constants c and δ. Then, H_{2n} converges to zero as n tends to ∞, where H_{2n} is as in (A.5).

Proof. Clearly, $|\Gamma_{k,c_k,c_t}| = \binom{p}{k}\binom{k}{c_k}\binom{t}{c_t}$.

$$\begin{aligned}
P\Big\{\sum_{\mathbf{k}:\mathbf{t}\not\subseteq\mathbf{k}} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon\Big\} &\leq P\Big\{\sum_{k=1}^{q_n}\sum_{c_k=0}^{k}\sum_{c_t=1}^{t}\sum_{\mathbf{k}\in\Gamma_{k,c_k,c_t}} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon\Big\} \\
&\leq P\Big\{\sum_{k=1}^{q_n}\sum_{c_k=0}^{k}\sum_{c_t=1}^{t}\sum_{\mathbf{k}\in\Gamma_{k,c_k,c_t}} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon\Big\} \\
&\leq \sum_{k=1}^{q_n}\sum_{c_k=0}^{k}\sum_{c_t=1}^{t} P\Big\{\sum_{\mathbf{k}\in\Gamma_{k,c_k,c_t}} \frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon n^{-3}\Big\} \\
&\leq \sum_{k=1}^{q_n}\sum_{c_k=0}^{k}\sum_{c_t=1}^{t} p^k n^{c_k} t^t \max_{\mathbf{k}\in\Gamma_{k,c_k,c_t}} P\Big\{\frac{\pi(\mathbf{k}\mid y)}{\pi(\mathbf{t}\mid y)} > \epsilon n^{-3} p^{-k} n^{-c_k} t^{-t}\Big\} \\
&\leq \sum_{k=1}^{q_n}\sum_{c_k=0}^{k}\sum_{c_t=1}^{t} p^k n^{c_k} t^t p^{-k(1+\delta)} \to 0,
\end{aligned}$$

as $n \to \infty$. □

Lemma A.1.4. *Suppose W follows a non-central chi-square distribution with the degree of freedom m_n that is a positive integer and the non-central parameter $\lambda_n \geq 0$, i.e, $W \sim \chi^2_{m_n}(\lambda_n)$. Also, consider w_n and t_n such that $w_n \to 0$ and $t_n \to \infty$ as n tends to ∞. Also, assume that $m_n \prec t_n$. Then,*

$$P(W \leq \lambda_n w_n) \leq c_1 \lambda_n^{-1} \exp\{-\lambda_n(1-w_n)^2\}, \tag{A.6}$$

And

$$P(W > \lambda_n + t_n) \leq c_2 \left(\frac{t_n}{2m_n}\right)^{m_n/2} \exp\{m_n/2 - t_n/2\} + c_3 \lambda_n^{1/2} t_n^{-1} \exp\left\{-\frac{t_n^2}{32\lambda_n}\right\}, \tag{A.7}$$

where c_1, c_2, and c_3 are some positive constants.

Proof. W can be expressed as $W = \sum_{i=1}^{m_n}\{Z_i + (\lambda_n/m_n)^{1/2}\}^2$, where $Z_i \overset{i.i.d}{\sim} N(0,1)$ for $i = 1, \ldots, m$. Then, by the fact that $P(Z > a) \leq (2\pi)^{-1/2} a^{-1} \exp\{-a^2/2\}$ for any $a > 0$, we

can show that there exist some positive constants c_1 such that

$$\begin{aligned} P\left(W \leq \lambda_n w_n\right) &= P\big\{ \sum_{i=1}^{m_n} Z_i^2 + 2(\lambda_n/m_n)^{1/2} \sum_{i=1}^{m_n} Z_i + \lambda_n \leq \lambda_n w_n \big\} \\ &\leq P\big\{ m_n^{-1/2} \sum_{i=1}^{m_n} Z_i \leq -\lambda_n^{1/2}(1-w_n)/2 \big\} \\ &= P\big\{ |Z_1| \geq \lambda_n^{1/2}(1-w_n)/2 \big\}/2 \\ &\leq c_1 \lambda_n^{-1} \exp\{-\lambda_n (1-w_n)^2/2\}, \end{aligned}$$

since Z_1 follows a standard normal distribution.

Also, by using Chernoffs's bound and the fact that $P(Z > a) \leq (2\pi)^{-1/2} a^{-1} \exp\{-a^2/2\}$ for any $a > 0$, one can show that

$$\begin{aligned} & P(W > \lambda_n + t_n) = P\left\{ \sum_{i=1}^{m_n} Z_i^2 + 2(\lambda_n/m_n)^{1/2} \sum_{i=1}^{m_n} Z_i > t_n \right\} \\ \leq\ & P\left(\sum_{i=1}^{m_n} Z_i^2 > t_n/2 \right) + P\left\{ m_n^{-1/2} \sum_{i=1}^{m_n} Z_i > \lambda_n^{-1/2} t_n/4 \right\} \\ \leq\ & c_2 \left(\frac{t_n}{2m_n} \right)^{m_n/2} \exp\left\{ m_n/2 - t_n/2 \right\} + c_3 \lambda_n^{1/2} t_n^{-1} \exp\left\{ -\frac{t_n^2}{32\lambda_n} \right\}, \end{aligned}$$

where c_2 and c_3 are some positive constants. □

Lemma A.1.5. *Consider $Q_{\mathbf{k}}$ defined in* (2.6) *for an arbitrary model* $\mathbf{k}$. *Fix any $\delta > 0$. For any* $\mathbf{k}$ *with* $\mathbf{t} \subsetneq \mathbf{k}$,

$$P\left[Q_{\mathbf{k}}/Q_{\mathbf{t}} > \exp\left\{ -|\mathbf{k} \setminus \mathbf{t}| \tau_{n,p}^{2/3} (n\nu_{\mathbf{k}*})^{1/3} + |\mathbf{t}| \tau_{n,p}^{1-\delta/8} (n\nu_{\mathbf{k}*})^{\delta/8} \right\} \right] \leq p^{-|\mathbf{k}\setminus\mathbf{t}|(1+\delta)}, \tag{A.8}$$

and for $\mathbf{k}$ *such that* $\mathbf{t} \nsubseteq \mathbf{k}$,

$$P\left[Q_{\mathbf{k}}/Q_{\mathbf{t}} > \exp\left\{ \left\| \beta_{\mathbf{t}}^0 \right\|_2^2 n\nu_{\mathbf{u}*} / \{2\log(\tau_{n,p}/\log p)\} \right\} \right] \leq p^{-|\mathbf{k}|(1+\delta)}. \tag{A.9}$$

Proof. By Lemma A.1.1, it is sufficient to show that

$$\begin{aligned} & P\left[\prod_{j\in\mathbf{t}}(Q^U_{\mathbf{k},j}/Q^L_{\mathbf{t},j}) > \exp\{|\mathbf{t}|\tau_{n,p}^{1-\delta/8}(n\nu_{\mathbf{k}*})^{\delta/8}\}\right] \\ & +P\left[\prod_{j\in\mathbf{k}\setminus\mathbf{t}} Q^U_{\mathbf{k},j} > \exp\{-|\mathbf{k}\setminus\mathbf{t}|\tau_{n,p}^{2/3}(n\nu_{\mathbf{k}*})^{1/3}\}\right] \\ \leq\ & p^{-|\mathbf{k}\setminus\mathbf{t}|(1+\delta)}. \end{aligned} \tag{A.10}$$

We first shall show that the first term in the left-hand side of (A.10) is bounded above by $\exp\{-cn\nu_{\mathbf{k}*}\}$ for some constant c.

$$\begin{aligned} & P\left[\prod_{j\in\mathbf{t}}\frac{Q^U_{\mathbf{k},j}}{Q^L_{\mathbf{t},j}} > \exp\left\{|\mathbf{t}|\tau_{n,p}^{1-\delta/8}(n\nu_{\mathbf{k}*})^{\delta/8}\right\}\right] \leq \sum_{j\in\mathbf{t}} P\left[\frac{Q^U_{\mathbf{k},j}}{Q^L_{\mathbf{t},j}} > \exp\left\{\tau_{n,p}^{1-\delta/8}(n\nu_{\mathbf{k}*})^{\delta/8}\right\}\right] \\ =\ & \sum_{j\in\mathbf{t}} P\Bigg[c'\left(\frac{n\nu_{\mathbf{k}*}+1/\tau_{n,p}}{n\nu^*_{\mathbf{t}}+1/\tau_{n,p}}\right)^{-1/2}\exp\left\{-\tau_{n,p}\left(1/(|\widetilde{\beta}_{\mathbf{k},j}|+\widetilde{\epsilon}_n)^2-1/\widetilde{\beta}^{*2}_{\mathbf{k},j}\right)\right\} \\ & > \exp\left\{\tau_{n,p}^{1-\delta/8}(n\nu_{\mathbf{k}*})^{\delta/8}\right\}\Bigg] \\ \leq\ & \sum_{j\in\mathbf{t}} P[|\widetilde{\beta}_{\mathbf{k},j}-\beta^*_{\mathbf{k},j}|>\epsilon'] + \sum_{j\in\mathbf{t}} P[|\widetilde{\beta}_{\mathbf{t},j}-\beta^*_{\mathbf{t},j}|>\epsilon'], \end{aligned} \tag{A.11}$$

for some small enough $\epsilon' > 0$ and some positive constant c' and $\widetilde{\beta}^*_{\mathbf{k},j} \in [\widetilde{\beta}_{\mathbf{k},j}-\widetilde{\epsilon}_n, \widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n] \setminus (-\widetilde{\epsilon}_n, \widetilde{\epsilon}_n)^c$ as defined in Lemma A.1.1, and $\widetilde{\beta}_{\mathbf{k},j}$ and $\beta^*_{\mathbf{k},j}$ defined in (A.4). The last inequality in the above display asymptotically holds, since

$$\tau_{n,p}^{1-\delta/8}(n\nu_{\mathbf{k}*})^{\delta/8} \succ \tau_{n,p}/(|\beta^*_{\mathbf{k},j}|-\epsilon'-\widetilde{\epsilon}_n)^2,$$

for any $\delta > 0$.

Since $(\widetilde{\beta}_{\mathbf{k},j}-\beta^*_{\mathbf{k},j})^2/\sigma^{*2}_{\mathbf{k},j} \sim \chi^2_1$ and $\sigma^{*2}_{\mathbf{k},j} \geq (n\nu_{\mathbf{k}*}+1/\tau_{n,p})^{-1}$, by using Lemma A.1.4, one can show that the first term in (A.11) bounded above by $\exp\{-c_1\epsilon'^2 n\nu_{\mathbf{k}*}\}$ for some constant c_1. Similarly, the second term in (A.11) is bounded above by $\exp\{-c_2\epsilon'^2 n\}$ for some constant

c_2, since *Assumption 5* states that $X_{\mathbf{t}}^{\mathrm{T}} X_{\mathbf{t}}/n$ is asymptically isotropic. Therefore, (A.11) is asymptotically bounded by $p^{-q_n(1+\delta)}$ by *Assumption 3*.

Next, we shall show that the second term in the left-hand side of (A.10) is bounded above by $\exp\{-c\tau_{n,p}^{1/3}(n\nu_{\mathbf{k}*})^{2/3}\}$ for some positive constant c. Since when $j \in \mathbf{k} \setminus \mathbf{t}$ and $\mathbf{t} \subsetneq \mathbf{k}$, $\beta^*_{\mathbf{k},j} \asymp n^{-1}$,

$$
\begin{aligned}
& P\left[\prod_{j\in\mathbf{k}\setminus\mathbf{t}} Q^U_{\mathbf{k},j} > \exp\{-|\mathbf{k}\setminus\mathbf{t}|\tau_{n,p}^{2/3}(n\nu_{\mathbf{k}*})^{1/3}\}\right] \\
\leq & \sum_{j\in\mathbf{k}\setminus\mathbf{t}} P\left[c'(n\nu_{\mathbf{k},j}+1/\tau_{n,p})^{-1/2}\exp\left\{-\frac{\tau_{n,p}}{(|\widetilde{\beta}_{\mathbf{k},j}|+\widetilde{\epsilon}_n)^2}\right\} > \exp\{-\tau_{n,p}^{2/3}(n\nu_{\mathbf{k}*})^{1/3}\}\right] \\
= & \sum_{j\in\mathbf{k}\setminus\mathbf{t}} P\left[\widetilde{\beta}^2_{\mathbf{k},j} > \left\{\tau_{n,p}^{1/2}\left((n\nu_{\mathbf{k}*})^{1/3}\tau_{n.p}^{2/3} - \log(n\nu_{\mathbf{k}*}+1/\tau_{n,p})/2 + \log c'\right)^{-1/2} - \widetilde{\epsilon}_n\right\}^2\right] \\
\leq & \sum_{j\in\mathbf{k}\setminus\mathbf{t}} P\left[(\widetilde{\beta}_{\mathbf{k},j}-\beta^*_{\mathbf{k},j})^2/\sigma^*_{\mathbf{k},j} > c''\left(\frac{\tau_{n,p}}{n\nu_{\mathbf{k}*}}\right)^{1/3}(n\nu_{\mathbf{k}*}+1/\tau_{n,p})/\sigma^2\right],
\end{aligned}
$$

for some positive contant c' and c''. Since $(\widetilde{\beta}_{\mathbf{k},j}-\beta^*_{\mathbf{k},j})^2/\sigma^*_{\mathbf{k},j} \sim \chi^2_1$, by Lemma A.1.4 the last quantity in the above display can be bounded by $\exp\{-c\tau_{n,p}^{1/3}(n\nu_{\mathbf{k}*})^{2/3}\}$ for some contant c. By *Assumption 3*, $\exp\{-c\tau_{n,p}^{1/3}(n\nu_{\mathbf{k}*})^{2/3}\} \prec p^{-q_n(1+\delta)} \leq p^{|\mathbf{k}\setminus\mathbf{t}|(1+\delta)|}$, which proves the statement (A.10).

We now shall show that the equation (A.9) holds for any $\delta > 0$. The left-hand side of (A.9)

can be bounded above by

$$
\begin{aligned}
& P\left[\prod_{j\in\mathbf{k}} Q^U_{\mathbf{k},j}\Big(\prod_{j\in\mathbf{t}} Q^L_{\mathbf{t},j}\Big)^{-1} > \exp\left\{\left\|\beta^0_{\mathbf{t}}\right\|_2^2 n\nu_{\mathbf{u}*}/\{2\log(\tau_{n,p}/\log p)\}\right\}\right] \\
\leq\ & \sum_{j\in\mathbf{k}} P\Bigg[c(n\nu_{\mathbf{k}*}+1/\tau_{n,p})^{-1/2}\exp\left\{-\tau_{n,p}/(|\widetilde{\beta}_{\mathbf{k},j}|+\widetilde{\epsilon}_n)^2\right\} \\
& > \exp\left\{\left\|\beta^0_{\mathbf{t}}\right\|_2^2 n\nu_{\mathbf{u}*}/\{4|\mathbf{k}|\log(\tau_{n,p}/\log p)\}\right\}\Bigg] \\
& + \sum_{j\in\mathbf{t}} P\Bigg[c'(n\nu_{\mathbf{t}*}+1/\tau_{n,p})^{1/2}\exp\left\{\tau_{n,p}/(\widetilde{\beta}^{*2}_{\mathbf{t},j})\right\} \\
& > \exp\left\{\left\|\beta^0_{\mathbf{t}}\right\|_2^2 n\nu_{\mathbf{u}*}/\{4|\mathbf{t}|\log(\tau_{n,p}/\log p)\}\right\}\Bigg] \\
\leq\ & \sum_{j\in\mathbf{k}} P\left[-\frac{\tau_{n,p}}{(|\widetilde{\beta}_{\mathbf{k},j}|+\widetilde{\epsilon}_n)^2} > \left\|\beta^0_{\mathbf{t}}\right\|_2^2 n\nu_{\mathbf{u}*}/\{4|\mathbf{k}|\log(\tau_{n,p}/\log p)\}+\log c\right] \quad \text{(A.12)} \\
& + \sum_{j\in\mathbf{t}} P\left[|\widetilde{\beta}^*_{\mathbf{t},j}| < c''\left\|\beta^0_{\mathbf{t}}\right\|_2^{-1}(n\nu_{\mathbf{u}*})^{-1/2}\{4|\mathbf{t}|\log(\tau_{n,p}/\log p)\}^{1/2}\tau_{n,p}^{1/2}\right], \quad \text{(A.13)}
\end{aligned}
$$

where c, c', and c'' are some positive constants.

(A.12) is always zero since the left-hand side in the probability is always negative and the right-hand side in the probability operator is always positive. So, we focus on (A.13) as below:

Since $\widetilde{\beta}_{\mathbf{t},j}-\widetilde{\epsilon}_n \leq \widetilde{\beta}^*_{\mathbf{t},j} \leq \widetilde{\beta}_{\mathbf{t},j}+\widetilde{\epsilon}_n$ implies $|\widetilde{\beta}_{\mathbf{t},j}|-\widetilde{\epsilon}_n \leq |\widetilde{\beta}^*_{\mathbf{t},j}| \leq |\widetilde{\beta}_{\mathbf{t},j}|+\widetilde{\epsilon}_n$, (A.13) can be bounded above by

$$
\begin{aligned}
& \sum_{j\in\mathbf{t}} P\left[|\widetilde{\beta}^*_{\mathbf{t},j}| < c''\left\|\beta^0_{\mathbf{t}}\right\|_2^{-1}(n\nu_{\mathbf{u}*})^{-1/2}\{4|\mathbf{t}|\log(\tau_{n,p}/\log p)\}^{1/2}\tau_{n,p}^{1/2}\right] \\
\leq\ & \sum_{j\in\mathbf{t}} P\left[|\widetilde{\beta}_{\mathbf{t},j}| < c''\left\|\beta^0_{\mathbf{t}}\right\|_2^{-1}(n\nu_{\mathbf{u}*})^{-1/2}\{4|\mathbf{t}|\log(\tau_{n,p}/\log p)\}^{1/2}\tau_{n,p}^{1/2}+\widetilde{\epsilon}_n\right],
\end{aligned}
$$

where $\beta^*_{\mathbf{t},j}$ is defined in (A.4). Since $\widetilde{\beta}^2_{\mathbf{t},j}/\sigma^2_{\mathbf{t},j} \sim \chi^2_1(\beta^{*2}_{\mathbf{t},j}/\sigma^2_{\mathbf{t},j})$ and $\sigma^2_{\mathbf{t},j} \asymp \sigma^2/n$ for $j \in \mathbf{t}$, by using Lemma A.1.4 and *Assumption 5*, one can show that the probability is bounded by $\exp\{-cn\}$ for some constant c, and it is evident that $\exp\{-cn\} \prec p^{-|\mathbf{k}|(1+\delta)}$, which completes

the proof of the Lemma. □

Proofs of Main Results

Proof of Theorem 1. We have $\pi(\mathbf{t} \mid \mathbf{y}) = \mathbf{m_t}(\mathbf{y})\pi(\mathbf{t})/\{\sum_{\mathbf{k}:|\mathbf{k}|\leq \mathbf{q_n}} \mathbf{m_k}(\mathbf{y})\pi(\mathbf{k})\}$, since $\pi(\mathbf{k}) = 0$ for any $\mathbf{k}$ with $|\mathbf{k}| > q_n$. Recall H_{1n} and H_{2n} from (A.5) and note that $\pi(\mathbf{t} \mid y) = (1 + H_{1n} + H_{2n})^{-1}$. Hence to show that $\pi(\mathbf{t} \mid y)$ converges to one in probability, it is sufficient to establish that H_{1n} and H_{2n} both converge in probability to zero as n tends to ∞. We shall prove the Theorem by showing:

For any $\delta \in (0, 8/3)$ and any model $\mathbf{k} \in \Gamma_d$ (defined in Lemma A.1.2),

$$P\left[\frac{\pi(\mathbf{k} \mid y)}{\pi(\mathbf{t} \mid y)} > \epsilon p^{-d} q_n^{-1}\right] \leq p^{-d(1+\delta)}, \tag{A.14}$$

and for any model $\mathbf{k} \in \Gamma_{k,c_k,c_t}$ (defined in Lemma A.1.3),

$$P\left[\frac{\pi(\mathbf{k} \mid y)}{\pi(\mathbf{t} \mid y)} > \epsilon n^{-3} p^{-k} n^{-c_k} t^{-t}\right] \leq c p^{-k(1+\delta)}. \tag{A.15}$$

Then, it is evident that H_{1n} and H_{2n} both converge to zero in probability by Lemma A.1.2 and A.1.3 respectively.

