Electronic Journal of Statistics
Vol. 8 (2014) 2804–2847
ISSN: 1935-7524
DOI: 10.1214/14-EJS963

High-dimensional Bayesian inference in nonparametric additive models

Zuofeng Shang

Purdue University
e-mail: shang9@purdue.edu

and

Ping Li

Rutgers University
e-mail: pingli@stat.rutgers.edu

Abstract: A fully Bayesian approach is proposed for ultrahigh-dimensional nonparametric additive models in which the number of additive components may be larger than the sample size, though ideally the true model is believed to include only a small number of components. Bayesian approaches can conduct stochastic model search and fulfill flexible parameter estimation by stochastic draws. The theory shows that the proposed model selection method has satisfactory properties. For instance, when the hyperparameter associated with the model prior is correctly specified, the true model has posterior probability approaching one as the sample size goes to infinity; when this hyperparameter is incorrectly specified, the selected model is still acceptable since asymptotically it is shown to be nested in the true model. To enhance model flexibility, two new g-priors are proposed and their theoretical performance is investigated. We also propose an efficient reversible jump MCMC algorithm to handle the computational issues. Several simulation examples are provided to demonstrate the advantages of our method.

AMS 2000 subject classifications: Primary 62G20, 62F25; secondary 62F15, 62F12.
Keywords and phrases: Bayesian group selection, ultrahigh-dimensionality, nonparametric additive model, posterior model consistency, size-control prior, generalized Zellner-Siow prior, generalized hyper-g prior, reversible jump MCMC.

Received January 2014.

1. Introduction

Suppose the data $\{Y_i, X_{1i}, \ldots, X_{pi}\}_{i=1}^n$ are *iid* copies of $Y, X_1, \ldots, X_p$ generated from the following model

$$Y_i = \sum_{j=1}^{p} f_j(X_{ji}) + \epsilon_i,\ i = 1, \ldots, n, \tag{1.1}$$

where ϵ_i's denote the zero-mean random errors, and for each $j = 1, \ldots, p$, X_j is a random variable taking values in $[0, 1]$, f_j is a function of X_j satisfying $E\{f_j(X_j)\} = 0$. The zero-expectation constraint is assumed for identifiability issue. Model (1.1) is called the additive component model; see [26, 41] for an excellent introduction. Suppose model (1.1) contains s_n significant covariates, and the remaining $p - s_n$ covariates are insignificant. Here we assume $p/n \to \infty$ as $n \to \infty$, denoted as $p \gg n$ or equivalently $n \ll p$, but ideally restrict $s_n = o(n)$, i.e., the true model is sparse. Our goal is to explore an automatic fully Bayesian procedure for selecting and estimating the significant (nonvanishing) f_j's in model (1.1).

When each f_j is linear in X_j, (1.1) reduces to a linear model. There have been a considerable number of frequentist approaches exploring issues on model selection in ultrahigh-dimensional situations, i.e., $\log p = O(n^k)$ for some $k > 0$. The representative ones include regularization-based approaches such as [24, 32, 37, 39, 48, 50, 53, 54, 56], and correlation-based approaches such as [13, 15, 51]. An insightful review is given by [14].

Model selection on the basis of a Bayesian framework is conceptually different. Specifically, Bayesian approaches conduct stochastic search of the models and evaluate each model by its posterior probability. Three major advantages of Bayesian selection methods are worth mentioning: (1) Bayesian approaches can perform model selection, parameter estimation and inference in a unified manner through posterior samples, no additional procedures such as prescreening, thresholding or data splitting are needed; (2) the choice of the hyperparameters is flexible by fulfilling stochastic draws; and (3) Bayesian methods allow the practitioners to incorporate prior information in the process of model search. The last feature might be attractive in small sample problems where prior information may be useful to address data insufficiency. There has been an amount of literature on Bayesian model selection in linear models. For example, when p is fixed, [1, 7, 17, 21, 34] show that, under certain regularity conditions, the posterior probability of the true model converges to one as n increases, in other words, *posterior model consistency* holds. This means that the proposed Bayesian selection method is asymptotically valid. Later on, these results were generalized by [30, 42] to the growing p situation with $p = O(n)$. In ultrahigh-dimensional situations, [44] considered a fully Bayesian hierarchical model with a prior controlling the model size and obtained posterior model consistency. A straightforward MCMC algorithm was developed for model search. Based on an extended Bayesian information criteria, [35] established posterior model consistency in generalized linear models.