First, we shall show that (A.14) holds. For any $\mathbf{k} \in \Gamma_d$, recall that

$$P\Big[\frac{\pi(\mathbf{k} \mid y)}{\pi(\mathbf{t} \mid y)} > \epsilon p^{-d} q_n^{-1}\Big] \leq P\Big[C_{n,p}^{-d} \frac{Q_\mathbf{k}}{Q_\mathbf{t}} \exp\Big\{ -\frac{1}{2\sigma^2}\big(\widetilde{R}_\mathbf{k} - \widetilde{R}_\mathbf{t}\big)\Big\} > \epsilon p^{-d}/q_n\Big].$$

Since $\widetilde{R}_\mathbf{k} > R^*_\mathbf{k}$ and $\widetilde{R}_\mathbf{t} < R^*_\mathbf{t} + \eta$, where $\eta = d_1 \widehat{\beta}_\mathbf{t}^T \widehat{\beta}_\mathbf{t}/\tau_{n,p}$ for some constant d_1 and $\widehat{\beta}_\mathbf{t}$ is the ordinary least square estimator of $\beta_\mathbf{t}$ in the true model $\mathbf{t}$, by using (A.3), the term in the last

display can be bounded above by

$$
\begin{aligned}
& P\Big[C_{n,p}^{-d}\frac{Q_{\mathbf{k}}}{Q_{\mathbf{t}}}\exp\big\{-\big(R_{\mathbf{k}}^{*}-R_{\mathbf{t}}^{*}\big)/(2\sigma^{2})+\eta/(2\sigma^{2})\big\}>\epsilon p^{-d}/q_{n}\Big] \\
\leq\ & P\Big[C_{n,p}^{-d}\frac{Q_{\mathbf{k}}}{Q_{\mathbf{t}}}p^{d(1+\delta)+\delta}>\epsilon p^{-d}/q_{n}\Big] && \text{(A.16)} \\
& +P\left[R_{\mathbf{t}}^{*}-R_{\mathbf{k}}^{*}>2\sigma^{2}d(1+\delta)\log p\right] && \text{(A.17)} \\
& +P\left[\exp\{\eta/(2\sigma^{2})\}>\epsilon p^{\delta}\right]. && \text{(A.18)}
\end{aligned}
$$

By using Lemma A.1.5, (A.16) is less than $p^{-d(1+\delta)}$ when $\delta < 8/3$. Since $(R_{\mathbf{t}}^{*}-R_{\mathbf{k}}^{*})/\sigma^{2} \sim \chi^{2}_{|\mathbf{k}\setminus\mathbf{t}|}$, by using (A.6) in Lemma A.1.4, we can show that (A.17) is bounded by $cp^{-d(1+\delta)}$ for some positive constant c. Since $\tau_{n,p}n\nu_{\mathbf{t}*}\eta/d_{1}\sigma^{2} \leq \widehat{\beta}_{\mathbf{t}}^{T}X_{\mathbf{t}}^{T}X_{\mathbf{t}}\widehat{\beta}_{\mathbf{t}}/\sigma^{2} \sim \chi^{2}_{|\mathbf{t}|}\left(\beta_{\mathbf{t}}^{0T}X_{\mathbf{t}}^{T}X_{\mathbf{t}}\beta_{\mathbf{t}}^{0}\right)$, by using the inequality (A.7) in Lemma A.1.4, (A.18) can be expressed as

$$
\begin{aligned}
& P\left[\exp\left\{\eta/2\sigma^{2}\right\}>\epsilon p^{\delta}\right] \\
\leq\ & P\left[\tau_{n,p}n\nu_{\mathbf{t}*}\eta/d_{1}\sigma^{2}>2\tau_{n,p}n\nu_{\mathbf{t}*}(\log\epsilon+\delta\log p)/d_{1}\right] \\
\leq\ & P\big[\widehat{\beta}_{\mathbf{t}}^{T}X_{\mathbf{t}}^{T}X_{\mathbf{t}}\widehat{\beta}_{\mathbf{t}}/\sigma^{2}>2\tau_{n,p}n\nu_{\mathbf{t}*}(\log\epsilon+\delta\log p)/d_{1}\big] \\
\leq\ & (n\delta\log p)^{|\mathbf{t}|/2}\exp\{-c_{1}\delta(n\log p)\}+n^{-1/2}(\delta\log p)^{-1}\exp\{-c_{2}(n\log p)^{2}/n\} \\
\leq\ & c_{3}p^{-|\mathbf{k}|(1+\delta)}, && \text{(A.19)}
\end{aligned}
$$

for some positive constant c_1, c_2, and c_3, which proves that (A.14) holds.

Next, we consider (A.15). Recall that $\mathbf{u} = \mathbf{k} \cup \mathbf{t}$. By using (A.3), it can be shown that

$$
\begin{aligned}
& P\Big[\frac{\pi(\mathbf{k} \mid y)}{\pi(\mathbf{t} \mid y)} > \epsilon n^{-3} p^{-|\mathbf{k}|} n^{-|\mathbf{k}\setminus\mathbf{t}|} |\mathbf{t}|^{-|\mathbf{t}|}\Big] \\
\leq\ & P\Big[C_{n,p}^{-(|\mathbf{k}|-|\mathbf{t}|)} \frac{Q_{\mathbf{k}}}{Q_{\mathbf{t}}} \exp\Big\{-(\widetilde{R}_{\mathbf{k}} - \widetilde{R}_{\mathbf{t}})/(2\sigma^2)\Big\} > \epsilon n^{-3} p^{-|\mathbf{k}|} n^{-|\mathbf{k}\setminus\mathbf{t}|} |\mathbf{t}|^{-|\mathbf{t}|}\Big] \\
\leq\ & P\Big[C_{n,p}^{-|\mathbf{k}|-|\mathbf{t}|)} \frac{Q_{\mathbf{k}}}{Q_{\mathbf{t}}} \exp\big\{-(R^*_{\mathbf{k}} - R^*_{\mathbf{u}})/(2\sigma^2)\big\} > n^{-3-|\mathbf{k}\setminus\mathbf{t}|} |\mathbf{t}|^{-|\mathbf{t}|} p^{-|\mathbf{k}|(2+\delta)+\delta}\Big] \\
& + P\big[\exp\big\{\big(R^*_{\mathbf{t}} - R^*_{\mathbf{u}}\big)/(2\sigma^2)\big\} \geq \epsilon p^{|\mathbf{k}|(1+\delta)}\big] + P\big[\exp\big(\eta/(2\sigma^2)\big) > p^{\delta}\big] \\
\leq\ & P\big[\exp\big\{\big(R^*_{\mathbf{t}} - R^*_{\mathbf{u}}\big)/2\sigma^2\big\} > \epsilon p^{|\mathbf{k}|(1+\delta)}\big] && \text{(A.20)} \\
& + P\big[\exp\big(\eta/2\sigma^2\big) > p^{\delta}\big] && \text{(A.21)} \\
& + P\Big[R^*_{\mathbf{k}} - R^*_{\mathbf{u}} < 2\sigma^2 \big\|\beta^0_{\mathbf{t}}\big\|_2^2 n\nu_{\mathbf{u}*}/\log(\tau_{n,p}/\log p)\Big] && \text{(A.22)} \\
& + P\Big[Q_{\mathbf{k}}/Q_{\mathbf{t}} > \exp\Big\{\big\|\beta^0_{\mathbf{t}}\big\|_2^2 n\nu_{\mathbf{u}*}/\{2\log(\tau_{n,p}/\log p)\}\Big\}\Big]. && \text{(A.23)}
\end{aligned}
$$

Since $(R^*_{\mathbf{t}} - R^*_{\mathbf{u}})/\sigma^2$ follows a $\chi^2_{|\mathbf{u}\setminus\mathbf{t}|}$ distribution, (A.20) is also bounded by $c_1 p^{-|\mathbf{k}|(1+\delta)}$ with some constant c_1. By following the same steps regarding (A.19), one can show that (A.21) is bounded by $c_2 p^{-|\mathbf{k}|(1+\delta)}$ for some constant c_2. We note that $(R^*_{\mathbf{k}} - R^*_{\mathbf{u}})/\sigma^2 \sim \chi^2_{|\mathbf{u}\setminus\mathbf{k}|}(\lambda_n)$ with $\lambda_n = \beta^{0T}_{\mathbf{t}} X^T_{\mathbf{t}} (P_{\mathbf{u}} - P_{\mathbf{k}}) X_{\mathbf{t}} \beta^0_{\mathbf{t}}$, where $P_{\mathbf{k}}$ is the projection matrix of $X_{\mathbf{k}}$. As discussed in Narisetty and He (2014), $\lambda_n \geq n\nu_{\mathbf{u}*}\big\|\beta^0_{\mathbf{t}}\big\|_2^2$. Hence, by using Lemma A.1.4, one can show that (A.22) is bounded by $\exp\{-c_3\big\|\beta^0_{\mathbf{t}}\big\|_2^2 n\nu_{\mathbf{u}*}/\log(\tau_{n,p}/\log p)\}$ for some constant c_3. Lemma A.1.5 states that (A.23) is bounded by $p^{-|\mathbf{k}|(1+\delta)}$. In summary, since $q_n \prec \tau_{n,p}/\log p$ by *Assumption 3*, there exists some positive constant c_4 such that $P[\pi(\mathbf{k} \mid y)/\pi(\mathbf{t} \mid y) > \epsilon n^{-3} p^{-|\mathbf{k}|} n^{-|\mathbf{k}\setminus\mathbf{t}|} |\mathbf{t}|^{-|\mathbf{t}|}] \leq c_4 p^{-|\mathbf{k}|(1+\delta)}$. which completes the proof of Theorem 1. □

Proof of Corollary 2. Recall the penalty term of a model $\mathbf{k}$, $Q^*_{\mathbf{k}}$, based on the piMoM priors is

$$
Q^*_{\mathbf{k}} = \int \exp\big\{ -(\beta_{\mathbf{k}} - \widehat{\beta}_{\mathbf{k}})^T \Sigma^{*-1}_{\mathbf{k}} (\beta_{\mathbf{k}} - \widehat{\beta}_{\mathbf{k}})/(2\sigma^2) - \sum_{j=1}^{|\mathbf{k}|} \tau_{n,p}/\beta^2_{\mathbf{k},j} - r\sum_{j=1}^{|\mathbf{k}|} \log(\beta^2_{\mathbf{k},j})\big\} d\beta_{\mathbf{k}},
$$

in (2.7). Since, for any $\epsilon > 0$, $\exp\big[-\sum_{j=1}^{|\mathbf{k}|}\{\epsilon\tau_{n,p}/\beta_{\mathbf{k},j}^2 + r\log(\beta_{\mathbf{k},j}^2)\}\big]$ is bounded above with respect to $\beta_{\mathbf{k},j}$, $Q_{\mathbf{k}}^* \leq C\int \exp\{-(\beta_{\mathbf{k}}-\widehat{\beta}_{\mathbf{k}})^T\Sigma_{\mathbf{k}}^{*-1}(\beta_{\mathbf{k}}-\widehat{\beta}_{\mathbf{k}})/(2\sigma^2) - \sum_{j=1}^{|\mathbf{k}|}(1-\epsilon)\tau_{n,p}/\beta_{\mathbf{k},j}^2\}d\beta_{\mathbf{k}}$ for some constant C. Following the exactly same steps in Lemma A.1.1,

$Q_{\mathbf{k}}^* \leq C'(n\nu_{\mathbf{k}}^*)^{-1/2}\prod_{j=1}^{|\mathbf{k}|}\exp\{-(1-\epsilon)\tau_{n,p}/(|\widehat{\beta}_{\mathbf{k},j}| + \widetilde{\epsilon}_n)^2\}$ for some constant $C' > 0$.

We shall show that the model selection procedure based on piMoM priors as in (2.4) assures consistency by proving that $Q_{\mathbf{k}}^*$ and $Q_{\mathbf{k}}$ are asymptotically equivalent.

Next, we shall show that $Q_{\mathbf{k}}^*$ is bounded below by

$C(n\nu_{\mathbf{k}}^*)^{-1/2}\prod_{j=1}^{|\mathbf{k}|}\exp\{-(1-\epsilon)\tau_{n,p}/\widehat{\beta}_{\mathbf{k},j}^{*2}\}$ for some constant $C > 0$ and $\widehat{\beta}_{\mathbf{k},j}^* \in [\widehat{\beta}_{\mathbf{k},j} - \widetilde{\epsilon}_n, \widehat{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n]$. Since $\exp\big\{-\epsilon\tau_{n,p}/\beta_{\mathbf{k},j}^2 + r\log(\beta_{\mathbf{k},j}^2)\big\}$ can be minimized in $[\widehat{\beta}_{\mathbf{k},j} - \widetilde{\epsilon}_n, \widehat{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n]$, by following the proof of Lemma A.1.1,

$$
\begin{aligned}
& \int_{-\infty}^{\infty}\exp\{-n\nu_{\mathbf{k}}^*(\beta-\widehat{\beta}_{\mathbf{k},j})^2/(2\sigma^2) - \tau_{n,p}/\beta^2 - r\log(\beta^2)\}d\beta \\
\geq\ & \int_{\widehat{\beta}_{\mathbf{k},j}-\widetilde{\epsilon}_n}^{\widehat{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n}\exp\{-n\nu_{\mathbf{k}}^*(\beta-\widehat{\beta}_{\mathbf{k},j})^2/(2\sigma^2) - (1-\epsilon)\tau_{n,p}/\beta^2\} \\
& \times\exp\{-\epsilon\tau_{n,p}/\beta^2 - r\log(\beta^2)\}d\beta \\
\geq\ & C(n\nu_{\mathbf{k}}^*)^{-1/2}\exp\left\{-(1-\epsilon)\tau_{n,p}/\widehat{\beta}_{\mathbf{k},j}^{*2}\right\},
\end{aligned}
$$

where C is some constant and $\widehat{\beta}_{\mathbf{k},j}^* \in [\widehat{\beta}_{\mathbf{k},j} - \widetilde{\epsilon}_n, \widehat{\beta}_{\mathbf{k},j} + \widetilde{\epsilon}_n] \setminus (-\widetilde{\epsilon}_n, \widetilde{\epsilon}_n)^c$.

Therefore, due to the asymptotic similarity between the ridge estimator and the least square estimator, the lower and upper bounds of $Q_{\mathbf{k}}^*$ are asymptotically equivalent to those of $Q_{\mathbf{k}}$ with the penalty parameter $(1-\epsilon)\tau_{n,p}$, which assures the strong consistency of the model selection based on the piMoM priors. □

Proof of Theorem 3. Under a situation where σ^2 is unknown, it is clear that

$$\begin{aligned} m_{\mathbf{k}}(y) &= \tau_{n,p}^{-\frac{|\mathbf{k}|}{2}} \int (2\pi\sigma^2)^{-\frac{n+|\mathbf{k}|}{2}} \int \exp\left\{ |\mathbf{k}| \left(\frac{2}{\sigma^2}\right)^{1/2} - \frac{(\beta_{\mathbf{k}} - \widetilde{\beta}_{\mathbf{k}})^T \widetilde{\Sigma}_{\mathbf{k}}^{-1} (\beta_{\mathbf{k}} - \widetilde{\beta}_{\mathbf{k}})}{2\sigma^2} \right\} \\ &\quad \times \exp\left\{ -\sum_{j=1}^{|\mathbf{k}|} \frac{\tau_{n,p}}{\beta_{\mathbf{k},j}^2} \right\} \pi(\sigma^2) d\beta_{\mathbf{k}} d\sigma^2, \end{aligned}$$

where $\pi(\sigma^2)$ is the prior for σ^2 (Inverse-gamma density with hyperparameters a_0 and b_0).

First, we shall show that the ratio between marginal likelihoods of a model $\mathbf{k}$ and the true model $\mathbf{t}$ can be bounded as

$$\begin{aligned} \frac{m_{\mathbf{k}}(y)}{m_{\mathbf{t}}(y)} &\leq c^{\frac{|\mathbf{k}|-|\mathbf{t}|}{2}} \left(\frac{\widetilde{R}_{\mathbf{k}} + 2b_0}{\widetilde{R}_{\mathbf{t}} + 2b_0} \right)^{-n/2-a_0} \exp\left\{ -\sum_{j=1}^{|\mathbf{k}|} \frac{\tau_{n,p}}{(|\widetilde{\beta}_{\mathbf{k},j}| + \widetilde{\epsilon}_n)^2} + \sum_{j=1}^{|\mathbf{t}|} \frac{\tau_{n,p}}{\widetilde{\beta}_{\mathbf{t},j}^{*2}} \right\} \\ &\quad \times \frac{(n\nu_{\mathbf{k}*}\tau_{n,p} + 1)^{-|\mathbf{k}|/2}}{(n\nu_{\mathbf{t}}^{*}\tau_{n,p} + 1)^{-|\mathbf{t}|/2}}, \end{aligned} \tag{A.24}$$

where $\widetilde{\beta}_{\mathbf{t},j}^{*} \in [\widetilde{\beta}_{\mathbf{t},j} - \widetilde{\epsilon}_n, \widetilde{\beta}_{\mathbf{t},j} + \widetilde{\epsilon}_n] \setminus (-\widetilde{\epsilon}_n, \widetilde{\epsilon}_n)^c$ for $j \in 1, \ldots, |\mathbf{t}|$ and c is some constant. Next, we shall show that $\{(\widetilde{R}_{\mathbf{k}} + 2b_0)/(\widetilde{R}_{\mathbf{t}} + 2b_0)\}^{-n/2-a_0} \leq \exp\{-(\widetilde{R}_{\mathbf{k}} - \widetilde{R}_{\mathbf{t}})/(2\sigma_0^2(1+u_n))\}$, where σ_0^2 is the true regression variance that involves in the data-generating process, and u_n is some random variable that is concentrated around a finite value with at least probability $1 - \exp\{-cn\}$ for some constant c. Then, by following the same steps in the proof of Theorem 1, the proof of Corollary 2 is completed.

By Lemma A.1.1, the marginal likelihood of a model $\mathbf{k}$ can be bounded by

$$\begin{aligned}
& m_{\mathbf{k}}(y) \\
\leq\ & \{c_1(n\nu_{\mathbf{k}*}\tau_{n,p}+1)\}^{-\frac{|\mathbf{k}|}{2}} \int (\sigma^2)^{-\frac{n+2a_0}{2}-1} \exp\left\{|\mathbf{k}|\left(\frac{2}{\sigma^2}\right)^{1/2} - \frac{\widetilde{R}_{\mathbf{k}}+2b_0}{2\sigma^2}\right\} \\
& \times \exp\left\{-\sum_{j=1}^{|\mathbf{k}|} \frac{\tau_{n,p}}{(|\widetilde{\beta}_{\mathbf{k},j}|+\widetilde{\epsilon}_n)^2}\right\} d\sigma^2 \\
\leq\ & \{c_1(n\nu_{\mathbf{k}*}\tau_{n,p}+1)\}^{-\frac{|\mathbf{k}|}{2}} \exp\left\{-\sum_{j=1}^{|\mathbf{k}|} \frac{\tau_{n,p}}{(|\widetilde{\beta}_{\mathbf{k},j}|+\widetilde{\epsilon}_n)^2}\right\} \\
& \times (1+\exp\{2|\mathbf{k}|\})\left(\widetilde{R}_{\mathbf{k}}+2b_0\right)^{-\frac{n+2a_0}{2}},
\end{aligned}$$

for some constant c_1.

Also, by using Lemma A.1.1, one can show that

$$\begin{aligned}
m_{\mathbf{k}}(y) \ \geq\ & \{c_2(n\nu_{\mathbf{k}*}\tau_{n,p}+1)\}^{-\frac{|\mathbf{k}|}{2}} \int (\sigma^2)^{-\frac{n+2a_0}{2}-1} \exp\left\{|\mathbf{k}|\left(\frac{2}{\sigma^2}\right)^{1/2} - \frac{\widetilde{R}_{\mathbf{k}}+2b_0}{2\sigma^2}\right\} \\
& \times \exp\left\{-\sum_{j=1}^{|\mathbf{k}|} \frac{\tau_{n,p}}{\widetilde{\beta}^{*2}_{\mathbf{k},j}}\right\} d\sigma^2 \\
\geq\ & \{c_2(n\nu_{\mathbf{k}*}\tau_{n,p}+1)\}^{-\frac{|\mathbf{k}|}{2}} \exp\left\{-\sum_{j=1}^{|\mathbf{k}|} \frac{\tau_{n,p}}{\widetilde{\beta}^{*2}_{\mathbf{k},j}}\right\} \left(\widetilde{R}_{\mathbf{k}}+2b_0\right)^{-\frac{n+2a_0}{2}},
\end{aligned}$$

where c_2 is some constant and $\widetilde{\beta}^*_{\mathbf{k},j} \in [\widetilde{\beta}_{\mathbf{k},j}-\widetilde{\epsilon}_n, \widetilde{\beta}_{\mathbf{k},j}+\widetilde{\epsilon}_n] \setminus (-\widetilde{\epsilon}_n, \widetilde{\epsilon}_n)^c$ for $j \in 1,\ldots,|\mathbf{k}|$. These results shows that (A.24) holds.

Next, we consider the asymptotic behavior of $\{(\widetilde{R}_{\mathbf{k}}+2b_0)/(\widetilde{R}_{\mathbf{t}}+2b_0)\}^{-n/2-a_0}$ in (A.24). Define ρ_n as the follows:

$$\rho_n = (\widetilde{R}_{\mathbf{t}}+2b_0)/(n\sigma_0^2) - 1.$$

Since $-\log(1-u) < u/(1-u)$ for $u \in \mathbb{R}$,

$$\begin{aligned} -\log\{(\widetilde{R}_{\mathbf{k}} + 2b_0)/(\widetilde{R}_{\mathbf{t}} + 2b_0)\} &= -\log[1 + (\widetilde{R}_{\mathbf{k}} - \widetilde{R}_{\mathbf{t}})/\{n(1+\rho_n)\sigma_0^2\}] \\ &\leq (\widetilde{R}_{\mathbf{t}} - \widetilde{R}_{\mathbf{k}})/\{n\sigma_0^2(1+u_n)\}, \end{aligned}$$

where $u_n = \rho_n + (\widetilde{R}_{\mathbf{k}} - \widetilde{R}_{\mathbf{t}})/(n\sigma_0^2)$.

Since $(R^*_{\mathbf{k}} - R^*_{\mathbf{u}})/\sigma_0^2 \sim \chi_{|\mathbf{u}\setminus\mathbf{k}|}(\lambda_n)$ with $\lambda_n = \beta^{0T}_{\mathbf{t}} X^T_{\mathbf{t}}(P_{\mathbf{u}} - P_{\mathbf{k}})X_{\mathbf{t}}\beta^0_{\mathbf{t}}/\sigma_0^2$, by using Lemma A.1.4 one can show that

$$\begin{aligned} P\left(|u_n - \lambda_n/n| > \epsilon\right) &\leq P\left(|\rho_n| > \epsilon/4\right) + P\left\{(R^*_{\mathbf{t}} - R^*_{\mathbf{u}})/(n\sigma_0^2) > \epsilon/4\right\} \\ &\quad + P\left\{\left|(R^*_{\mathbf{k}} - R^*_{\mathbf{u}})/(n\sigma_0^2) - \lambda_n/n\right| > \epsilon/4\right\} + P\left(\eta/2n\sigma_0^2 > \epsilon/4\right) \\ &\leq \exp\{-c'n\} + P\left\{\left|(R^*_{\mathbf{k}} - R^*_{\mathbf{u}})/(n\sigma_0^2) - \lambda_n/n\right| > \epsilon/4\right\} \\ &\leq \exp\{-c''n\}, \end{aligned}$$

for some constant c' and c'', and η is defined in the proof of Theorem 1. Also, by Assumption 5, λ_n/n will be bounded below and above. □

Proof of Corollary 4. Since we showed that the asymptotic equivalence between $Q_{\mathbf{k}}$ and $Q^*_{\mathbf{k}}$ in the proof of Corollary 2, by following exactly same steps in the proof of Theorem 3 we can prove the model selection consistency under piMoM prior densities. □

Proof of Proposition 5. We shall show that for any $\alpha_{\mathbf{k}} = \widehat{\beta}_{\mathbf{k}} + \boldsymbol{\epsilon}_n$ with $\boldsymbol{\epsilon}_n = \{\epsilon_{n,j}\}_{j=1,\ldots,|\mathbf{k}|}$ and $|\epsilon_{n,j}| \succ \epsilon^*_n$ for at least one $j \in \{1,\ldots,|\mathbf{k}|\}$, $P\{g(\alpha_{\mathbf{k}};\mathbf{k}) < g(\widetilde{\beta}^*_{\mathbf{k}};\mathbf{k})\} \to 0$ as n tends to ∞, where $\widetilde{\beta}^*_{\mathbf{k}} \in B(\widehat{\beta}_{\mathbf{k}};\epsilon^*_n)$ with $\epsilon^*_n \asymp (\tau_{n,p}/n)^{1/3}$. More specifically, we set $\widetilde{\beta}^*_{\mathbf{k},j} = \widehat{\beta}_{\mathbf{k},j} + \epsilon^*_n$ for $j \in \mathbf{t}$ and $\widetilde{\beta}^*_{\mathbf{k},j} = \widehat{\beta}_{\mathbf{k},j}$ for $j \in \mathbf{t}^c$. Without loss of generality, we assume that $X_j^{\mathrm{T}} X_j = n$ for $j = 1,\ldots,p$.

Note that

$$\begin{aligned} g(\alpha_{\mathbf{k}};\mathbf{k}) &= ||X_{\mathbf{k}}\alpha_{\mathbf{k}} - X_{\mathbf{k}}\widehat{\beta}_{\mathbf{k}}||_2^2 + \sum_{j=1}^{|\mathbf{k}|}\tau_{n,p}/|\alpha_{\mathbf{k},j}| + D_n \\ &= \sum_{j=1}^{|\mathbf{k}|}\{c_j n\epsilon_{n,j}^2 + \tau_{n,p}/|\widehat{\beta}_{\mathbf{k},j} + \epsilon_{n,j}|\} + D_n, \end{aligned}$$

for some constants c_j such that $C_L < c_j < C_U$ for $j = 1, \ldots, |\mathbf{k}|$, and some randome variable D_n that are not relevant to $\alpha_{\mathbf{k}}$. Then,

$$\begin{aligned} & P\{g(\alpha_{\mathbf{k}};\mathbf{k}) < g(\widetilde{\beta}^*_{\mathbf{k}};\mathbf{k})\} \\ \leq\ & P\left[\sum_{j=1}^{|\mathbf{k}|}\left\{c_j n\epsilon_{n,j}^2 + \frac{\tau_{n,p}}{|\widehat{\beta}_{\mathbf{k},j} + \epsilon_{n,j}|}\right\} < \sum_{j=1}^{|\mathbf{k}|}\left\{c_j n\epsilon_n^{*2} + \frac{\tau_{n,p}}{|\widetilde{\beta}^*_{\mathbf{k},j}|}\right\}\right] \\ \leq\ & P\Bigg[\sum_{j\in S^*\cap S_{\mathbf{k},n}}\left\{c_j n\epsilon_{n,j}^2 + \frac{\tau_{n,p}}{|\widehat{\beta}_{\mathbf{k},j}| + |\epsilon_{n,j}|} - t_{n,j}\right\} \\ & < \sum_{j\in S^*\cap S_{\mathbf{k},n}}\left\{c_j n\epsilon_n^{*2} + \frac{\tau_{n,p}}{|\widetilde{\beta}^*_{\mathbf{k},j}|}\right\}\Bigg] \qquad \text{(A.25)} \\ & + P\Bigg[\sum_{j\in S^*\cap S^c_{\mathbf{k},n}}\left\{c_j n\epsilon_{n,j}^2 + \frac{\tau_{n,p}}{|\widehat{\beta}_{\mathbf{k},j}| + |\epsilon_{n,j}|} - t_{n,j}\right\} \\ & < \sum_{j\in S^*\cap S^c_{\mathbf{k},n}}\left\{c_j n\epsilon_n^{*2} + \frac{\tau_{n,p}}{|\widetilde{\beta}^*_{\mathbf{k},j}|}\right\}\Bigg] \qquad \text{(A.26)} \\ & + P\Bigg[\sum_{j\in S^{*c}}\left\{c_j n\epsilon_{n,j}^2 + \frac{\tau_{n,p}}{|\widehat{\beta}_{\mathbf{k},j}| + |\epsilon_{n,j}|} + \sum_{j\in S^*}\frac{t_{n,j}}{|S^{*c}|}\right\} \\ & < \sum_{j\in S^{*c}}\left\{c_j n\epsilon_n^{*2} + \frac{\tau_{n,p}}{|\widetilde{\beta}^*_{\mathbf{k},j}|}\right\}\Bigg], \qquad \text{(A.27)} \end{aligned}$$

where t_n is an arbitrary sequence such that $t_{n,j} = n^{2/3}\tau_{n,p}^{1/3}\epsilon_{n,j}$, and $S^* = \{j \in \{1, \ldots, p\} : |\epsilon_{n,j}| \succ \epsilon_n^*\}$, and $S_{\mathbf{k},n} = \{j \in \mathbf{k} : |\widehat{\beta}_{\mathbf{k},j}| < \epsilon_n^*\}$. Then, to complete the proof, it is sufficient to

show that each of (A.25), (A.26), and (A.27) converges to zero.

Since $n(\widehat{\beta}_{\mathbf{k},j} - \beta^0_{\mathbf{t},j})^2/\sigma^2 \sim \chi^2_1$ for $j = 1, \ldots, |\mathbf{k}|$,

$$P(|\widehat{\beta}_{\mathbf{t},j} - \beta^0_{\mathbf{t},j}| > \zeta_n) \leq (\pi n \zeta_n^2/2)^{-1/2} \exp\{-n\zeta_n^2/(2\sigma^2)\},$$

for any $\zeta_n > 0$. This implies that $S_{\mathbf{k},n} = \mathbf{t}$ at least probability

$1 - |\mathbf{t}^c|(\pi n \epsilon_n^{*2}/2)^{-1/2} \exp\{-n\epsilon_n^{*2}/(2\sigma^2)\}$. Therefore, the equation (A.25) can be asymptotically bounded by

$$\begin{aligned} & \sum_{j \in S^* \cap \mathbf{t}} P\left[c_j n \epsilon_{n,j}^2 + \frac{\tau_{n,p}}{2|\epsilon_{n,j}|} - t_{n,j} < c_j n \epsilon_n^{*2} + \frac{\tau_{n,p}}{|\widehat{\beta}_{\mathbf{k},j} + \epsilon_n^*|} \right] \\ \leq & \sum_{j \in S^* \cap \mathbf{t}} P\left[|\widehat{\beta}_{\mathbf{k},j} + \epsilon_n^*| < c\tau_{n,p}(n\epsilon_{n,j}^2 - t_{n,j} + \tau_{n,p}/|\epsilon_{n,j}|)^{-1} \right], \end{aligned}$$

for some positive constant c. Consider Lemma A.1.4 with $\lambda_n = n\epsilon_n^{*2}/\sigma^2$ and

$w_n = c^2 \tau_{n,p}^2 / \{\epsilon_n^{*2}(n\epsilon_{n,j}^2 - t_{n,j} + \tau_{n,p}/|\epsilon_{n,j}|)^2\}$ for $j \in S^* \cap \mathbf{t}$. Since $n\epsilon_{n,j}^2 \succ n^{1/3}\tau_{n,p}^{2/3}$ for $j \in S^*$ implies $w_n \to 0$, Lemma A.1.4 guarantees that the last display is bounded by $c'|S^* \cap \mathbf{t}|\lambda_n^{-1} \exp\{-\lambda_n(1 - w_n)^2\}$ for some constant c', which means that (A.25) converges to zero as n tends to 0. By following the same steps, one can show that (A.26) converges to zero.

Also, (A.27) can be asymptotically bounded by

$$\begin{aligned} & \sum_{j \in S^{*c} \cap \mathbf{t}} P\left[c_j n \epsilon_{n,j}^2 + \frac{\tau_{n,p}}{2|\epsilon_{n,j}|} + c \min_{j \in S^*} t_{n,j} < c_j n \epsilon^{*2} + \frac{\tau_{n,p}}{|\widehat{\beta}_{\mathbf{k},j} + \epsilon_n^*|} \right] \\ & + \sum_{j \in S^{*c} \cap \mathbf{t}^c} P\left[c_j n \epsilon_{n,j}^2 + \frac{\tau_{n,p}}{2|\widehat{\beta}_{\mathbf{k},j} + \epsilon_n^*|} + c \min_{j \in S^*} t_{n,j} < c_j n \epsilon^{*2} + \frac{\tau_{n,p}}{|\widehat{\beta}_{\mathbf{k},j} + \epsilon_n^*|} \right] \\ \leq & \sum_{j \in S^{*c} \cap \mathbf{t}} P\left[|\widehat{\beta}_{\mathbf{k},j} + \epsilon_n^*| < c'\tau_{n,p}(n\epsilon_{n,j}^2 - n\epsilon_n^{*2} + c \min_{j \in S^*} t_{n,j} + \tau_{n,p}/|\epsilon_{n,j}|)^{-1} \right] \\ & + \sum_{j \in S^{*c} \cap \mathbf{t}^c} P\left[|\widehat{\beta}_{\mathbf{k},j} + \epsilon_n^*| < c''\tau_{n,p}(n\epsilon_{n,j}^2 - n\epsilon_n^{*2} + c \min_{j \in S^*} t_{n,j} + \tau_{n,p}/|\epsilon_{n,j}|)^{-1}/2 \right], \end{aligned}$$

where c, c', and c'' are some positive constants. For the first term in the last line of the above display, by setting $\lambda_n = n\epsilon^{*2}/\sigma^2$ and $w_n = c^2\tau_{n,p}^2/\{\epsilon_n^{*2}(n\epsilon_{n,j}^2 - n\epsilon_n^* + c\min_{j\in S^*} t_{n,j} + \tau_{n,p}/|\epsilon_{n,j}|)^2\}$, we can apply Lemma A.1.4. Since $w_n \prec \tau_{n,p}^2(\epsilon_n^* \min_{j\in S^*} t_{n,j})^{-2}$ implies $w_n \to 0$, the first term in the above display converges to zero by Lemma A.1.4. Similarly, the second term also converges to zero. □

A.2 Functional Horseshoe Prior for Nonparametric Subspace Shrinkage

Proof of Lemma 1. As discussed in the paragraphs following Lemma 1 when $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$, we can generate a new basis $\widetilde{\Phi} = [\Phi_0, \Phi_1]$ such that $\Phi_0^{\mathrm{T}}\Phi_1 = \mathbf{0}$ and $\mathfrak{L}(\Phi) = \mathfrak{L}(\widetilde{\Phi})$, which implies $Q_{\widetilde{\Phi}} = Q_{\Phi}$. Then,

$$
\begin{aligned}
& \Phi\left(\Phi^T\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\right)^{-1}\Phi^{\mathrm{T}} \\
= \;& \widetilde{\Phi}\left(\widetilde{\Phi}^T\widetilde{\Phi} + \frac{\omega}{1-\omega}\widetilde{\Phi}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\widetilde{\Phi}\right)^{-1}\widetilde{\Phi}^{\mathrm{T}} \\
= \;& [\Phi_0, \Phi_1]\begin{bmatrix} (\Phi_0^{\mathrm{T}}\Phi_0)^{-1} & \mathbf{0} \\ \mathbf{0} & (1-\omega)(\Phi_1^{\mathrm{T}}\Phi_1)^{-1} \end{bmatrix}\begin{bmatrix} \Phi_0^{\mathrm{T}} \\ \Phi_1^{\mathrm{T}} \end{bmatrix} \\
= \;& (1-\omega)Q_{\widetilde{\Phi}} + \omega Q_0 \\
= \;& (1-\omega)Q_{\Phi} + \omega Q_0.
\end{aligned}
$$