However, in many practical applications there might be little evidence confirming linearity of the f_j's, for which a nonparametric assumption on the f_j's will largely enhance model flexibility, leading to the so-called nonparametric additive models. Surprisingly, theoretical studies relating to model selection in nonparametric additive models are almost all in frequentist settings. For instance, [31, 36, 40] explored issues relating to component selection with smoothing constraints assumed on the nonparametric functions. [25, 38] proposed penalty-based approaches and studied their asymptotic properties. [12] proposed a learn-

ing approach based on independent correlation and proved selection consistency. To the best of our knowledge, theoretical studies in Bayesian settings are nonexistent, especially when $p \gg n$. In terms of empirical evaluation, [46] proposed an objective Bayesian approach using penalized splines, [8] proposed a Bayesian framework based on adaptive regression trees, and [45] proposed a Bayesian framework based on a spike-and-slab prior induced from normal-mixture-inverse-gamma distributions. However, theoretical validity of these methods in ultrahigh-dimensional scenarios has not been justified.

In this paper, we propose a fully Bayesian hierarchical model which involves a new spike-and-slab prior on the function coefficients and a novel prior controlling the model size, namely, the *size-control prior*. The spike-and-slab prior has two important features: first, it either removes or includes the entire block of function coefficients, which is useful for model selection purpose; second, within each block, suitable decay rates are assumed on the function coefficients via their prior variances to produce smooth estimate of the nonparametric function. The size-control prior, which involves a *size-control parameter*, effectively restricts the scope of the target models, and facilitates both theoretical and computing issues. Based on the proposed Bayesian framework, we show that when the size-control parameter is correctly specified, posterior model consistency uniformly holds when the hyperparameters are confined by suitable ranges; when the size-control parameter is incorrectly specified, the selection results are still acceptable in the sense that the selected model is asymptotically nested in the true model, in other words, the number of false positives asymptotically vanishes. Interestingly, the asymptotic results are shown to be true even in the hyper-g prior settings. Furthermore, a novel and nontrivial blockwise MCMC procedure is proposed for computation. Our MCMC procedure allows stochastic search of all critical hyperparameters including the blocks of the function coefficients, the indicator variables representing inclusion/exclusion of the variables, the size-control parameter, and even the number of basis functions used for model fitting. The most challenging part in computation is the so-called trans-dimensional problem, which is successfully resolved by a novel and nontrivial variation of the "dimension-matching" technique proposed by [19] in the reversible jump MCMC approach. Simulation results demonstrate satisfactory selection and estimation accuracy of the proposed method. Performance under different basis structures is also examined. To the best of our knowledge, our work is the first one establishing a both theoretically and empirically effective fully Bayesian procedure for function component selection in ultrahigh-dimensional settings.

The rest of the paper is organized as follows. In Section 2, we carefully describe our fully Bayesian model and the prior distributions on the model parameters. In Section 3, asymptotic results are provided for both well specified and misspecified model spaces. In the meantime, two new types of g-priors are constructed and their theoretical properties are carefully studied. Section 4 contains the details of the MCMC algorithm. Section 5 includes the simulation examples showing the satisfactory performance of the proposed method. Section 7 summarizes the conclusions. Technical arguments are provided in Appendix.

2. A Bayesian nonparametric size-control model

Before describing our model, we introduce some notation and assumptions that are used frequently in this paper. We associate each f_j, $j = 1, \ldots, p$, a $0\backslash 1$ variable γ_j indicating the exclusion\inclusion of f_j in the model (1.1). Specifically, when $\gamma_j = 0$, $f_j = 0$ implies that f_j is not included in model (1.1); when $\gamma_j = 1$, $f_j \neq 0$ implies that f_j is included in model (1.1). Define $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$. For simplicity, we denote $j \in \boldsymbol{\gamma}$ to mean that $\gamma_j = 1$, and denote $j \in -\boldsymbol{\gamma}$ to mean $\gamma_j = 0$. Throughout we use $|\boldsymbol{\gamma}|$ to denote the number of ones in $\boldsymbol{\gamma}$, which is called the size of $\boldsymbol{\gamma}$. It is clear that there are totally 2^p possible $\boldsymbol{\gamma}$'s representing 2^p different models, each of which determines a subset of $\{f_j : j = 1, \ldots, p\}$ that are included in model (1.1). In other words, under $\boldsymbol{\gamma}$, model (1.1) is equivalent to $Y_i = \sum_{j \in \boldsymbol{\gamma}} f_j(X_{ji}) + \epsilon_i$, $i = 1, \ldots, n$. For any $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ and $\boldsymbol{\gamma}' = (\gamma_1', \ldots, \gamma_p')^T$, let $(\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}')_j = I(\gamma_j = 1, \gamma_j' = 0)$, and $(\boldsymbol{\gamma} \cap \boldsymbol{\gamma}')_j = I(\gamma_j = 1, \gamma_j' = 1)$. Thus, $\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}'$ is the $0\backslash 1$ vector indicating the functional components present in model $\boldsymbol{\gamma}$ but absent in model $\boldsymbol{\gamma}'$, and $\boldsymbol{\gamma} \cap \boldsymbol{\gamma}'$ is the $0\backslash 1$ vector indicating the functional components present in both models $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$. We say that $\boldsymbol{\gamma}$ is nested in $\boldsymbol{\gamma}'$ (denoted by $\boldsymbol{\gamma} \subset \boldsymbol{\gamma}'$) if $\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}'$ is zero. We further assume $\{f_j^0, j = 1, \ldots, p\}$ to be the true functional components, and denote $\boldsymbol{\gamma}^0 = (\gamma_1^0, \ldots, \gamma_p^0)^T$ with $\gamma_j^0 = I(f_j^0 \neq 0)$. That is, the data $\{Y_i, X_{1i}, \ldots, X_{pi}\}_{i=1}^n$ are truly sampled from model $Y_i = \sum_{j \in \boldsymbol{\gamma}^0} f_j^0(X_{ji}) + \epsilon_i$, $i = 1, \ldots, n$. Thus, $\boldsymbol{\gamma}^0$ represents the true model where data are generated, and $s_n = |\boldsymbol{\gamma}^0|$ denotes the size of the true model, i.e., the number of components f_j's included in the true model.

For $j = 1, \ldots, p$, define an inner product $\langle f_j, \tilde{f}_j \rangle_j = E\{f_j(X_j)\tilde{f}_j(X_j)\}$ for any $f_j, \tilde{f}_j \in \mathcal{H}_j$, where $\mathcal{H}_j$ is the class of functions on $[0, 1]$ satisfying $E\{|f_j(X_j)|^2\} < \infty$ and $E\{f_j(X_j)\} = 0$. This inner product induces a norm denoted by $\|\cdot\|_j$, that is, $\|f_j\|_j = \sqrt{E\{|f_j(X_j)|^2\}}$. Suppose the density function $d_j(x_j)$ of X_j satisfies $0 < K_1 \leq d_j(x_j) \leq K_2 < \infty$ for any $x_j \in [0, 1]$ and $j = 1, \ldots, p$, where K_1, K_2 are constants. Clearly, under $\langle \cdot, \cdot \rangle_j$, $\mathcal{H}_j$ is a well-defined Hilbert space. Let $\{\varphi_{jl}, l = 1, 2, \ldots\} \subset \mathcal{H}_j$ be the orthonormal basis functions for $\mathcal{H}_j$ under $\langle \cdot, \cdot \rangle_j$. Any function $f_j \in \mathcal{H}_j$ thus admits the Fourier series $f_j = \sum_{l=1}^{\infty} \beta_{jl}\varphi_{jl}$, with $\beta_{jl} = \langle f_j, \varphi_{jl} \rangle_j$ being the Fourier coefficients. It can be shown that $f_j = 0$ if and only if all the Fourier coefficients β_{jl}'s are zero. Therefore, to detect whether f_j vanishes or not, it is sufficient to detect whether the β_{jl}'s are zero. In general, f_j might correspond to infinitely many Fourier coefficients. Handling all the Fourier coefficients is computationally infeasible. Furthermore, it is commonly believed that only a finite subset of the Fourier coefficients capture most of the information possessed by f_j. Thus, we consider the *partial Fourier series* $f_j \approx \sum_{l=1}^{m} \beta_{jl}\varphi_{jl}$ with truncation parameter m, where $m = m_n$ is a sequence increasing with n. General theory on Fourier analysis leads to that $\|f_j - \sum_{l=1}^{m} \beta_{jl}\varphi_{jl}\|_j$ approaches zero as $m \to \infty$, showing the validity of such approximation. We introduce some additional matrix notation to simplify the expression of our model. For $j = 1, \ldots, p$ and $l = 1, \ldots, m$, define $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm})^T$, $\boldsymbol{\Phi}_{jl} = (\varphi_{jl}(X_{j1}), \ldots, \varphi_{jl}(X_{jn}))^T$, and $\mathbf{Z}_j = (\boldsymbol{\Phi}_{j1}, \ldots, \boldsymbol{\Phi}_{jm})$.