□

Proof of Lemma 2. From Polson and Scott (2012) it follows that

$$
\int_0^1 \omega^{A_n-1}(1-\omega)^{B_n-1}\exp\{-H_n\omega\}d\omega = \frac{\Gamma(A_n)\Gamma(B_n)}{\Gamma(A_n+B_n)}\exp\{-H_n\}\sum_{m=0}^{\infty}\frac{(A_n)_{(m)}}{(A_n+B_n)_{(m)}}\frac{{H_n}^m}{m!},
$$

where $(a)_{(m)} = a(a+1)\ldots(a+m-1)$. We shall show that $\sum_{m=0}^{\infty}\left\{\frac{(B_n)_{(m)}}{(A_n+B_n)_{(m)}}\frac{{H_n}^m}{m!}\right\} \geq 1+Q_n^L$. By using Lemma A.2.1 and Stirling's approximation, i.e., $m! \asymp m^{m+1/2}\exp\{-m\}$, it follows

that

$$
\begin{aligned}
& \sum_{m=0}^{\infty} \left\{ \frac{(B_n)_{(m)}}{(A_n+B_n)_{(m)}} \frac{{H_n}^m}{m!} \right\} \\
= \;& 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + \sum_{m=1}^{\infty} \left[\frac{(B_n+1)_{(m)}}{(A_n+B_n+1)_{(m)}} \frac{{H_n}^{m+1}}{(m+1)!} \right] \right\} \\
\geq \;& 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + \sum_{m=1}^{\infty} \left[\frac{(B_n+m)!}{(A_n+B_n+m)!} \frac{{H_n}^{m+1}}{(m+1)!} \right] \right\} \\
\geq \;& 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + D \sum_{m=1}^{\infty} \left[\left(\frac{B_n+m}{A_n+B_n+m} \right)^{A_n+B_n+m+1/2} (B_n+m)^{-A_n} \right. \right. \\
& \left. \left. e^{A_n} \frac{{H_n}^{m+1}}{(m+1)!} \right] \right\} \\
\geq \;& 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + D \sum_{m=1}^{T_n} \left[\left(\frac{B_n+1}{A_n+B_n+1} \right)^{1/2} (B_n+m)^{-A_n} \right. \right. \\
& \left. \left. \times \left(\frac{B_n+m}{A_n+B_n+m} \right)^{A_n+B_n+m} e^{A_n} \frac{{H_n}^{m+1}}{(m+1)!} \right] \right\} \\
\geq \;& 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + D \left(\frac{B_n+1}{A_n+B_n+1} \right)^{1/2} (B_n+T_n)^{-A_n} \right. \\
& \left. \times \exp\left\{ \frac{A_n^2}{2(A_n+B_n+T_n)} \right\} \sum_{m=2}^{T_n+1} \frac{{H_n}^m}{m!} \right\}, \qquad \text{(A.28)}
\end{aligned}
$$

where $T_n = \max\{A_n^2, 3\lceil H_n \rceil\}$, and D is some positive constant.

Since $H_n < (T_n+2)\exp\{1\}$, by using the Stirling's approximation, the term $\sum_{m=2}^{T_n+1} H_n/m!$ in (A.28) can be expressed as follows:

$$
\begin{aligned}
\sum_{m=2}^{T_n+1} \frac{H_n^m}{m!} &= \exp\{H_n\} - 1 - H_n - \sum_{m=T_n+2}^{\infty} \frac{H_n^m}{m!} \\
&\preceq \exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \sum_{m=T_n+2}^{\infty} \left(\frac{\exp\{1\} H_n}{T_n+2} \right)^m \\
&\leq \exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2}
\end{aligned}
$$

Therefore, (A.28) can be bounded by

$$\begin{aligned} & 1+\frac{B_n}{A_n+B_n}\Bigg\{H_n+D\left(\frac{B_n+1}{A_n+B_n+1}\right)^{1/2}(B_n+T_n)^{-A_n} \\ & \times\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+\Bigg\} \\ \geq\ & 1+\frac{B_nH_n}{A_n+B_n}+\frac{DB_n}{(A_n+B_n)^{3/2}}(B_n+T_n)^{-A_n}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+, \end{aligned}$$

where $(\cdot)_+$ denotes the positive hinge function (i.e., for any $t\in\mathbb{R}$, $(t)_+=t$, if $t>0$, and $(t)_+=0$, otherwise).

Also, since $(B_n+m)!/(A_n+B_n+m)!<1$ for any positive integer m, it follows that

$$H_n+\sum_{m=1}^{\infty}\left[\frac{(B_n+m)!}{(A_n+B_n+m)!}\frac{{H_n}^{m+1}}{(m+1)!}\right]\leq\exp\{H_n\},$$

which completes the proof. □

Lemma A.2.1. *For arbitrary positive sequences u_n and w_n,*

$$\left(1-\frac{u_n}{u_n+w_n}\right)^{u_n+w_n}\geq\exp\left\{-u_n+\frac{u_n^2}{2(u_n+w_n)}\right\}. \tag{A.29}$$

Proof. By Talyor's theorem, there exists $q_n^*\in(0,u_n/(u_n+w_n))$ such that

$$\begin{aligned} \left(1-\frac{u_n}{u_n+w_n}\right)^{u_n+w_n} &= \exp\left\{(u_n+w_n)\log\left(1-\frac{u_n}{u_n+w_n}\right)\right\} \\ &= \exp\left\{(u_n+w_n)\left(-\frac{u_n}{u_n+w_n}+\frac{1}{(1-q_n^*)^2}\frac{u_n^2}{2(u_n+w_n)^2}\right)\right\} \\ &\geq \exp\left\{-u_n+\frac{u_n^2}{2(u_n+w_n)}\right\}. \end{aligned}$$

□ □

Lemma A.2.2.

$$n\left\|Q_0\Phi\beta - Q_0Y\right\|_{n,2}^2/\sigma^2 \mid Y,\omega \sim \chi^2_{d_0},$$

and

$$n\left\|Q_1\Phi\beta - (1-\omega)Q_1Y\right\|_{n,2}^2/\{(1-\omega)\sigma^2\} \mid Y,\omega \sim \chi^2_{k_n-d_0}.$$

Proof. Recall that

$$\beta \mid Y,\omega \sim \mathbf{N}(\widetilde{\beta}_\omega, \widetilde{\Sigma}_\omega),$$

where

$$\widetilde{\beta}_\omega = \left(\Phi^{\mathrm{T}}\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\right)^{-1}\Phi^{\mathrm{T}}Y, \quad \widetilde{\Sigma}_\omega = \sigma^2\left(\Phi^{\mathrm{T}}\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\right)^{-1}.$$

As shown in the proof of Lemma 1, $\Phi\left(\Phi^T\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\right)^{-1}\Phi^{\mathrm{T}} = (1-\omega)Q_\Phi + \omega Q_0$, so

$$\mathbb{E}\left[Q_0\Phi\beta \mid Y,\omega\right] = Q_0Y$$
$$\mathrm{Var}\left[Q_0\Phi\beta \mid Y,\omega\right] = \sigma^2Q_0,$$

which shows that $n\left\|Q_0\Phi\beta - Q_0Y\right\|_{n,2}^2/\sigma^2 \mid Y,\omega \sim \chi^2_{d_0}$.

Similarly,

$$\mathbb{E}\left[Q_1\Phi\beta \mid Y,\omega\right] = (1-\omega)Q_1Y$$
$$\mathrm{Var}\left[Q_1\Phi\beta \mid Y,\omega\right] = \sigma^2(1-\omega)Q_1,$$

which proves that $n\left\|Q_1\Phi\beta - (1-\omega)Q_1Y\right\|_{n,2}^2/\{(1-\omega)\sigma^2\} \mid Y,\omega \sim \chi^2_{k_n-d_0}$. □

Proof of Theorem 6. Let β^* denote the projection of the true F_0 on the basis $\{\phi_j\}_{1\leq j\leq k_n}$, i.e.,

$$\beta^* = \text{argmin}_{\beta\in\mathbb{R}^{k_n}} \left\|F_0 - \Phi\beta\right\|_{2,n}. \tag{A.30}$$

We shall treat β^* as the *pseudo-true* parameter and study the posterior concentration of $\Phi\beta$ in the posterior around $\Phi\beta^*$.

To prove Theorem 6, it is sufficient to show that the posterior probability in the equation (4.10) converges in probability to zero. The quantity in (4.10) can be decomposed as follows:

$$\begin{aligned} & P\left[\left\|\Phi\beta - F_0\right\|_{n,2} > M_n^{1/2} \mid Y\right] \\ \leq\ & P\left[\left\|\Phi\beta - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/2 \mid Y\right] + \mathbf{1}\left[\left\|\Phi\beta^* - F_0\right\|_{n,2} > M_n^{1/2}/2\right], \end{aligned}$$

where β^* is defined in (A.30) and $\mathbf{1}(\cdot)$ is the indicator function. The second term on the right-hand side of this expression is always zero when $F_0 \in \mathfrak{L}(\Phi_0)$, since we assume that the column space of Φ_0 is contained in the column space of Φ, and its expectation with respect to the true density is asymptotically zero when $F_0^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F_0 \asymp n$ from (4.9). Therefore, we focus on the first term on the right-hand side. Since $\Phi\beta = Q_1\Phi\beta + Q_0\Phi\beta$, by Lemma 1. the first term can be decomposed as

$$\begin{aligned} & P\left[\left\|\Phi\beta - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/2 \mid Y\right] = E_{\omega|Y}\left[P\left(\left\|\Phi\beta - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/2 \mid Y,\omega\right)\right] \\ \leq\ & \mathbb{E}_{\omega|Y}\left[P\left(\left\|\Phi\beta - \Phi\widetilde{\beta}_\omega\right\|_{n,2} > M_n^{1/2}/4 \mid Y,\omega\right)\right] \\ & +\mathbb{E}_{\omega|Y}\left[P\left(\left\|\Phi\widetilde{\beta}_\omega - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/4 \mid Y,\omega\right)\right] \\ \leq\ & \mathbb{E}_{\omega|Y}\left[P\left(\left\|\mathrm{Q}_1\Phi\beta - (1-\omega)\mathrm{Q}_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y,\omega\right)\right] \\ & +\mathbb{E}_{\omega|Y}\left[P\left(\left\|Q_1\Phi\beta^* - (1-\omega)Q_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y,\omega\right)\right] \\ & +\mathbb{E}_{\omega|Y}\left[P\left(\left\|Q_0\Phi\beta - Q_0 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y,\omega\right)\right] \\ & +\mathbf{1}\left[\left\|Q_0\Phi\beta^* - Q_0 Y\right\|_{n,2} > M_n^{1/2}/8\right], \end{aligned}$$

where $\Phi\widetilde{\beta}_\omega = (1-\omega)\mathrm{Q}_\Phi Y + \omega \mathrm{Q}_0 Y = (1-\omega)\mathrm{Q}_1 Y + \mathrm{Q}_0 Y$.

We denote

$$\begin{aligned} W_1 &= P\left(\left\|\mathrm{Q}_1\Phi\beta - (1-\omega)\mathrm{Q}_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right), \\ W_2 &= P\left(\left\|Q_1\Phi\beta^* - (1-\omega)Q_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right), \\ W_3 &= P\left(\left\|Q_0\Phi\beta - Q_0 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right). \end{aligned}$$

The indicator function in the fourth term converges to zero in probability, since $\left\|Q_0 Y - Q_0\Phi\beta^*\right\|_{2,n}^2$ achieves the parametric optimal rate. To complete the proof we show that the expectations of W_1, W_2, and W_3 with respect to the marginal posterior distribution of ω converge to zero in probability.

First consider W_3. Since $n\left\|\mathrm{Q}_0\Phi\beta - \mathrm{Q}_0 Y\right\|_{2,n}^2/\sigma^2 \mid Y, \omega \sim \chi^2_{d_0}$ by Lemma A.2.2, by using Lemma A.1.4 it follows that

$$\begin{aligned} & E_{\omega|Y}\left[W_3\right] = E_{\omega|Y}\left[P\left\{\left\|\mathrm{Q}_0\Phi\beta - \mathrm{Q}_0 Y\right\|_{2,n} > M_n^{1/2}/8 \mid Y, \omega\right\}\right] \\ \leq\ & C\left(\frac{nM_n}{64\sigma d_0}\right)^{d_0/2} \exp\{-nM_n/(128\sigma^2)\}, \end{aligned}$$

for some constant C.

The last quantity converges to zero as n tends to ∞, which implies that $\mathbb{E}_{\omega|Y}[W_3] = o_p(1)$. Now we obtain the bounds on W_1. By Lemma A.2.2 $n\left\|\mathrm{Q}_1\Phi\beta - (1-\omega)\mathrm{Q}_1 Y\right\|_{2,n}^2/\{(1-\omega)\sigma^2\} \mid Y \sim \chi^2_{k_n - d_0}$. By using Lemma A.1.4, it follows that

$$\begin{aligned} W_1 \leq\ & \left[\frac{nM_n}{64\sigma^2(k_n - d_0)}(1-\omega)^{-1}\right]^{\frac{k_n - d_0}{2}} \exp\left\{\frac{k_n - d_0}{2} - \frac{nM_n}{128\sigma^2}(1-\omega)^{-1}\right\} \\ & \times \mathbf{1}\left[\frac{nM_n}{64\sigma^2}(1-\omega)^{-1} > k_n - d_0\right] + \mathbf{1}\left[\frac{nM_n}{64\sigma^2}(1-\omega)^{-1} \leq k_n - d_0\right]. \end{aligned}$$

We denote the two terms in this expression as $W_{1,1}$ and $W_{1,2}$.

By using Lemma 2 and defining $\widehat{\omega} = (k_n - d_0)/\{nM_n/(64\sigma^2) + k_n - d_0\}$, it follows that

$$
\begin{aligned}
& \mathbb{E}_{\omega|Y}\left[W_{1,1}\right] \\
= \; & \frac{1}{m(Y)}\left[\frac{nM_n \exp\{1\}}{64\sigma^2(k_n-d_0)}\right]^{\frac{k_n-d_0}{2}} \int_{m_n}^{1} \omega^{a+\frac{k_n-d_0}{2}-1}(1-\omega)^{b-\frac{k_n-d_0}{2}-1} \\
& \times \exp\left\{-\frac{nM_n}{128\sigma^2}(1-\omega)^{-1} - H_n\omega\right\} d\omega \\
\leq \; & \frac{1}{m(Y)}\left[\frac{nM_n \exp\{1\}}{64\sigma^2(k_n-d_0)}\right]^{\frac{k_n-d_0}{2}} \int_{m_n}^{1} \omega^{a-1}(1-\omega)^{b-1}\exp\left\{-H_n\omega\right\} d\omega \\
& \times \widehat{\omega}^{\frac{k_n-d_0}{2}}(1-\widehat{\omega})^{-\frac{k_n-d_0}{2}} \exp\left\{-\frac{nM_n}{128\sigma^2}(1-\widehat{\omega})^{-1}\right\} \\
= \; & \frac{1}{m(Y)}\exp\left\{-\frac{nM_n}{128\sigma^2}\right\} \int_{m_n}^{1} \omega^{a-1}(1-\omega)^{b-1}\exp\left\{-H_n\omega\right\} d\omega, \qquad \text{(A.31)}
\end{aligned}
$$

where $m_n = \max[0, 1 - nM_n/\{16\sigma^2(k_n - d_0)\}]$.

Also,

$$
\begin{aligned}
& \mathbb{E}_{\omega|Y}[W_{1,2}] = P_{\omega|Y}\left[\omega < 1 - \frac{nM_n}{64\sigma^2(k_n-d_0)}\right] \\
= \; & \frac{1}{m(Y)} \int_0^{1-\frac{nM_n}{64\sigma^2(k_n-d_0)}} \omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}\exp\{-H_n\omega\}d\omega \\
\leq \; & \frac{1}{m(Y)}\left(\frac{nM_n}{64\sigma^2(k_n-d_0)}\right)^{b-1} \int_0^1 \omega^{a+(k_n-d_0)/2-1}\exp\{-H_n\omega\}d\omega \\
\leq \; & \left(\frac{nM_n}{64\sigma^2(k_n-d_0)}\right)^{b-1} \frac{\Gamma(a+b+(k_n-d_0)/2)}{\Gamma(a+(k_n-d_0)/2)\Gamma(b)} H_n^{-1}\mathbf{1}\left(1 - \frac{nM_n}{64\sigma^2(k_n-d_0)} \geq 0\right) \\
& \times \exp\{H_n\}\left[1 + \frac{bH_n}{a+b+(k_n-d_0)/2} + D\frac{b(b+T_n)^{-a-(k_n-d_0)/2}}{(a+b+(k_n-d_0)/2)^{3/2}}\right. \\
& \left. \times \left(\exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2}\right)_+\right]^{-1}, \qquad \text{(A.32)}
\end{aligned}
$$

where $T_n = \max\{(a + (k_n - d_0)/2)^2, 3\lceil H_n\rceil\}$ and D is some constant.

We now consider two cases: (i) when $F_0 \in \mathfrak{L}(\Phi_0)$ and (ii) when $F_0^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F_0 \asymp n$.

Case (i) $F_0 \in \mathfrak{L}(\Phi_0)$:

Recall that in this case $M_n = \zeta_n n^{-1}$ for any arbitrary diverging sequence ζ_n. First, we show that $\mathbb{E}_{\omega|Y}[W_1] \xrightarrow{p} 0$ by proving that $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{p} 0$ and $\mathbb{E}_{\omega|Y}[W_{1,2}] \xrightarrow{p} 0$.

Applying Lemma 2, it follows that (A.31) is bounded above by

$$\begin{aligned} \mathbb{E}_{\omega|Y}\left[W_{1,1}\right] &\leq \frac{C\exp\left\{-nM_n/(128\sigma^2)\right\}\left(1+\frac{b}{a+b}\exp\{H_n\}\right)}{1+\delta_n+u_n\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+} \\ &\leq C\exp\left\{-\frac{nM_n}{128\sigma^2}\right\}\left(1+\frac{b}{a+b}\exp\{H_n\}\right), \end{aligned} \tag{A.33}$$

where $\delta_n = bH_n/(a+b+(k_n-d_0)/2)$ and $u_n = (a+b)(b+T_n)^{-a_n-(k_n-d_0)/2}/(a+b+(k_n-d_0)/2)^{3/2}$ with $T_n = \max\{(a+(k_n-d_0)/2)^2, 3\lceil H_n\rceil\}$, and C and D are some constants.