Thus, each $\mathbf{Z}_j$ is n by m. For and $\boldsymbol{\gamma}$, define $\mathbf{Z}_{\boldsymbol{\gamma}} = (\mathbf{Z}_j, j \in \boldsymbol{\gamma})$, the n by $m|\boldsymbol{\gamma}|$ matrix formed by $\mathbf{Z}_j$'s with $j \in \boldsymbol{\gamma}$, and define $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ to be the $m|\boldsymbol{\gamma}|$-vector of Fourier coefficients formed by $\boldsymbol{\beta}_j$'s with $j \in \boldsymbol{\gamma}$. Define $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ to be the response vector.

We assume that the model errors ϵ_i's are *iid* zero-mean Gaussian with variance σ^2, therefore, model (1.1), given $\boldsymbol{\gamma}$, f_j's and σ^2, becomes

$$Y_i \sim N(\sum_{j\in\boldsymbol{\gamma}} f_j(X_{ji}), \sigma^2), \ i = 1, \ldots, n. \tag{2.1}$$

Since each f_j can be well approximated by $\sum_{l=1}^m \beta_{jl}\varphi_{jl}$ for some sufficiently large m, the mean of Y_i is approximately $\sum_{j\in\boldsymbol{\gamma}}\sum_{l=1}^m \beta_{jl}\varphi_{jl}(X_{ji})$. Thus, (2.1) is approximately $Y_i \sim N(\sum_{j\in\boldsymbol{\gamma}}\sum_{l=1}^m \beta_{jl}\varphi_{jl}(X_{ji}), \sigma^2)$. In matrix form, this becomes

$$\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \sim N(\mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2\mathbf{I}_n). \tag{2.2}$$

When $\gamma_j = 0$, $f_j = 0$ implies that all the Fourier coefficients β_{jl}'s are zero. When $\gamma_j = 1$, $f_j \neq 0$, we place normal prior distributions over its Fourier coefficients. Explicitly, for $j = 1, \ldots, p$, we adopt the spike-and-slab prior for β_{jl}'s, i.e,

$$\beta_{jl}|\gamma_j, \sigma^2 \sim (1-\gamma_j)\delta_0 + \gamma_j N(0, c_j\sigma^2\tau_l^2), \ l = 1, \ldots, m, \tag{2.3}$$

where $\delta_0(\cdot)$ is the point mass measure concentrating on zero, $\{\tau_l^2, l \geq 1\}$ is a fixed nonincreasing sequence, and c_j's are temporarily assumed to be fixed. Note that the c_j's are used to control the variance of the nonzero coefficients, and therefore can be viewed as the *variance-control parameters*. In many applications we may choose $\tau_l^2 = l^{-(2\omega+1)}$ for $l \geq 1$, where $\omega > 1/2$ is a fixed constant characterizing the degree of smoothness; see, e.g., [4]. The prior (2.3) can be viewed as a direct multivariate extension of the conventional spike-and-slab prior on scalar coefficients considered by [9]. Note that $\gamma_j = 0$ or 1 will exclude or include the entire block of the coefficients β_{jl}'s, and within the nonvanishing block, the coefficients follow the zero-mean Gaussian priors with variances decaying at the rates τ_l^2's, which may be useful to produce smooth estimates of the functions. In [45], a different type of spike-and-slab prior was considered. Specifically, each coefficient block is represented as the product of a scalar with normal-mixture-inverse-gamma prior and a vector whose entries follow the bivariate mixture normal priors with a constant variance.

A variety of priors can be assumed on σ^2. For convenience, we consider the inverse χ^2 prior, i.e.,

$$1/\sigma^2 \sim \chi_\nu^2, \tag{2.4}$$

where ν is a fixed hyperparameter. Other priors such as the noninformative priors or the inverse Gamma priors can also be applied. All the results developed in this paper can be extended to such situations without substantial difficulty.