Since $2H_n \sim \chi^2_{k_n-d_0}$, by Lemma A.1.4 and defining $q_n = k_n^{-1/2}(\log k_n)^{1/2}(-\log b)^{1/2}$, it follows that

$$P\left[H_n > k_n q_n/2\right] \leq \exp\{-ck_nq_n\}, \tag{A.34}$$

for some constant c. Hence, by the condition that $k_n \log k_n \prec -\log b$, it is clear that $b\exp\{H_n\} = o_p(1)$, which shows that $\mathbb{E}_{\omega|Y}[W_{1,1}] = o_p(1)$.

Similarly, since $\Gamma(b)^{-1} \asymp b$, (A.32) is bounded by

$$C'b\exp\{H_n\}\left(\frac{nM_n}{64\sigma^2(k_n-d_0)}\right)^{b-1},$$

for some constant C'. By (A.34), $b\exp\{H_n\} = o_p(1)$, which implies $\mathbb{E}_{\omega|Y}[W_{1,2}] = o_p(1)$.

We next show that $E_{\omega|Y}[W_2]$ converges in probability to zero. Applying Lemma 2, it fol-

lows that

$$
\begin{aligned}
& \mathbb{E}_{\omega|Y}[W_2] = \mathbb{E}_{\omega|Y}\left[P\left[\left\|(1-\omega)Q_1 Y - Q_1\Phi\beta^*\right\|_{n,2} > M_n^{1/2}/8 \mid Y,\omega\right]\right] \\
= \;& P_{\omega|Y}\left[\omega < 1 - \left(\frac{nM_n}{64\sigma^2 H_n}\right)^{1/2}\right] \\
= \;& \frac{1}{m(Y)}\int_0^{1-\left(\frac{nM_n}{128\sigma^2 H_n}\right)^{1/2}} \omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}\exp\{-H_n\omega\}d\omega \\
\leq \;& \mathbf{1}\left\{1 - \left(\frac{nM_n}{128\sigma^2 H_n}\right)^{1/2} \geq 0\right\}\frac{1}{m(Y)}\left(\frac{nM_n}{64\sigma^2 H_n}\right)^{(b-1)/2} \\
& \times \int_0^1 \omega^{a+(k_n-d_0)/2-1}\exp\{-H_n\omega\}d\omega \\
\leq \;& \mathbf{1}\left\{1 - \left(\frac{nM_n}{128\sigma^2 H_n}\right)^{1/2} \geq 0\right\}\frac{\Gamma(a+b+(k_n-d_0)/2)}{\Gamma(b)\Gamma(a+(k_n-d_0)/2)}\left(\frac{nM_n}{128\sigma^2 H_n}\right)^{(b-1)/2} \\
& \times \exp\{H_n\}\left\{1+\delta_n+u_n\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+\right\}^{-1} \\
\leq \;& Cb\left(\frac{nM_n}{128\sigma^2}\right)^{(b-1)/2} H_n^{1/2}\exp\{H_n\},
\end{aligned}
$$

where C is some constant, and δ_n and u_n are defined following (A.33).

From (A.34), it follows that $b\{nM_n/(128\sigma^2)\}^{(b-1)/2}H_n^{1/2}\exp\{H_n\}$ is bounded by $b\{nM_n/(128\sigma^2)\}^{(b-1)/2}(k_nq_n/2)^{1/2}\exp\{k_nq_n/2\}$ with probability greater than $1-\exp\{-ck_nq_n\}$ from which it follows that $\mathbb{E}_{\omega|Y}[W_2] = o_p(1)$.

Case (ii) $F_0^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F_0 \asymp n$:

Recall that in this case $M_n = \zeta_n n^{-2\alpha/(1+2\alpha)}\log n$ for any arbitrary diverging sequence ζ_n,

and δ_n and u_n are defined following (A.33). From (A.31) it follows that

$$\begin{aligned}
&\mathbb{E}_{\omega|Y}\left[W_{1,1}\right] \leq \frac{1}{m(Y)} \exp\left\{-\frac{nM_n}{128\sigma^2}\right\} \int_{m_n}^{1} \omega^{a-1}(1-\omega)^{b-1} \exp\left\{-H_n\omega\right\} d\omega \\
\leq\ & C \exp\left\{-\frac{nM_n}{128\sigma^2}\right\} \frac{1+\frac{b}{a+b}\exp\{H_n\}}{1+\delta_n+u_n\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+},
\end{aligned}$$

for some constant C.

By Lemma A.1.4, for any sequence $w_n \to 0$, H_n is larger than $w_n F_0^{\mathrm{T}} Q_1 F_0/\sigma^2$ with probability greater than $1-\exp\{-cF_0^{\mathrm{T}}Q_1F_0(1-w_n)^2/\sigma^2\}$ for some constant c. Since $F_0^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F_0 \asymp n$ implies $F_0^{\mathrm{T}}Q_1F_0 \asymp n$, the last line in the above display can be expressed as

$$C' \exp\left\{-\frac{nM_n}{128\sigma^2}(k_n-d_0)^{3/2}(b+T_n)^{(k_n-d_0)/2}\right\} + o_p(1),$$

where $T_n = \max\{(a+(k_n-d_0)/2)^2, 3H_n\}$ and C' is some positive constant. Therefore, to show $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{p} 0$, it is sufficient to prove that $T_n^{(k_n-d_0)/2}\exp\{-nM_n/(128\sigma^2)\} = o_p(1)$. For any $\epsilon > 0$,

$$\begin{aligned}
&P\left[T_n^{(k_n-d_0)/2}\exp\left\{-\frac{nM_n}{128\sigma^2}\right\} > \epsilon\right] \\
\leq\ & P\left[(3H_n)^{(k_n-d_0)/2}\exp\left\{-\frac{nM_n}{128\sigma^2}\right\} > \epsilon\right] + P\left[3H_n < (a+(k_n-d_0)/2)^2\right] \\
\leq\ & P\left[\log H_n > \zeta_n \log n\right] + P\left[3H_n < (a+(k_n-d_0)/2)^2\right].
\end{aligned}$$

Since $\zeta_n \to \infty$ as n tends to ∞, from (A.7) in Lemma A.1.4, it follows that the first term in the above display can be bounded above by $\exp\{-c'(n_n^{\zeta} - F_0^{\mathrm{T}}Q_1F_0/\sigma^2)\}$ for some constant c'. Similarly, from (A.6) in Lemma A.1.4, the second term is bounded by $\exp\{-c''F_0^{\mathrm{T}}Q_1F_0/\sigma^2\}$ with some constant c'', which proves that $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{p} 0$.

Since $nM_n \succ k_n$, the indicator function $\mathbf{1}(1-nM_n/(64\sigma^2(k_n-d_0)) \geq 0)$ in (A.32) is zero when n is large enough, which results in $\mathbb{E}_{\omega|Y}[W_{1,2}] \xrightarrow{p} 0$.

The marginal posterior mean of W_2 can be decomposed as

$$\begin{aligned}\mathbb{E}_{\omega|Y}[W_2] &\leq P_{\omega|Y}\left[\left\|(1-\omega)\mathrm{Q}_1 Y - \mathrm{Q}_1 Y\right\|_{n,2} > \frac{1}{16}M_n^{1/2}\right] \\ &\quad + \mathbf{1}\left[\left\|\mathrm{Q}_1 Y - \mathrm{Q}_1\Phi\beta^*\right\|_{n,2} > \frac{1}{16}M_n^{1/2}\right].\end{aligned}$$

Results provided by Zhou et al. (1998) (see equation (4.9) on page 66) show that the second term in the previous expression is $o_p(1)$. The first term can be expressed as

$$\begin{aligned}&P_{\omega|Y}\left[\omega > \left(\frac{nM_n}{256\sigma^2 H_n}\right)^{1/2}\right] \\ &= \frac{1}{m(Y)}\int_{\left(\frac{nM_n}{256\sigma^2 H_n}\right)^{1/2}}^{1} \omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}\exp\{-H_n\omega\}d\omega \\ \leq\ & \frac{1}{m(Y)}\exp\left\{-H_n^{1/2}\left(nM_n/(256\sigma^2)\right)^{1/2}\right\}\int_0^1 \omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}d\omega \\ \leq\ & \left[u_n\exp\{-H_n\}\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+\right]^{-1} \\ &\times\exp\left\{-H_n^{1/2}\left(nM_n/(256\sigma^2)\right)^{1/2}\right\},\end{aligned}$$

for some positive constant D. Since $H_n/n = O_p(1)$ and $-\log b \prec n^{1/2}k_n^{1/2}$, the above quantity converges in probability to zero, which completes the proof. □

A.3 Nonlocal Functional Priors for Nonparametric Hypothesis Testing and High-dimensional Model Selection

A set of technical results follow that are used in the proof of the main results. For a given model $\mathbf{k}$,

$$\widetilde{\beta}_{\mathbf{k}} = (\Phi_{\mathbf{k}}^{\mathrm{T}}\Phi_{\mathbf{k}} + 1/\tau_n I)^{-1}\Phi_{\mathbf{k}}^{\mathrm{T}}\mathbf{y}, \quad \widetilde{F}_{\mathbf{k}} = \Phi_{\mathbf{k}}\widetilde{\beta}_{\mathbf{k}}, \quad \widetilde{P}_{\mathbf{k}} = \Phi_{\mathbf{k}}(\Phi_{\mathbf{k}}^{\mathrm{T}}\Phi_{\mathbf{k}} + 1/\tau_n I)^{-1}\Phi_{\mathbf{k}}^{\mathrm{T}},$$
$$D_{\mathbf{k}}(\mathbf{y}) = E_{\beta_{\mathbf{k}}|\mathbf{y},\mathbf{k}}\left[\exp\left\{-\sum_{j\in\mathbf{k}}\frac{\sigma^2\tau_n}{\beta_j^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_j}\right\}\right], \tag{A.35}$$

where $\Phi_{\mathbf{k}}$ is defined in the second paragraph of Section 5.5.

For $1 \leq j \leq p$, the subvector of $\widetilde{\beta}_{\mathbf{k}}$ corresponding to the covariate x_j is denoted by $\widetilde{\beta}_{\mathbf{k},j}$, and define

$$\widetilde{\beta}_j = (\Phi_j^{\mathrm{T}}\Phi_j + 1/\tau_n I)^{-1}\Phi_j^{\mathrm{T}}\mathbf{y}, \quad \widetilde{F}_j = \Phi_j\widetilde{\beta}_j,$$
$$\widetilde{F}_{\mathbf{k},j} = \Phi_j\widetilde{\beta}_{\mathbf{k},j}, \quad \text{and} \quad \widetilde{P}_j = \Phi_j(\Phi_j^{\mathrm{T}}\Phi_j + 1/\tau_n I)^{-1}\Phi_j^{\mathrm{T}}, \tag{A.36}$$

where Φ_j is defined in the second paragraph of Section 5.5. Similarly, we define $\widehat{\beta}_{\mathbf{k},j}$ as the subvector of $\widehat{\beta}_{\mathbf{k}}$ defined in (5.14) corresponding the covariate x_j for $j \in \mathbf{k}$, and $\widehat{F}_{\mathbf{k},j} = \Phi_j\widehat{\beta}_{\mathbf{k},j}$.

Recall that P_0 denotes the probability measure that generates data $\mathbf{y}$.

For univariate settings, we simply denote the basis matrix by Φ and the corresponding coefficients by $\beta \in \mathbb{R}^{K_n}$. The ridge solution of β is defined by $\widetilde{\beta} = (\Phi^{\mathrm{T}}\Phi + 1/\tau_n I)^{-1}\Phi^{\mathrm{T}}\mathbf{y}$.

Lemma A.3.1. *Suppose* $\beta^* \mid \mathbf{y} \sim N\left(\widetilde{\beta}, \sigma_n^{*2}(\Phi^{\mathrm{T}}\Phi + 1/\tau_n I)^{-1}\right)$ *for some arbitrary* $\widetilde{\beta} \in \mathbb{R}^{K_n}$

and $\sigma_n^{*2} > 0$. *Let* $\widetilde{d}_n = \widetilde{F}^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\widetilde{F}$, *where* $\widetilde{F} = \Phi\widetilde{\beta}$. *Suppose* $K_n \prec \tau_n n^{-1/2}/\log n$. *Then,*

$$\begin{aligned} & E_{\beta^*|\mathbf{y}}\left[\exp\left\{-\sigma^2\tau_n\left(\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^*\right)^{-1}\right\}\right] \\ \leq\ & \exp\left\{-\sigma^2\left\{\frac{c_1\sigma_n^{*2}}{n^{1/2}} + \widetilde{d}_n/\tau_n + c_2\sigma_n^*(n\widetilde{d}_n)^{1/2}\tau_n^{-1}\right\}^{-1}\right\} \\ & + \exp\{-c_3 n^{-1/2}\tau_n\} + \exp\{-c_4 n\}, \end{aligned} \tag{A.37}$$

and

$$\begin{aligned} & E_{\beta^*|\mathbf{y}}\left[\exp\left\{-\sigma^2\tau_n\left(\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^*\right)^{-1}\right\}\right] \\ \geq\ & \exp\left\{-\tau_n\widetilde{d}_n^{-1}\log n\right\}\left(1 - c_5\widetilde{d}_n^{-1}\exp\left\{-\widetilde{d}_n\left(1-1/\log n\right)^2\right\}\right), \end{aligned} \tag{A.38}$$

for some positive constants c_i *for* $i = 1, \ldots, 5$.

Proof. First, we shall show the upper bound (A.37). Since $\exp\{-\sigma^2\tau_n\left(\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^*\right)^{-1}\} \leq 1$ for any $\beta^* \in \mathbb{R}^{K_n}$ and $\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^* \leq n\|\Phi\beta^* - \widetilde{F}\|_{n,2}^2 + 2|\widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)(\Phi\beta^* - \widetilde{F})| + \widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\widetilde{F}$, it follows that

$$\begin{aligned} & E_{\beta^*|\mathbf{y}}\left[\exp\left\{-\sigma^2\tau_n\left(\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^*\right)^{-1}\right\}\right] \\ \leq\ & E_{\beta^*|\mathbf{y}}\left[\exp\left\{-\sigma^2\tau_n\left(n\|\Phi\beta^* - \widetilde{F}\|_{n,2}^2 + 2|\widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)(\Phi\beta^* - \widetilde{F})| + \widetilde{d}_n\right)^{-1}\right\}\right] \\ \leq\ & \exp\left\{-\sigma^2\left\{\frac{c_1\sigma_n^{*2}}{n^{1/2}} + \widetilde{d}_n/\tau_n + c_2\sigma_n^*(n\widetilde{d}_n)^{1/2}\tau_n^{-1}\right\}^{-1}\right\} \\ & + P_{\beta^*|\mathbf{y}}\left[n\|\Phi\beta^* - \widetilde{F}\|_{n,2}^2 > c_1\sigma_n^{*2}\tau_n n^{-1/2}\right] \\ & + P_{\beta^*|\mathbf{y}}\left[|\widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)(\Phi\beta^* - \widetilde{F})| > c_2\sigma_n^*(n\widetilde{d}_n)^{1/2}/2\right], \end{aligned}$$

for some constants c_1 and c_2.

Since $n\|\Phi\beta^* - \widetilde{F}\|_{n,2}^2/\sigma_n^{*2} = (\beta^* - \widetilde{\beta})^{\mathrm{T}}\Phi^{\mathrm{T}}\Phi(\beta^* - \widetilde{\beta}) \leq (\beta^* - \widetilde{\beta})^{\mathrm{T}}(\Phi^{\mathrm{T}}\Phi + 1/\tau_n I)(\beta^* - \widetilde{\beta})$, $(\beta^* - \widetilde{\beta})^{\mathrm{T}}(\Phi^{\mathrm{T}}\Phi + 1/\tau_n I)(\beta^* - \widetilde{\beta}) \mid \mathbf{y} \sim \chi^2_{K_n}$, and $K_n \prec \tau_n n^{-1/2}/\log n$, Lemma A.1.4 implies

that

$$\begin{aligned} & P_{\beta^*|\mathbf{y}}\left[n\|\Phi\beta^* - \widetilde{F}\|_{n,2}^2 > c_1\sigma_n^{*2}\tau_n n^{-1/2}\right] \\ \leq \;& \exp\{-c_3 n^{-1/2}\tau_n\}, \end{aligned}$$

for some constant c_3.

Since $\widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)(\Phi\beta^* - \widetilde{F}) \mid \mathbf{y} \sim N(0, \sigma_n^{*2}\widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\widetilde{Q}_\Phi(\mathrm{I}-\mathrm{Q}_0)\widetilde{F})$, where $\widetilde{Q}_\Phi = \Phi(\Phi^{\mathrm{T}}\Phi + 1/\tau_n I)^{-1}\Phi^{\mathrm{T}}$, and $\widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\widetilde{Q}_\Phi(\mathrm{I}-\mathrm{Q}_0)\widetilde{F} \leq \widetilde{d}_n$, it follows that

$$\begin{aligned} & P_{\beta^*|\mathbf{y}}\left[|\widetilde{F}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)(\Phi\beta^* - \widetilde{F})| > c_2\sigma_n^*(n\widetilde{d}_n)^{1/2}/2\right] \\ \leq \;& \exp\{-c_4 n\}, \end{aligned}$$

for some constant c_4, by the fact that for z , $P(|Z| > z) \leq (2\pi)^{-1/2}z^{-1}\exp\{-z^2/2\}$, where Z follows a standard Gaussian distribution.