In high-dimensional inference, it is commonly believed that only a small subset of covariates contribute substantially to the model. Treating this as prior information, the models with larger sizes should be assigned with zero prior

probabilities, and only the models with smaller sizes should be assigned with positive weights. We call this a size-control prior on the model space. Namely, we choose the prior on $\boldsymbol{\gamma}$ as

$$p(\boldsymbol{\gamma}) = \begin{cases} \pi_{\boldsymbol{\gamma}}, & \text{if } |\boldsymbol{\gamma}| \le t_n, \\ 0, & \text{otherwise,} \end{cases} \tag{2.5}$$

where $\pi_{\boldsymbol{\gamma}}$ for $|\boldsymbol{\gamma}| \le t_n$ are fixed positive numbers summing to one, i.e., $\sum_{|\boldsymbol{\gamma}| \le t_n} \pi_{\boldsymbol{\gamma}} = 1$, and $t_n \in (0, n)$ is an integer-valued hyperparameter controlling the sizes of the candidate models. We name the set of models whose sizes are not exceeding t_n as the *target model space*.

Denote $\mathbf{D}_n = \{Y_i, X_{1i}, \ldots, X_{pi}\}_{i=1}^n$ to be the full data variable. It can be shown by direct calculations that, based on the above hierarchical model (2.2)–(2.5), the joint posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)$ is

$$\begin{aligned} & p(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 | \mathbf{D}_n) \\ \propto\ & p(\mathbf{Y} | \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \mathbf{X}_j\text{'s}) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma}, \sigma^2) p(\boldsymbol{\gamma}) p(\sigma^2) \\ \propto\ & \sigma^{-(n+\nu+2)} \exp\left(-\frac{\|\mathbf{Y} - \mathbf{Z}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + 1}{2\sigma^2} \right) p(\boldsymbol{\gamma}) \\ & \times \prod_{j \in \boldsymbol{\gamma}} \left(\sqrt{2\pi c_j} \sigma \right)^{-m} \det(\mathbf{T}_m)^{-1/2} \exp\left(-\frac{\boldsymbol{\beta}_j^T \mathbf{T}_m^{-1} \boldsymbol{\beta}_j}{2 c_j \sigma^2} \right), \end{aligned} \tag{2.6}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector, and $\mathbf{T}_m = \mathrm{diag}(\tau_1^2, \ldots, \tau_m^2)$. Integrating out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2 in (2.6), it can be checked that the marginal posterior of $\boldsymbol{\gamma}$ is

$$p(\boldsymbol{\gamma} | \mathbf{D}_n) \propto \det(\mathbf{W}_{\boldsymbol{\gamma}})^{-1/2} p(\boldsymbol{\gamma}) \left(1 + \mathbf{Y}^T (\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}} \mathbf{U}_{\boldsymbol{\gamma}}^{-1} \mathbf{Z}_{\boldsymbol{\gamma}}^T) \mathbf{Y} \right)^{-(n+\nu)/2}, \tag{2.7}$$

where $\mathbf{W}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{1/2} \mathbf{U}_{\boldsymbol{\gamma}} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{1/2}$, $\mathbf{U}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \mathbf{Z}_{\boldsymbol{\gamma}}^T \mathbf{Z}_{\boldsymbol{\gamma}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \mathrm{diag}(c_j \mathbf{T}_m, j \in \boldsymbol{\gamma})$ is the $m|\boldsymbol{\gamma}|$ by $m|\boldsymbol{\gamma}|$ diagonal matrix with diagonal elements $(c_j \tau_1^2, \ldots, c_j \tau_m^2)$ for $j \in \boldsymbol{\gamma}$. We adopt the convention $\mathbf{Z}_\emptyset = 0$ and $\boldsymbol{\Sigma}_\emptyset = \mathbf{U}_\emptyset = \mathbf{W}_\emptyset = 1$, where $\emptyset$ means the null model, i.e., the vector $\boldsymbol{\gamma}$ with all elements being zero. So (2.7) is meaningful for $\boldsymbol{\gamma} = \emptyset$. The selected model $\widehat{\boldsymbol{\gamma}}$ is defined to be the one that maximizes $p(\boldsymbol{\gamma} | \mathbf{D}_n)$. Clearly, $\widehat{\boldsymbol{\gamma}}$ belongs to the target model space since any model outside the target space has zero posterior probability.