Second, we consider the lower bound (A.38). By Markov's inequality, it follows that

$$\begin{aligned} & E_{\beta^*|\mathbf{y}}\left[\exp\left\{-\sigma^2\tau_n\left(\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^*\right)^{-1}\right\}\right] \\ \geq \;& \exp\{-\sigma^2\tau_n\widetilde{d}_n^{-1}\log n\} \\ & \times P_{\beta^*|\mathbf{y}}\left[\exp\left\{-\sigma^2\tau_n\left(\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^*\right)^{-1}\right\} > \exp\left\{-\sigma^2\tau_n\widetilde{d}_n^{-1}\log n\right\}\right] \\ \geq \;& \exp\{-\sigma^2\tau_n\widetilde{d}_n^{-1}\log n\}\left(1 - P_{\beta^*|\mathbf{y}}\left[\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^* < \widetilde{d}_n/\log n\right]\right). \end{aligned}$$

Since $\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^*/\sigma^{*2} \mid \mathbf{y} \sim \chi^2_{K_n-d_0}(\widetilde{F}(\mathrm{I}-\mathrm{Q}_0)\widetilde{F}/\sigma^{*2})$, by Lemma A.1.4, it follows that

$$P_{\beta^*|\mathbf{y}}\left[\beta^{*\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^* < \widetilde{d}_n/\log n\right] \leq c_5\widetilde{d}_n^{-1}\exp\left\{-\widetilde{d}_n(1-1/\log n)^2\right\},$$

for some constant c_5. $\square$

Lemma A.3.2. *Define*

$$t_{n,j} = \begin{cases} q_n F_0^{\mathrm{T}} P_j F_0, & \text{if } j \in \mathbf{t}, \\ u_n(\log p + K_n + \zeta_{n*}^{-1}), & \text{if } j \notin \mathbf{t}, \end{cases}$$

where $u_n = q_n^2(\log n)^2$.

Then, for an arbitrary $\delta > 0$,

$$P_0\left[\max_{\mathbf{k}:|\mathbf{k}|\le q_n} \max_{j\in\mathbf{k}} \widetilde{F}_{\mathbf{k},j}^{\mathrm{T}} \widetilde{F}_{\mathbf{k},j}/\sigma^2 > t_{n,j}\right] \le p^{-|\mathbf{k}|(1+\delta)}, \tag{A.39}$$

for large enough n. *Also,*

$$P_0\left[\min_{j\in\mathbf{t}} \widetilde{F}_{\mathbf{t},j}^{\mathrm{T}} \widetilde{F}_{\mathbf{t},j}/\sigma^2 < F_0^{\mathrm{T}} P_j F_0/(\sigma^2 \log n)\right] \le |\mathbf{t}| \exp\{-cn\}, \tag{A.40}$$

for some constant c.

Proof. We note that $\widehat{F}_{\mathbf{k},j}^{\mathrm{T}} \widehat{F}_{\mathbf{k},j}^{\mathrm{T}} = A^{\mathrm{T}}\Phi_j P_{\mathbf{k}\setminus\{j\}} \Phi_j^{\mathrm{T}} A + A^{\mathrm{T}}\Phi_j (I - P_{\mathbf{k}\setminus\{j\}})\Phi_j^{\mathrm{T}} A$,
where $A = (\Phi_j^{*\mathrm{T}}\Phi_j^*)^{-1}\Phi_j^{*\mathrm{T}}\mathbf{y}$ and Φ_j^* is a full column-rank matrix of $(I - P_{\mathbf{k}\setminus\{j\}})\Phi_j$ created by the Gram-Schmidt procedure. Since $A^{\mathrm{T}}\Phi_j(I - P_{\mathbf{k}\setminus\{j\}})\Phi_j^{\mathrm{T}} A = \mathbf{y}^{\mathrm{T}} P_j^* \mathbf{y}$, where P_j^* is a projection matrix of Φ_j^*, (A4) implies that

$$\mathbf{y}^{\mathrm{T}} P_j^* \mathbf{y} \le \widehat{F}_{\mathbf{k},j}^{\mathrm{T}} \widehat{F}_{\mathbf{k},j}^{\mathrm{T}} \le 2\mathbf{y}^{\mathrm{T}} P_j^* \mathbf{y}.$$

Since $\mathbf{y}^{\mathrm{T}} P_j^* \mathbf{y}/\sigma^2 \sim \chi^2_{K_n}(F_0^{\mathrm{T}} P_j^* F_0/\sigma^2)$, $F_0^{\mathrm{T}} P_j F_0 \prec \log p$ for $j \notin \mathbf{t}$ by (A5), and $F_0^{\mathrm{T}} P_j^* F_0 \le$

$F_0^{\mathrm{T}} P_j F_0$, Lemma A.1.4 and (A5) imply that

$$
\begin{aligned}
& P_0\left[\max_{\mathbf{k}:|\mathbf{k}|\le q_n}\max_{j\in\mathbf{k}} \widetilde{F}_{\mathbf{k},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{k},j}/\sigma^2 > t_{n,j}\right] \le \sum_{\mathbf{k}:|\mathbf{k}|\le q_n}\sum_{j\in\mathbf{k}} P_0\left[\widehat{F}_{\mathbf{k},j}^{\mathrm{T}}\widehat{F}_{\mathbf{k},j}/\sigma^2 > t_{n,j}\right] \\
\le & \sum_{\mathbf{k}:|\mathbf{k}|\le q_n}\sum_{j\in\mathbf{k}}\left[\left(\frac{t_{n,j}}{2K_n}\right)^{K_n/2}\exp\{K_n/2 - c_1 t_{n,j}\} + c_2(F_0^{\mathrm{T}}P_j^* F_0/\sigma^2)^{1/2} t_{n,j}^{-1}\right. \\
& \left.\times\exp\{-c_3 t_{n,j}^2/F_0^{\mathrm{T}}P_j^* F_0\}\right] \\
\le & \sum_{\mathbf{k}:|\mathbf{k}|\le q_n}\sum_{j\in\mathbf{k}\cap\mathbf{t}}\exp\{-c_4 q_n n\} + \sum_{\mathbf{k}:|\mathbf{k}|\le q_n}\sum_{j\in\mathbf{k}\setminus\mathbf{t}}\exp\left\{-c_5\frac{u_n^2(\log p + K_n + \zeta_{n*}^{-1})^2}{q_n\log p}\right\} \\
\le & p^{-|\mathbf{k}|(1+\delta)},
\end{aligned}
$$

for some positive constants c_i for $i = 1, \ldots, 5$.

Similarly, by Lemma A.1.4, it follows that

$$
\begin{aligned}
& P_0\left[\min_{j\in\mathbf{t}}\widetilde{F}_{\mathbf{t},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{t},j}/\sigma^2 < F_0^{\mathrm{T}}P_j F_0/(\sigma^2\log n)\right] \\
\le & \sum_{j\in\mathbf{t}} P_0\left[\widetilde{F}_{\mathbf{t},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{t},j}/\sigma^2 < F_0^{\mathrm{T}}P_j F_0/(\sigma^2\log n)\right] \\
\le & |\mathbf{t}|\exp\{-cn\},
\end{aligned}
$$

for some constant c.

□

Lemma A.3.3. *Recall that $D_{\mathbf{k}}(\mathbf{y})$ is defined in* (A.35)*. Assume that (A2)–(A3) hold. Let $u_n = q_n^2(\log n)^2$. For $\mathbf{k} \neq \mathbf{t}$ and a given $\delta > 0$, there exist some positive constant c and c' such that*

$$
P_0\left[\frac{D_{\mathbf{k}}(\mathbf{y})}{D_{\mathbf{t}}(\mathbf{y})} > \exp\left\{-\frac{c|\mathbf{k}|\tau_n u_n^{-1/2}\zeta_{n*}^{1/2}}{(K_n + \log p + +\zeta_{n*}^{-1})^{1/2} n^{1/2}} + c'|\mathbf{t}|\tau_n\log/n + z_n\right\}\right] \le p^{-|\mathbf{k}|(1+\delta)},
$$

where $z_n = K_n\log(\zeta_n^*/\zeta_{n*})/2$.

Proof. We note that

$$\begin{aligned}
\frac{D_{\mathbf{k}}(\mathbf{y})}{D_{\mathbf{t}}(\mathbf{y})} &= \frac{E_{\beta_{\mathbf{k}}|\mathbf{y},\mathbf{k}}\left[\exp\left\{-\sum_{j\in\mathbf{k}}\frac{\sigma^2\tau_n}{\beta_j^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_j}\right\}\right]}{E_{\beta_{\mathbf{t}}|\mathbf{y},\mathbf{t}}\left[\exp\left\{-\sum_{j\in\mathbf{t}}\frac{\sigma^2\tau_n}{\beta_j^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_j}\right\}\right]} \\
&\leq \prod_{j\in\mathbf{k}}\zeta_*^{-K_n/2}E_{\beta_{*j}|\mathbf{y},\mathbf{k}}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta_{*j}^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_{*j}}\right\}\right] \\
&\quad\times\prod_{j\in\mathbf{t}}\zeta^{*K_n/2}E_{\beta_j^*|\mathbf{y},\mathbf{t}}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta_j^{*\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_j^*}\right\}\right]^{-1}, \qquad \text{(A.41)}
\end{aligned}$$

where $\beta_j^* \mid \mathbf{y},\mathbf{k} \sim N(\widetilde{\beta}_{\mathbf{k},j}, \sigma^2\zeta_n^{*-1}(\Phi_j^{\mathrm{T}}\Phi_j + 1/\tau_n I)^{-1})$ and $\beta_{*j} \mid \mathbf{y},\mathbf{k} \sim N(\widetilde{\beta}_{\mathbf{k},j}, \sigma^2\zeta_{n*}^{-1}(\Phi_j^{\mathrm{T}}\Phi_j + 1/\tau_n I)^{-1})$ for $j \in \mathbf{k}$.

By using the upper bound and the lower bound provided in Lemma A.3.1, it follows that

$$\begin{aligned}
&E_{\beta_{*j}|\mathbf{y},\mathbf{k}}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta_{*j}^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_{*j}}\right\}\right] \\
\leq\ &\exp\left\{-\sigma^2\left\{\frac{c_1\sigma^2}{\zeta_{n*}n^{1/2}} + \widetilde{F}_{\mathbf{k},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{k},j}/\tau_n + c_2\sigma\zeta_{n*}^{-1/2}\left(n\widetilde{F}_{\mathbf{k},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{k},j}\right)^{1/2}\tau_n^{-1}\right\}^{-1}\right\} \\
&+\exp\{-c_3 n\} + \exp\{-c_4 n^{-1/2}\tau_n\},
\end{aligned}$$

for $j \in \mathbf{k}$ and some constants c_i for $i = 1,\ldots,4$. Also, by (A.38) in Lemma A.3.1, it follows that

$$\begin{aligned}
&E_{\beta_j^*|\mathbf{y},\mathbf{t}}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta_j^{*\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_j^*}\right\}\right] \\
\geq\ &\exp\left\{-\frac{\sigma^2\tau_n\log n}{\widetilde{F}_{\mathbf{k},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{k},j}}\right\}\left(1 - c_5\left(\widetilde{F}_{\mathbf{k},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{k},j}\right)^{-1}\exp\left\{-\widetilde{F}_{\mathbf{k},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{j}}(1-1/\log n)^2\right\}\right),
\end{aligned}$$

for $j \in \mathbf{t}$ and some constant c_5.

Plugging the bounds in (A.41), it follows that

$$
\begin{aligned}
&\frac{D_{\mathbf{k}}(\mathbf{y})}{D_{\mathbf{t}}(\mathbf{y})} \\
\leq\ & \prod_{j\in\mathbf{k}}\Bigg[\exp\left\{-\sigma^2\left\{\frac{c_1\sigma^2}{\zeta_{n*}n^{1/2}}+\widetilde{d}_{\mathbf{k},j}/\tau_n+c_2\sigma\zeta_{n*}^{-1/2}\left(n\widetilde{d}_{\mathbf{k},j}\right)^{1/2}\tau_n^{-1}\right\}^{-1}\right\} \\
&+\exp\{-c_3 n\}+\exp\{-c_4 n^{-1/2}\tau_n\}\Bigg]\times\frac{\zeta_{n*}^{-|\mathbf{k}|K_n/2}}{\zeta_n^{*-|\mathbf{t}|K_n/2}} \\
&\times\prod_{j\in\mathbf{t}}\left[\exp\left\{\frac{\sigma^2\tau_n\log n}{\widetilde{d}_{\mathbf{k},j}}\right\}\left(1-c_5\widetilde{d}_{\mathbf{k},j}^{-1}\exp\left\{-\widetilde{d}_{\mathbf{k},j}(1-1/\log n)^2\right\}\right)^{-1}\right],
\end{aligned}
\tag{A.42}
$$

where $\widetilde{d}_{\mathbf{k},j}=\widetilde{F}_{\mathbf{k},j}^{\mathrm{T}}\widetilde{F}_{\mathbf{k},j}$.

Since $F_0^{\mathrm{T}}P_jF_0\prec u_n(K_n+\log p+\zeta_{n*}^{-1})\prec n$ for $j\in\mathbf{k}\setminus\mathbf{t}$, by Lemma A.3.2, it follows that

$$
\begin{aligned}
&D_{\mathbf{k}}(\mathbf{y})/D_{\mathbf{t}}(\mathbf{y}) \\
\leq\ & \exp\Bigg\{-c_6|\mathbf{k}|\sigma^2\tau_n\Big\{n^{-1/2}\zeta_{n*}^{-1}+u_n(K_n+\log p+\zeta_{n*}^{-1}) \\
&+u_n^{1/2}(K_n+\log p+\zeta_{n*}^{-1})^{1/2}n^{1/2}\zeta_{n*}^{-1/2}\Big\}^{-1}\Bigg\}\exp\{c_7|\mathbf{t}|\sigma^2\tau_n\log n/n+z_n\} \\
\leq\ & \exp\left\{-\frac{c_8|\mathbf{k}|\sigma^2\tau_n u_n^{-1/2}\zeta_{n*}^{1/2}}{(K_n+\log p+\zeta_{n*}^{-1})^{1/2}n^{1/2}}+c_7|\mathbf{t}|\sigma^2\tau_n\log n/n+z_n\right\},
\end{aligned}
$$

with probability greater than $1-p^{-|\mathbf{k}|(1+\delta)}$ for some constants c_6, c_7, and c_8. □

Proof of Theorem 7. Since

$$
\log\left\{\frac{m_1^L(\mathbf{y})}{m_0(\mathbf{y})}\right\}=-\frac{K_n-d_0}{2}\log(1+g_n)-\frac{g_n}{2\sigma^2(1+g_n)}\mathbf{y}^{\mathrm{T}}(Q_\Phi-Q_0)\mathbf{y},
$$

it is sufficient to show that for any diverging sequence $v_n \to \infty$,

$$\begin{aligned} & P_0\Big(\left|\mathbf{y}^{\mathrm{T}}(Q_\Phi - Q_0)\mathbf{y}/\sigma^2 - \left\{F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0/\sigma^2 + K_n - d_0\right\}\right| \\ & > 2\{2F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0/\sigma^2 + K_n - d_0\}^{1/2} v_n \Big) \\ = \; & o(1). \end{aligned} \tag{A.43}$$

We note that Birgé (2001) showed the following statements:

When $W \sim \chi^2_{m_n}(\lambda_n)$, for all $t > 0$,

$$\begin{aligned} P\left[W \geq m_n + \lambda_n + 2\left\{(2\lambda_n + m_n)t\right\}^{1/2} + 2t\right] &\leq \exp\{-t\} \\ P\left[W \leq m_n + \lambda_n - 2\left\{(2\lambda_n + m_n)t\right\}^{1/2}\right] &\leq \exp\{-t\}. \end{aligned}$$

These results completes the proof, since $\mathbf{y}^{\mathrm{T}}(Q_\Phi - Q_0)\mathbf{y}/\sigma^2 \sim \chi^2_{K_n - d_0}(F_0^{\mathrm{T}}(Q_\Phi - Q_0)F_0/\sigma^2)$.

□

Proof of Theorem 8. It is well-known that when $Z \sim N(m, V)$, $E(Z^{\mathrm{T}}WZ) = tr(WV) + m^{\mathrm{T}}Wm$ and $Var(Z^{\mathrm{T}}WZ) = 2tr\{(WV)^2\} + 4m^{\mathrm{T}}WVWm$ for some symmetric matrix W. So, since the posterior distribution of β from the local prior follows $N\{g_n\widehat{\beta}/(1+g_n), \sigma^2 g_n(\Phi^{\mathrm{T}}\Phi)^{-1}/(1+g_n)\}$, where $\widehat{\beta} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\mathbf{y}$, it follows that

$$\begin{aligned} D_n(h_M^{r=1}; \mathbf{y}) &= \frac{E_{\beta|\mathbf{y}}\left[\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I - Q_0)\Phi\beta\right]}{E_{\pi^L}\left[\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I - Q_0)\Phi\beta\right]} \\ &= (1+g_n)^{-1} + \left(\frac{g_n}{1+g_n}\right)^2 \widehat{F}^{\mathrm{T}}(Q_\Phi - Q_0)\widehat{F}/\left(\sigma^2 g_n(K_n - d_0)\right), \end{aligned}$$

where $\widehat{F} = \Phi\widehat{\beta} = Q_\Phi\mathbf{y}$. Also,

$$D_n(h_M^{r=2}; \mathbf{y}) = \frac{E_{\beta|\mathbf{y}}\left[\{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I - Q_0)\Phi\beta\}^2\right]}{E_{\pi^L}\left[\{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I - Q_0)\Phi\beta\}^2\right]} = \frac{U_1 + U_2}{U_3},$$

where

$$\begin{aligned} U_1 &= \frac{2(\sigma^2 g_n)^2}{(1+g_n)^2}(K_n - d_0)^2 + 4\sigma^2 \left(\frac{g_n}{1+g_n}\right)^3 \widehat{F}^{\mathrm{T}}(Q_\Phi - Q_0)\widehat{F} \\ U_2 &= \left[\frac{\sigma^2 g_n}{1+g_n}(K_n - d_0) + \left(\frac{g_n}{1+g_n}\right)\widehat{F}^{\mathrm{T}}(Q_\Phi - Q_0)\widehat{F}\right]^2 \\ U_3 &= 2(\sigma^2 g_n)^2(K_n - d_0) + \left(\sigma^2 g_n (K_n - d_0)\right)^2. \end{aligned}$$

Since $\widehat{F}^{\mathrm{T}}(Q_\Phi - Q_0)\widehat{F} = \mathbf{y}^{\mathrm{T}}(Q_\Phi - Q_0)\mathbf{y}$, (A.43) completes the proof. □

Proof of Theorem 10. Since the local prior π^L on β follows $N(0, \sigma^2\tau_n(\Phi^T\Phi)^{-1})$, by using Markov's inequality, it follows that for any fixed $\epsilon > 0$,

$$\begin{aligned} E_{\pi^L}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta}\right\}\right] &\geq \epsilon P_{\pi^L}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta}\right\}\right] \\ &\geq \epsilon P_{\pi^L}\left[\frac{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta}{\sigma^2\tau_n} > (-\log\epsilon)^{-1}\right]. \end{aligned}$$

Since $\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta/(\sigma^2\tau_n) \sim \chi^2_{K_n-d_0}$, the right-hand side of the last equation of the above display is bounded below by $\epsilon P\left[Z^2 > (-\log\epsilon)^{-1}\right]$, where $Z \sim N(0,1)$, which concludes that $E_{\pi^L}[\exp\{-\sigma^2\tau_n/\{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta\}\}]$ is strictly bounded from zero. So, there exists a strictly positive constant C such that $E_{\pi^L}[\exp\{-\sigma^2\tau\{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta\}^{-1}\}] > C > 0$.