3. Main results

Suppose the data are truly drawn from the model $Y_i = \sum_{j \in \boldsymbol{\gamma}^0} f_j^0(X_{ji}) + \epsilon_i$, where ϵ_i's $\overset{iid}{\sim} N(0, \sigma_0^2)$ are independent of X_{ji}'s, σ_0^2 is a fixed (unknown) positive number, and $f_j^0 \in \mathcal{H}_j$ for $j \in \boldsymbol{\gamma}^0$. Recall that $\boldsymbol{\gamma}^0$ is a p-dimensional $0\backslash 1$-vector representing the true model, and $s_n = |\boldsymbol{\gamma}^0|$ denotes its size. We only consider $s_n > 0$, i.e., the true model is non-null. Any f_j^0 for $j \in \boldsymbol{\gamma}^0$ admits the Fourier

expansion $f_j^0 = \sum_{l=1}^{\infty} \beta_{jl}^0 \varphi_{jl}$, where β_{jl}^0's represent the "true" unknown Fourier coefficients of f_j^0. When m is sufficiently large, f_j^0 is approximated by its partial Fourier series, that is, $f_j^0 \approx \sum_{l=1}^{m} \beta_{jl}^0 \varphi_{jl}$. To insure that such partial Fourier series is a valid approximation, we assume a uniform error rate on the tails of the Fourier series. Specifically, we assume that there are some positive constants $a > 1$ and C_β such that

$$\max_{m \geq 1} \max_{j \in \boldsymbol{\gamma}^0} m^a \sum_{l=m+1}^{\infty} |\beta_{jl}^0|^2 \leq C_\beta. \tag{3.1}$$

It is easy to see that (3.1) is equivalent to $\max_{j \in \boldsymbol{\gamma}^0} \|f_j^0 - \sum_{l=1}^{m} \beta_{jl}^0 \varphi_{jl}\|_j^2 = O(m^{-a})$, uniformly for $m \geq 1$. That is, the errors of the partial Fourier series of the nonzero f_j^0's uniformly decrease to zero at rate m^{-a}. For instance, when f_j^0's uniformly belong to the Sobolev's ellipsoid of order $a/2$, i.e., $\max_{j \in \boldsymbol{\gamma}^0} \sum_{l=1}^{\infty} l^a |\beta_{jl}^0|^2 < \infty$, for some constant $a > 0$, it can be checked that (3.1) holds. Namely, a measures the degree of smoothness of the nonzero functions. A larger a implies that the nonzero functions are more smooth.

Define $l_n = \sum_{j \in \boldsymbol{\gamma}^0} \|f_j^0\|_j^2$ and $\theta_n = \min_{j \in \boldsymbol{\gamma}^0} \|f_j^0\|_j$. Define $\mathbf{P}_{\boldsymbol{\gamma}} = \mathbf{Z}_{\boldsymbol{\gamma}} (\mathbf{Z}_{\boldsymbol{\gamma}}^T \mathbf{Z}_{\boldsymbol{\gamma}})^{-1} \mathbf{Z}_{\boldsymbol{\gamma}}^T$ to be the n by n projection (or smoothing) matrix corresponding to $\boldsymbol{\gamma}$. We adopt the convention $\mathbf{P}_{\emptyset} = 0$. Let $\lambda_-(\mathbf{A})$ and $\lambda_+(\mathbf{A})$ be the minimal and maximal eigenvalues of matrix $\mathbf{A}$. Suppose the truncation parameter m is chosen within the range $[m_1, m_2]$, where $m_1 = m_{1n}$, $m_2 = m_{2n}$ with $m_1 \leq m_2$ are positive sequences approaching infinity as $n \to \infty$. The variance-control parameters c_j's are chosen within $[\underline{\phi}_n, \bar{\phi}_n]$ for some positive sequences $\underline{\phi}_n$, $\bar{\phi}_n$.

3.1. Well specified target model space

In this section we present our first theorem on posterior consistency of our model selection procedure. We consider the situation $t_n \geq s_n$, that is, the hyperparameter t_n is correctly specified as being no less than the size of the true model. Thus, the true model is among our target model space, for which we say that the target model space is well specified. We will present a set of sufficient conditions and show that under these conditions, the posterior probability of the true model converges to one in probability. Thus, the selection procedure asymptotically yields the true model.

Define

$$S_1(t_n) = \{\boldsymbol{\gamma} | \boldsymbol{\gamma}^0 \subset \boldsymbol{\gamma}, \boldsymbol{\gamma} \neq \boldsymbol{\gamma}^0, |\boldsymbol{\gamma}| \leq t_n\},$$

and

$$S_2(t_n) = \{\boldsymbol{\gamma} | \boldsymbol{\gamma}^0 \text{ is not nested in } \boldsymbol{\gamma}, |\boldsymbol{\gamma}| \leq t_n\}.$$

It is clear that $S_1(t_n)$ and $S_2(t_n)$ are disjoint, and $S(t_n)$ defined by $S(t_n) = S_1(t_n) \bigcup S_2(t_n) \bigcup \{\boldsymbol{\gamma}^0\}$ is the class of all models with size not exceeding t_n, i.e., the target model space. We first list some conditions that are used to show our theorem.

Assumption A.1. *There exists a positive constant c_0 such that, as $n \to \infty$, with probability approaching one*

$$1/c_0 \le \min_{m\in[m_1,m_2]} \min_{\boldsymbol{\gamma}\in S_2(t_n)} \lambda_- \left(\frac{1}{n}\boldsymbol{Z}^T_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{\gamma}})\boldsymbol{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\right)$$
$$\le \max_{m\in[m_1,m_2]} \max_{\boldsymbol{\gamma}\in S_2(t_n)} \lambda_+ \left(\frac{1}{n}\boldsymbol{Z}^T_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\boldsymbol{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\right) \le c_0,$$

and

$$\min_{m\in[m_1,m_2]} \min_{\boldsymbol{\gamma}\in S_1(t_n)} \lambda_- \left(\frac{1}{n}\boldsymbol{Z}^T_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{\gamma}^0})\boldsymbol{Z}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}\right) \ge 1/c_0.$$

Assumption A.2. $\sup_n \max_{\boldsymbol{\gamma}\in S(t_n)} \frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}^0)} < \infty$.

Assumption A.2 clearly holds if $p(\boldsymbol{\gamma})$ is chosen to be constant for $\boldsymbol{\gamma} \in S(t_n)$. More generally, one can use the following prior

$$p(\boldsymbol{\gamma}) = Cq^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|},\ |\boldsymbol{\gamma}| \le t_n; 0,\ \text{otherwise}, \tag{3.2}$$

where $C > 0$ is normalizing constant such that $p(\boldsymbol{\gamma})$ defines a valid probability measure, and $q = q_n \in (0,1)$. We choose q such that $\frac{q}{1-q} = 1 + 1/a_n$ for some positive sequence a_n satisfying $t_n = O(a_n)$. The prior (3.2) satisfies Assumption A.2. Next we verify this argument. If $t_n \ge s_n$, then it can be shown that for any $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| \le t_n$,

$$\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}^0)} = \left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|-s_n} = (1+1/a_n)^{|\boldsymbol{\gamma}|-s_n} \le (1+1/a_n)^{t_n} = O(1),$$

where the $O(1)$ term in the final equation is not depending on $\boldsymbol{\gamma}$. This shows that Assumption A.2 holds.

Assumption A.3. *There exists a positive sequence $\{h_m, m \ge 1\}$ such that, as $m, m_1, m_2 \to \infty$, $h_m \to \infty$, $m^{-a}h_m$ decreasingly converges to zero, mh_m increasingly converges to ∞, and $\sum_{m_1\le m\le m_2} 1/h_m = o(1)$. Furthermore, the sequences $m_1, m_2, h_m, s_n, t_n, \theta_n, l_n, \underline{\phi}_n, \bar{\phi}_n$ satisfy*

(1). $m_2 h_{m_2} s_n = o(n\min\{1,\theta_n^2\})$ *and* $m_1^{-a}h_{m_1}s_n^2 = o(\min\{1, n^{-1}m_1\log(\underline{\phi}_n), \theta_n^2, \theta_n^4\})$;
(2). $t_n \ge s_n$ *and* $t_n \log p = o(n\log(1+\min\{1,\theta_n^2\}))$;
(3). $l_n = O(\underline{\phi}_n\tau^2_{m_2})$ *and* $\log p = o(m_1\log(n\underline{\phi}_n\tau^2_{m_2}))$;
(4). $m_2 s_n \log(1+n\bar{\phi}_n) = o(n\log(1+\min\{1,\theta_n^2\}))$.

In the following proposition we show that Assumption A.1 holds under suitable dependence assumption among the predictors X_j's. To clearly describe this assumption, let $\{X_j\}_{j=1}^\infty$ be a stationary sequence taking values in $[0,1]$, and define its ρ-mixing coefficient to be $\rho(|j-j'|) = \sup_{f,g}|E\{f(X_j)g(X_{j'})\} - E\{f(X_j)\}E\{g(X_{j'})\}|$, where the supremum is taken over the measurable functions f and g with $E\{f(X_j)^2\} = E\{g(X_{j'})^2\} = 1$. Ideally we assume that the predictors $X_1, \ldots, X_p$ in model (1.1) are simply the first p elements of $\{X_j\}_{j=1}^\infty$.