Therefore,

$$
\begin{aligned}
& P_0\left[\frac{M_{n,I}}{-\log D_n(h_I;\mathbf{y})} > v_n\right] \\
= \; & P_0\Bigg[\log E_{\beta|\mathbf{y}}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta}\right\}\right] \\
& \quad > \log E_{\pi^L}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta}\right\}\right] - M_{n,I}v_n^{-1}\Bigg] \\
\leq \; & P_0\left[\left\{\log E_{\beta|\mathbf{y}}\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(I-Q_0)\Phi\beta}\right\}\right] > C - M_{n,I}v_n^{-1}\right\} \cap A_n\right] \\
& + P_0\left[A_n^c\right], \qquad \text{(A.44)}
\end{aligned}
$$

where $A_n = \{\mathbf{y} : \mathbf{y}^{\mathrm{T}}(I-Q_0)\mathbf{y} < F_0^{\mathrm{T}}(I-Q_0)F_0 + \sigma^2(K_n - d_0) + 2\sigma^2\{F_0^{\mathrm{T}}(I-Q_0)F_0/\sigma^2 + (K_n-d_0)\}^{1/2}\log n\}$.

Since $P_0[A_n^c] = o(1)$ by (A.43), it is sufficient to show that the first term in (A.44) is $o(1)$. By (A.37) in Lemma A.3.1, the first term can be bounded above by

$$
\begin{aligned}
& P_0\Bigg[\Bigg\{\log\left[\exp\left\{-\frac{\sigma^2\tau_n}{c_1\sigma_n^{*2}n^{-1/2} + \widetilde{d}_n + c_2\sigma_n^*(n\widetilde{d}_n)^{1/2}}\right\} + \exp\left\{-c_3\frac{\tau_n}{n^{1/2}}\right\}\right] \\
& + \exp\{-c_4 n\} > C - M_{n,I}v_n^{-1}\Bigg\} \cap A_n\Bigg] \\
\leq \; & I\left[-\frac{c_5\tau_n}{d_n + \sigma^2(nd_n)^{1/2}} > C - M_{n,I}v_n^{-1}\right], \qquad \text{(A.45)}
\end{aligned}
$$

where $I(\cdot)$ is the indicator function, $\sigma_n^{*2} = \tau_n\sigma^2/(1+\tau_n)$ and $\widetilde{d}_n = \tau_n^2\widehat{F}^{\mathrm{T}}(Q_\Phi - Q_0)\widehat{F}/(1+\tau_n)^2 = \tau_n^2\widehat{\mathbf{y}}^{\mathrm{T}}(Q_\Phi - Q_0)\mathbf{y}/(1+\tau_n)^2$ for some constants c_i for $i = 1,\ldots,5$. Since $M_{n,I} = \tau_n\{d_n + \sigma^2(nd_n)^{1/2}\}^{-1}$ and $v^{-1} \to 0$, it is clear that (A.45) is $o(1)$, which completes the proof. □

Proof of Proposition 12. The asymptotic property of the B-spline approximation De Boor

(1978) guarantees that if $f_0 \in C^{\alpha}[0,1]$, there exists some $\beta^{\infty} \in \mathbb{R}^{K_n}$, $\|\Phi\beta^{\infty} - F_0\|_{\infty} \preceq K_n^{-\alpha}\|f_0\|_{\alpha}$. By using this asymptotic inequality, it follows that

$$\begin{aligned} F_0^{\mathrm{T}}(Q_{\Phi} - Q_0)F_0 &= F_0^{\mathrm{T}}(I - Q_0)F_0 - F_0^{\mathrm{T}}(I - Q_{\Phi})F_0 \\ &\geq F_0^{\mathrm{T}}(I - Q_0)F_0 - n\|\Phi\beta^{\infty} - F_0\|_{\infty} \\ &\succeq F_0^{\mathrm{T}}(I - Q_0)F_0 - nK_n^{-\alpha}\|f_0\|_{\alpha}. \end{aligned}$$

Also, it is clear that $F_0^{\mathrm{T}}(Q_{\Phi} - Q_0)F_0 \leq F_0^{\mathrm{T}}(I - Q_0)F_0$, which completes the proof. □

Proof of Theorem 13. To show that $\pi^{NL}(\mathbf{t} \mid \mathbf{y})$ converges to one in probability, it is sufficient to show that H_{1n} and H_{2n} in (A.5) both converge zero in probability as n tends to ∞. We shall prove the Theorem by showing the follows:

For any fixed $\delta > 0$, $\epsilon > 0$ and any model $\mathbf{k} \in \Gamma_d$ (defined in Lemma A.1.2),

$$P_0\left[\frac{m_{\mathbf{k}}(\mathbf{y})}{m_{\mathbf{t}}(\mathbf{y})} > \epsilon p^{-d} q_n^{-1}\right] \leq p^{-d(1+\delta)}, \tag{A.46}$$

and for any model $\mathbf{k} \in \Gamma_{k,c_k,c_t}$ (defined in Lemma A.1.3),

$$P_0\left[\frac{m_{\mathbf{k}}(\mathbf{y})}{m_{\mathbf{t}}(\mathbf{y})} > \epsilon n^{-3} p^{-k} n^{-c_k} t^{-t}\right] \leq p^{-k(1+\delta)}. \tag{A.47}$$

Then, it is clear that H_{1n} and H_{2n} both converge to zero in probability by Lemma A.1.2 and Lemma A.1.3 respectively.

We first show that the normalizing constant of the nonlocal functional prior densities is asymptotically at rate of $(2\pi\tau_n)^{K_n/2}$, i.e.,
$\int \exp\{-\beta^{\mathrm{T}}\beta/(2\tau_n) - \sigma^2\tau_n/\beta^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta\}d\beta \asymp (2\pi\tau_n)^{K_n/2}$. Let $E_{\beta}[\cdot]$ and $P_{\beta}[\cdot]$ be the expectation and the probability induced by a random variable $\beta \sim N(0, \sigma^2\tau_n I)$. Then, by using the

Markov inequality and **(A3)**, it follows that

$$
\begin{aligned}
&\int \exp\{-\beta^{\mathrm{T}}\beta/(2\sigma^2\tau_n) - \tau_n/\beta^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta\}d\beta = (2\pi\tau_n)^{K_n/2}E_\beta\left[\exp\left\{-\frac{\sigma^2\tau_n}{\beta^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta}\right\}\right] \\
\geq\ & (2\pi\tau_n)^{K_n/2}\exp\{-\sigma^2\log nK_n/(\lambda_* n)\} \\
& \times P_\beta\left[\exp\{-\sigma^2\tau_n/\beta^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta\} > \exp\{-\sigma^2(\log n)K_n/(\lambda_* n)\}\right] \\
\geq\ & (2\pi\tau_n)^{K_n/2}\exp\{-\sigma^2(\log n)K_n/(\lambda_* n)\}P_\beta\left[\beta^{\mathrm{T}}\beta/(\sigma^2\tau_n) > 1/(\sigma^2\log n)\right] \\
\geq\ & (2\pi\tau_n)^{K_n/2}(1-o(1)),
\end{aligned}
$$

where λ_* is defined in **(A3)**. The last inequality is derived by using the fact that $P(W_{K_n} < x) \leq (x/K_n)^{K_n/2}\exp\{K_n/2 - x/2\}$ for $W_{K_n} \sim \chi^2_{K_n}$ and $x < K_n$. It is also clear that $\int \exp\{-\beta^{\mathrm{T}}\beta/(2\sigma^2\tau_n) - \sigma^2\tau_n/\beta^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta\}d\beta \leq (2\pi\tau_n)^{K_n/2}$, since $\exp\{-\sigma^2\tau_n/\beta^{\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta\} \leq 1$ for any $\beta \in \mathbb{R}^{K_n}$.

Second, we shall show that (A.46) holds. Recall that $\Gamma_d = \{\mathbf{k} : |\mathbf{k}| \leq q_n\ \mathbf{t} \subsetneq \mathbf{k},\ |\mathbf{k}| - |\mathbf{t}| = d\}$. By Lemma A.3.3, **(A2)** and **(A3)**, it follows that for any $\mathbf{k} \in \Gamma_d$, there exists $\delta > 0$ such

that

$$
\begin{aligned}
& P_0\left[\frac{m_{\mathbf{k}}(\mathbf{y})}{m_{\mathbf{t}}(\mathbf{y})} > \epsilon p^{-d} q_n^{-1}\right] \\
\leq\ & P_0\Bigg[(c_1\tau_n)^{-dK_n/2}\left(\frac{|\Phi_{\mathbf{k}}^{\mathrm{T}}\Phi_{\mathbf{k}} + 1/\tau_n I|}{|\Phi_{\mathbf{t}}^{\mathrm{T}}\Phi_{\mathbf{t}} + 1/\tau_n I|}\right)^{-1/2}\left(\frac{D_{\mathbf{k}}(\mathbf{y})}{D_{\mathbf{t}}(\mathbf{y})}\right) \\
& \times \exp\left\{\frac{\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{k}} - \widetilde{P}_{\mathbf{t}})\mathbf{y}}{2\sigma^2}\right\} > \epsilon p^{-d} q_n^{-1}\Bigg] \\
\leq\ & P_0\Bigg[(c_1\tau_n)^{-dK_n/2} Q_{n*}^{|\mathbf{k}|} Q_n^{*-|\mathbf{t}|} \exp\left\{\frac{\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{k}} - \widetilde{P}_{\mathbf{t}})\mathbf{y}}{2\sigma^2}\right\} \\
& \times \exp\left\{-\frac{c_2 d\tau_n u_n^{-1/2}\zeta_{n*}^{1/2}}{(K_n + \log p + \zeta_{n*}^{-1})^{1/2} n^{1/2}}\right\} > \epsilon p^{-d} q_n^{-1}\Bigg] \\
& + P_0\left[\frac{D_{\mathbf{k}}(\mathbf{y})}{D_{\mathbf{t}}(\mathbf{y})} > \exp\left\{-\frac{c_2 d\tau_n u_n^{-1/2}\zeta_{n*}^{1/2}}{(K_n + \log p + \zeta_{n*}^{-1})^{1/2} n^{1/2}}\right\}\right] \\
\leq\ & P_0\Bigg[\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{k}} - \widetilde{P}_{\mathbf{t}})\mathbf{y}/\sigma^2 > -2d\log p - 2\log q_n + dK_n \log\tau_n \\
& \times + \frac{c_2 d\tau_n u_n^{-1/2}\zeta_{n*}^{1/2}}{(K_n + \log p + \zeta_{n*}^{-1})^{1/2} n^{1/2}} + Z_n\Bigg] \\
& + p^{-d(1+\delta)},
\end{aligned}
$$

where $Q_{n*} = (\zeta_{n*}\lambda_* n/K_n + 1/\tau_n)^{-K_n/2}$, $Q_n^* = (\zeta_n^*\lambda^* n/K_n + 1/\tau_n)^{-K_n/2}$ and $Z_n = |\mathbf{k}|K_n \log(\zeta_{n*}\lambda_* n/K_n) - |\mathbf{t}|K_n \log(\zeta_n^*\lambda^* n/K_n)$, for some positive constants c_1 and c_2.

Let

$$
t_n = -2d\log p - 2\log q_n + dK_n\log(1+g_n) + \frac{c_2 d\tau_n u_n^{-1/2}\zeta_{n*}^{1/2}}{(K_n + \log p + \zeta_{n*}^{-1})^{1/2} n^{1/2}} + Z_n - F_0^{\mathrm{T}}(P_{\mathbf{k}} - P_{\mathbf{t}})F_0.
$$

We note that $\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{k}} - \widetilde{P}_{\mathbf{t}})\mathbf{y} \leq \mathbf{y}^{\mathrm{T}}(P_{\mathbf{k}} - P_{\mathbf{t}})\mathbf{y} + c_0 K_n (n\tau_n\lambda_*\zeta_{n*})^{-1}\mathbf{y}^{\mathrm{T}}P_{\mathbf{t}}\mathbf{y}$ for some constant c_0. Therefore, since $\mathbf{y}^{\mathrm{T}}(P_{\mathbf{k}} - P_{\mathbf{t}})\mathbf{y}/\sigma^2 \sim \chi^2_{dK_n}(F_0^{\mathrm{T}}(P_{\mathbf{k}} - P_{\mathbf{t}})F_0)$, Lemma A.1.4 and **(A5)** implies

that

$$
\begin{aligned}
& P_0\left[\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{k}}-\widetilde{P}_{\mathbf{t}})\mathbf{y}/\sigma^2 > F_0^{\mathrm{T}}(P_{\mathbf{k}}-P_{\mathbf{t}})F_0/\sigma^2 + t_n\right] \\
\leq\ & P_0\left[\mathbf{y}^{\mathrm{T}}(P_{\mathbf{k}}-P_{\mathbf{t}})\mathbf{y}/\sigma^2 > F_0^{\mathrm{T}}(P_{\mathbf{k}}-P_{\mathbf{t}})F_0/\sigma^2 + t_n/2\right] + P_0\left[\frac{c_0 K_n \mathbf{y}^{\mathrm{T}} P_{\mathbf{t}}\mathbf{y}}{n\tau_n\lambda_*\zeta_{n*}} > t_n/2\right] \\
\leq\ & c_3(t_n/(2dK_n))^{dK_n/2}\exp\{dK_n/2 - t_n/4\} \\
& + c_4\{F_0^{\mathrm{T}}(P_{\mathbf{k}}-P_{\mathbf{t}})F_0\}^{1/2} t_n^{-1}\exp\left\{-\frac{\sigma^2 t_n^2}{128 F_0^{\mathrm{T}}(P_{\mathbf{k}}-P_{\mathbf{t}})F_0}\right\} + \exp\{-c_5 n\},
\end{aligned}
$$

for some constant c_l with $l \in \{3,4,5\}$. Since $\tau_n\zeta_{n*}^{1/2}u_n^{-1/2}(K_n+\log p+\zeta_{n*}^{-1})^{-1/2}n^{-1/2} \succ \max\{F_0^{\mathrm{T}}(P_{\mathbf{k}}-P_{\mathbf{t}})F_0, Z_n, \log p\}$ by (A5), it follows that the last equation in the above display is bounded above by $p^{-d(1+\delta)}$ for any $\delta > 0$, which proves (A.46).

Third, we shall show that (A.47) holds. Recall that $\Gamma_{k,c_k,c_t} = \{\mathbf{k} : |\mathbf{k}| \leq q_n, |\mathbf{k}| = k,\ |\mathbf{k}\backslash\mathbf{t}| = c_k,\ \ |\mathbf{t}\backslash\mathbf{k}| = c_t\}$. By following similar steps used in the previous proof for (A.46), it follows that for any model $\mathbf{k} \in \Gamma_{k,c_k,c_t}$, there exists some $\delta > 0$ such that

$$
\begin{aligned}
& P_0\left[\frac{m_{\mathbf{k}}(\mathbf{y})}{m_{\mathbf{t}}(\mathbf{y})} > \epsilon n^{-3}p^{-k}n^{-c_k}t^{-t}\right] \\
\leq\ & P_0\Bigg[(2\pi\tau_n)^{-(c_k-c_t)K_n/2}\left(\frac{|\Phi_{\mathbf{k}}^{\mathrm{T}}\Phi_{\mathbf{k}} + 1/\tau_n I|}{|\Phi_{\mathbf{t}}^{\mathrm{T}}\Phi_{\mathbf{t}} + 1/\tau_n I|}\right)^{-1/2}\left(\frac{D_{\mathbf{k}}(\mathbf{y})}{D_{\mathbf{t}}(\mathbf{y})}\right)\exp\left\{\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{k}}-\widetilde{P}_{\mathbf{t}})\mathbf{y}/(2\sigma^2)\right\} \\
& > \epsilon n^{-3}p^{-k}n^{-c_k}t^{-t}\Bigg] \\
\leq\ & P_0\Bigg[(cn)^{-(c_k-c_t)K_n}Q_{n*}^{|\mathbf{k}|}Q_n^{*-|\mathbf{t}|}\exp\left\{\frac{\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{k}}-\widetilde{P}_{\mathbf{u}})\mathbf{y}}{2\sigma^2} - \frac{c' c_k\tau_n\zeta_{n*}^{1/2}u_n^{-1/2}}{(K_n+\log p+\zeta_{n*}^{-1})^{1/2}n^{1/2}}\right. \\
& \left. + q_n\log p + Z_n\right\} > p^{-k-\delta}\Bigg] \qquad \text{(A.48)} \\
& + P_0\left[\frac{D_{\mathbf{k}}(\mathbf{y})}{D_{\mathbf{t}}(\mathbf{y})} > \exp\left\{-\frac{c' c_k\tau_n\zeta_{n*}^{1/2}u_n^{-1/2}}{(K_n+\log p+\zeta_{n*}^{-1})^{1/2}n^{1/2}} + Z_n\right\}\right] \qquad \text{(A.49)} \\
& + P_0\left[\frac{\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{u}}-\widetilde{P}_{\mathbf{t}})\mathbf{y}}{2\sigma^2} > q_n\log p\right], \qquad \text{(A.50)}
\end{aligned}
$$

where $\mathbf{u} = \mathbf{k} \cup \mathbf{t}$ and $t = |\mathbf{t}|$ for some constant c and c'.

We are going to show that the three terms (A.48), (A.49), and (A.50) all are bounded above by $p^{-|\mathbf{k}|(1+\delta)}$ for some $\delta > 0$. Then, the proof is competed by Lemma A.1.3.

By Lemma A.3.3, (A.49) is bounded above by $p^{-|\mathbf{k}|(1+\delta)}$. Let $z_n = q_n \log p - F_0^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{t}})F_0/\sigma^2$. Then, since $\mathbf{y}^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{t}})\mathbf{y}/\sigma^2 \sim \chi^2_{c_k K_n}(F_0^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{t}})F_0/\sigma^2)$, by Lemma A.1.4, (A.50) also can be shown as

$$
\begin{aligned}
& P_0\left[\mathbf{y}^{\mathrm{T}}(\widetilde{P}_{\mathbf{u}} - \widetilde{P}_{\mathbf{t}})\mathbf{y}/\sigma^2 > 2q_n \log p\right] \\
\leq\ & P_0\left[\mathbf{y}^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{t}})\mathbf{y}/\sigma^2 > q_n \log p\right] + P_0\left[\frac{c_0 K_n \mathbf{y}^{\mathrm{T}} P_{\mathbf{t}}\mathbf{y}}{n\tau_n \lambda_* \zeta_{n*}} > q_n \log p\right] \\
\leq\ & c_1 \left(z_n/\{c_k K_n\}\right)^{c_k K_n/2} \exp\left\{|\mathbf{k} \setminus \mathbf{t}|K_n/2 - z_n/2\right\} \\
& + c_2 \left(F_0^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{t}})F_0/\sigma^2\right)^{1/2} z_n^{-1} \exp\left\{-z_n^2/\{32 F_0^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{t}})F_0/\sigma^2\}\right\} \\
& + P_0\left[\frac{c_0 K_n \mathbf{y}^{\mathrm{T}} P_{\mathbf{t}}\mathbf{y}}{n\tau_n \lambda_* \zeta_{n*}} > q_n \log p\right],
\end{aligned}
$$

for some constant c_1 and c_2. Since $P\left[c_0 K_n/(n\tau_n \lambda_* \zeta_{n*})\mathbf{y}^{\mathrm{T}} P_{\mathbf{t}}\mathbf{y}/\mathbf{y} > q_n \log p\right] \leq p^{-|\mathbf{k}|(1+\delta)}$ by Lemma A.1.4, $|\mathbf{k}| \leq q_n$ and $z_n \asymp q_n \log p$ by (A5), (A.50) is bounded above by $p^{-|\mathbf{k}|(1+\delta)}$ for any fixed $\delta > 0$.

Also, (A.48) can be bounded above by

$$
\begin{aligned}
& P_0\left[\frac{\mathbf{y}^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{k}})\mathbf{y}}{\sigma^2} < (2+\delta)q_n \log p - 2\log\left(\frac{Q_{n*}^{|\mathbf{k}|}}{Q_n^{*|\mathbf{t}|}}\right)\right. && \text{(A.51)} \\
& \left. + \frac{c' c_k \tau_n \zeta_{n*}^{1/2} u_n^{-1/2}}{(K_n + \log p + \zeta_{n*}^{-1})^{1/2} n^{1/2}}\right] && \text{(A.52)} \\
+\ & P_0\left[\frac{c_0 K_n \mathbf{y}^{\mathrm{T}} P_{\mathbf{u}}\mathbf{y}}{n\tau_n \lambda_* \zeta_{n*}} > \sigma^2 K_n \log n\right], && \text{(A.53)}
\end{aligned}
$$

for some constant c_3. We note that $\mathbf{y}^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{k}})\mathbf{y}/\sigma^2 \sim \chi^2_{c_t}(F_0^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{k}})F_0/\sigma^2)$. Since $F_0^{\mathrm{T}}(P_{\mathbf{u}} - P_{\mathbf{k}})F_0 \succ q_n \tau_n \zeta_{n*}^{1/2} u_n^{-1/2}/\{(K_n + \log p + \zeta_{n*}^{-1})^{1/2} n^{1/2}\}$ by (A5), Lemma A.1.4 implies that (A.51) is bounded above by $p^{-|\mathbf{k}|(1+\delta)}$ for some $\delta > 0$. Also, since $\mathbf{y}^{\mathrm{T}} P_{\mathbf{u}}\mathbf{y}/\sigma^2 \sim$

$\chi^2_{|\mathbf{u}|K_n}(F_0^{\mathrm{T}} P_{\mathbf{u}} F_0/\sigma^2)$, it follows that (A.53) is bounded above by $p^{-|\mathbf{k}|(1+\delta)}$ by Lemma A.1.4.

□

APPENDIX B

DETAILS OF COMPUTATION

B.1 Nonlocal Prior Densities for High-dimensional Linear Model Selection

In this section, we provide the Laplace approximation of the marginal likelihoods based on the nonlocal priors. Because closed form expressions for posterior model probabilities based on modified peMoM priors and modified piMoM priors are not available, we estimate the posterior model probabilities using Laplace approximations. For posterior probabilities based on the peMoM priors, an inverse-Gamma density with parameters (a_0, b_0) on σ^2 the Laplace approximation to the marginal density of the data for model $\mathbf{k}$ can be expressed as

$$\pi(\mathbf{k} \mid y) \propto (2\pi)^{|\mathbf{k}|/2} \left|V(\beta^*_{\mathbf{k}}, \sigma^{2*})\right|^{-1/2} \exp\{f(\beta^*_{\mathbf{k}}, \sigma^{2*})\} p(\mathbf{k}), \tag{B.1}$$

where

$$\begin{aligned}
(\beta^*_{\mathbf{k}}, \sigma^{2*}) &= \underset{(\beta_{\mathbf{k}}, \sigma^2)}{\operatorname{argmax}} f(\beta_{\mathbf{k}}, \sigma^2) \\
f(\beta_{\mathbf{k}}, \sigma^2) &= -(n/2 + |\mathbf{k}|/2 + a_0 + 1)\log\sigma^2 - (y - X_{\mathbf{k}}\beta_{\mathbf{k}})^T(y - X_{\mathbf{k}}\beta_{\mathbf{k}})/(2\sigma^2) \\
&\quad -\beta_{\mathbf{k}}^T\beta_{\mathbf{k}}/(2\sigma^2\tau_{n,p}) - \sum_{j=1}^{|\mathbf{k}|} \tau_{n,p}/\beta^2_{\mathbf{k},j} + |\mathbf{k}|(2/\sigma^2)^{1/2} - b_0/\sigma^2 + |\mathbf{k}|(\log\tau_{n,p})/2,
\end{aligned}$$

and $V(\beta_{\mathbf{k}}, \sigma^2)$ is a $(|\mathbf{k}|+1) \times (|\mathbf{k}|+1)$ matrix with the following blocks:

$$\begin{aligned}
V_{11} &= X_{\mathbf{k}}^T X_{\mathbf{k}}/\sigma^2 + I_{\mathbf{k}}/\sigma^2\tau_{n,p} + diag\left\{6\tau_{n,p}/\beta_{\mathbf{k},j}^4\right\}_{j=1,\ldots,|\mathbf{k}|} \\
V_{12} &= X_{\mathbf{k}}^T(y - X_{\mathbf{k}}\beta_{\mathbf{k}})/\sigma^4 - \beta_{\mathbf{k}}/\{\sigma^4\tau_{n,p}\} \\
V_{22} &= -(n/2 + |\mathbf{k}|/2 + a_0 + 1)/\sigma^4 + (y - X_{\mathbf{k}}\beta_{\mathbf{k}})^T(y - X_{\mathbf{k}}\beta_{\mathbf{k}})/\sigma^6 - \beta_{\mathbf{k}}^T\beta_{\mathbf{k}}/\tau_{n,p} \\
&\quad -3|\mathbf{k}|2^{1/2}\sigma^{-5}/4 + 2b_0/\sigma^6.
\end{aligned}$$

For the piMoM priors on $\beta_{\mathbf{k}}$, the Laplace approximation of the posterior model probability can be expressed as in (B.3), but with

$$\begin{aligned}
f(\beta_{\mathbf{k}}, \sigma^2) &= -(n/2 + a_0 + 1)\log\sigma^2 - (y - X_{\mathbf{k}}\beta_{\mathbf{k}})^T(y - X_{\mathbf{k}}\beta_{\mathbf{k}})/(2\sigma^2) - b_0/\sigma^2 \\
&\quad - \sum_{j=1}^{|\mathbf{k}|}\left\{r\log(\beta_{\mathbf{k},j}^2) + \tau_{n,p}/\beta_{\mathbf{k},j}^2\right\} + |\mathbf{k}|\left\{(r-1/2)\log\tau_{n,p} - \log\Gamma(r-1/2)\right\},
\end{aligned}$$

and $V(\beta_{\mathbf{k}}, \sigma^2)$ a $(|\mathbf{k}|+1) \times (|\mathbf{k}|+1)$ matrix with the following blocks:

$$\begin{aligned}
V_{11} &= X_{\mathbf{k}}^T X_{\mathbf{k}}/\sigma^2 + diag\left\{6\tau_{n,p}/\beta_{\mathbf{k},j}^4 - 2r/\beta_{\mathbf{k},j}^2\right\}_{j=1,\ldots,|\mathbf{k}|} \\
V_{12} &= X_{\mathbf{k}}^T(y - X_{\mathbf{k}}\beta_{\mathbf{k}})/\sigma^4 \\
V_{22} &= -(n/2 + a_0 + 1)/\sigma^4 + (y - X_{\mathbf{k}}\beta_{\mathbf{k}})^T(y - X_{\mathbf{k}}\beta_{\mathbf{k}})/\sigma^6 + 2b_0/\sigma^6.
\end{aligned}$$

B.2 Functional Horseshoe Prior for Nonparametric Subspace Shrinkage

In model (4.1), the conditional posterior distribution of τ based on the functional horseshoe prior can be expressed as

$$\pi(\tau \mid Y, \beta) \propto (\tau^2)^{-(k_n - d_0)/2 + b - 1/2}(1+\tau^2)^{-a-b}\exp\{-\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\beta/(2\sigma^2)\}.$$

By reparameterizing $\eta = 1/\tau^2$, the resulting conditional posterior distribution of η can be expressed as

$$\pi(\eta \mid Y, \beta) \propto \eta^{a+(k_n-d_0)/2-1} \exp\{-\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta/(2\sigma^2)\}\frac{1}{(1+\eta)^{a+b}}.$$

As in Polson et al. (2014), a slice sampling method (Neal, 2003) can be used to sample η from its conditional posterior distribution. The resulting MCMC algorithm is described in Algorithm 3.

Algorithm 3 MCMC algorithm for simple nonparametric regression models

Choose an initial value $\beta^{(0)}$ and $\tau^{(0)}$.
For l in $0 : (L-1)$
  Sample $\beta^{(l+1)}$ from $\mathrm{N}(\widetilde{\beta}_{\omega^{(l)}}, \sigma^2\widetilde{\Sigma}_{\omega^{(l)}})$, where $\widetilde{\beta}_{\omega}$ and $\widetilde{\Sigma}_{\omega}$ are defined in (4.7).
  (Slice sampling step) Set $\eta = 1/\tau^{2(l)}$ and $t = (\eta+1)^{-a-b}$.
    Sample $u \sim Unif(0,t)$ and set $t^* = u^{-(a+b)^{-1}} - 1$.
    Sample $\eta^* \sim truncated\ Gamma(a + (k_n - d_0)/2, \beta^{(l+1)\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^{(l+1)}/(2\sigma^2))$
      on $(0, t^*)$,
  Update $\tau^{(l+1)}$ by $\eta^{*-1/2}$.
End.

In the additive model in (4.14) with a product of the functional horseshoe priors, the conditional posterior distribution of β_j given ω_j and the other coefficients $\beta_{(-j)}$, for $j = 1, \ldots, p$, can be expressed as

$$\beta_j \mid \omega_j, \beta_{(-j)}, Y \sim \mathrm{N}\left(\widetilde{\beta}_{j,\omega}, \sigma^2\widetilde{\Sigma}_{j,\omega}\right),$$

where

$$\widetilde{\beta}_{j,\omega} = \widetilde{\Sigma}_{j,\omega}\Phi_j^{\mathrm{T}} r_j, \quad \widetilde{\Sigma}_{j,\omega} = (1-\omega_j)\left(\Phi_j^{\mathrm{T}}\Phi_j\right)^{-1}, \quad r_j = Y - \sum_{l\neq j}\Phi_l\beta_l. \tag{B.2}$$

It follows that sampling Algorithm 3 can be extended to additive regression models to obtain Algorithm 4 below.

Algorithm 4 MCMC algorithm for additive regression models

Choose an initial value $\beta_j^{(0)}$ and $\tau_j^{(0)}$ for $j = 1, \cdots, p$.
For l in $0 : (L-1)$
  For j in $1 : p$
    Sample $\beta_j^{(l+1)}$ from $\mathrm{N}(\widetilde{\beta}_{j,\omega^{(l)}}, \sigma^2\widetilde{\Sigma}_{j,\omega^{(l)}})$, where $\widetilde{\beta}_{j,\omega}$ and $\widetilde{\Sigma}_{j,\omega}$ are defined in (B.2).
  End.
  For j in $1 : p$
  (Slice sampling step)
    Set $\eta = 1/\tau_j^{2(l)}$ and $t = (\eta+1)^{-a-b}$.
    Sample $u \sim Unif(0,t)$ and set $t^* = u^{-(a+b)^{-1}} - 1$.
    Sample $\eta^* \sim truncated\ Gamma(a + k_n/2, \beta_j^{(l+1)\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_j^{(l+1)}/(2\sigma^2))$ on $(0, t^*)$,
    Update $\tau_j^{(l+1)}$ by $\eta^{*-1/2}$.
  End.
End.

B.3 Nonlocal Functional Priors for Nonparametric Hypothesis Testing and High-dimensional Model Selection

B.3.1 Modified Simplified Shotgun Stochastic Search with Screening (S5) for Additive Models

We consider a sequence of L number of temperatures $\{t_l\}_{l=1,\ldots,L}$ such that $t_1 > t_2 > \ldots > t_L > 0$. Also, we define a screened set by marginal correlations as $\mathbf{S}_{\mathbf{k}}^L(M) = \{j \in$

$\{1, \ldots, p\} : rank(|r_{\mathbf{k}}^T X_j| \leq M)\}$, where $r_{\mathbf{k}}$ is the residual of a model $\mathbf{k}$. Then, letting $\text{nbd}(\mathbf{k}) = \{\Gamma^-, \Gamma^+_{scr}\}$, where $\Gamma^- = \{\mathbf{k} \setminus \{j\} : j \in \mathbf{k}\}$ and $\Gamma^+_{scr} = \{\mathbf{k} \cup \{j\} : j \in \mathbf{k}^c \cap \mathbf{S}_{\mathbf{k}}(M)\}$, the modified S5 for additive model is illustrated in **Algorithm** 5.

Algorithm 5 Modified S5 for Additive Models

Set a temperature schedule $t_1 > t_2 > \ldots > t_L > 0$
Choose an initial model $\mathbf{k}^{(1,1)}$ and a set of variables after screening $\mathbf{S}_{\mathbf{k}^{(1,1)}}$ based on $\mathbf{k}^{(1,1)}$
For $l = 1$ in $l = L$
  For i in $1, \ldots, J-1$
    Compute all $\pi(\mathbf{k} \mid y)$ for all $\mathbf{k} \in \text{nbd}_{scr}(\mathbf{k}^{(i,l)})$
    Sample $\mathbf{k}^+$ and $\mathbf{k}^-$, from Γ^+_{scr} and Γ^-, with probabilities proportional to $\pi(\mathbf{k} \mid y)^{1/t_l}$
    Sample $\mathbf{k}^{(i+1,l)}$ from $\{\mathbf{k}^+, \mathbf{k}^-\}$,
      with probability proportional to $\{\pi(\mathbf{k}^+ \mid y)^{1/t_l}, \pi(\mathbf{k}^- \mid y)^{1/t_l}\}$
    Update the set of considered variables $\mathbf{S}_{\mathbf{k}^{(i+1,l)}}$ to be the union of variables in $\mathbf{k}^{(i+1,l)}$ and $\mathbf{S}^{INIS}_{\mathbf{k}^{(i+1,l)}}(M) \cup \mathbf{S}^{L}_{\mathbf{k}^{(i+1,l)}}(M)$.

B.3.2 Laplace Approximations of Marginal Likelihoods Based on Nonlocal Functional Prior Densities

In this section, we provide the Laplace approximation of the marginal likelihoods based on the inverse moment functional priors and $\pi(\sigma^2) \propto 1/\sigma^2$. Because explicit expressions for marginal likelihoods are not available, we estimate the marginal likelihoods using Laplace approximations. Letting $\eta = 1/\sigma^2$, the Laplace approximation to the marginal density of the data for model $\mathbf{k}$ can be expressed as

$$m_{\mathbf{k}}(\mathbf{y}) \approx (2\pi)^{|\mathbf{k}|K_n/2} \, |V(\beta^*_{\mathbf{k}}, \eta^*)|^{-1/2} \exp\{f(\beta^*_{\mathbf{k}}, \eta^*)\},$$

where

$$
\begin{aligned}
(\beta^*_{\mathbf{k}}, \eta^*) &= \underset{(\beta_{\mathbf{k}}, \eta)}{\operatorname{argmax}} f(\beta_{\mathbf{k}}, \eta) \\
f(\beta_{\mathbf{k}}, \eta) &= (n/2 + |\mathbf{k}|/2) \log \eta - \eta (\mathbf{y} - \Phi_{\mathbf{k}} \beta_{\mathbf{k}})^{\mathrm{T}} (\mathbf{y} - \Phi_{\mathbf{k}} \beta_{\mathbf{k}})/2 - \eta \beta_{\mathbf{k}}^{\mathrm{T}} \beta_{\mathbf{k}}/(2\tau_n) \\
&\quad - \sum_{j \in \mathbf{k}} \tau_n / \left(\eta \beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j \beta_j \right) - \sum_{j \in \mathbf{k}} \log Z_j,
\end{aligned}
$$

and $V(\beta_{\mathbf{k}}, \eta)$ is a $(|\mathbf{k}|K_n + 1) \times (|\mathbf{k}|K_n + 1)$ matrix with the following blocks:

$$
\begin{aligned}
V_{11} &= diag \left[\left\{ \frac{8\tau_n \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j \beta_j \beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j}{\eta (\beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j \beta_j)^3} - \frac{2\tau_n \Phi_j^{\mathrm{T}} \Phi_j}{\eta (\beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j \beta_j)^2} \right\}_{j \in \mathbf{k}} \right] \\
&\quad + \eta \Phi_{\mathbf{k}}^{\mathrm{T}} \Phi_{\mathbf{k}} + \frac{\eta}{\tau_n} I \\
V_{12} &= -\Phi_{\mathbf{k}}^{\mathrm{T}} (\mathbf{y} - \Phi_{\mathbf{k}} \beta_{\mathbf{k}}) + \frac{1}{\tau_n} \{\beta_j\}_{j \in \mathbf{k}} + \left\{ \frac{2\tau_n \Phi_j^{\mathrm{T}} \Phi_j \beta_j}{\eta \beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j \beta_j} \right\}_{j \in \mathbf{k}} \\
V_{22} &= (n/2 + |\mathbf{k}|K_n/2)/(\eta^2) + \sum_{j \in \mathbf{k}} \frac{2\tau_n}{\eta^2 \beta_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} (I - Q_0) \Phi_j \beta_j}.
\end{aligned}
$$

The prior normalizing constant Z_j for $j = 1, \ldots, p$ can be approximated by using important sampling procedures, since $Z_j = E_{\pi^L}[\exp\{\sigma^2 \tau_n / (\beta_j^{\mathrm{T}} \Phi_j (I - Q_0) \Phi_j \beta_j)\}]$ and $\pi^L \sim N(0, \sigma^2 \tau_n I)$.