Proposition 3.1. *Suppose*

$$\sum_{r=1}^{\infty}\rho(r)<1/2,\quad t_n^2m_2^2\log p=o(n),\quad \textit{and}\quad \max_{1\le j\le p}\sup_{l\ge 1}\|\varphi_{jl}\|_{\sup}<\infty,$$

where $\|\cdot\|_{\sup}$ *denotes the supnorm. Then there is a constant* $c_0>0$ *such that with probability approaching one*

$$\begin{aligned} c_0^{-1} &\le \min_{m\in[m_1,m_2]}\min_{0<|\boldsymbol{\gamma}|\le 2t_n}\lambda_-\left(\frac{1}{n}\boldsymbol{Z}_{\boldsymbol{\gamma}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}}\right)\\ &\le \max_{m\in[m_1,m_2]}\max_{0<|\boldsymbol{\gamma}|\le 2t_n}\lambda_+\left(\frac{1}{n}\boldsymbol{Z}_{\boldsymbol{\gamma}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}}\right)\le c_0. \end{aligned}\tag{3.3}$$

Furthermore, (3.3) implies Assumption A.1.

Assumption A.2 holds if we choose $p(\boldsymbol{\gamma})$ to be constant for all $|\boldsymbol{\gamma}|\le t_n$, i.e., we adopt an indifference prior over the target model space. To see when Assumption A.3 holds, we look at a special example. We choose $\tau_l^2=l^{-5}$ for $l\ge 1$. Suppose $\log p\propto n^k$ for $0<k<1$, $\psi_n\propto 1$, and the smoothness parameter $a=4$. Choose $m_1=\zeta n^{1/5}+c_{1n}$ and $m_2=\zeta n^{1/5}+c_{2n}$, where $\zeta>0$ is constant, $c_{1n}=o((\log n)^r)$, $c_{2n}=o((\log n)^r)$, $c_{1n}\le c_{2n}$, and $r>0$ is a constant. Note that such choice of m_1 and m_2 yields minimax error rate in univariate regression. Let $h_m=(\log m)^r$ for $m\ge 1$. Ideally we suppose that the selected t_n is greater than s_n. Choose $\underline{\phi}_n$ and $\bar{\phi}_n$ as $\log(\underline{\phi}_n)\propto n^{k_1}$ and $\log(\bar{\phi}_n)\propto n^{k_2}$ with $\max\{0,k-1/5\}<k_1<k_2<4/5$. In this simple situation, it can be directly verified that Assumption A.3 holds if the true model size s_n satisfies $s_n=o(n^{\min\{1-k,k_1/2\}}(\log n)^{-r})$ and $t_n\ge s_n$ with $t_n=O(s_n)$. Furthermore, Proposition 3.1 says that to satisfy Assumption A.1, an additional sufficient condition is $t_n^2m_2^2\log p=o(n)$, which implies $k<3/5$. Therefore, if we further require that $t_n\propto 1$, then the dimension p cannot exceed the order $\exp(O(n^{3/5}))$, which coincides with the finding by [40].

Theorem 3.2. *Under Assumptions A.1 to A.3, as* $n\to\infty$,

$$\min_{m\in[m_1,m_2]}\inf_{\underline{\phi}_n\le c_1,\ldots,c_p\le\bar{\phi}_n}p(\boldsymbol{\gamma}^0|\boldsymbol{D}_n)\to 1,\ \textit{in probability.}\tag{3.4}$$

Theorem 3.2 says that under mild conditions the posterior probability of the true model converges to one in probability. This means, with probability approaching one, our Bayesian method selects the true model, which guarantees the validity of the proposed approach. Here, convergence holds uniformly over c_j's $\in[\underline{\phi}_n,\bar{\phi}_n]$ and $m\in[m_1,m_2]$. This means, the selection result is insensitive to the choice of c_j's and m when they belong to suitable ranges. It is well known that choosing the truncation parameter m is a practically difficult problem in nonparametrics; see [12, 40]. Therefore, a method that is insensitive to the choice of the truncation parameter within certain range will be highly useful. In Theorem 3.2 we theoretically show that the proposed Bayesian selection method is among the ones which provide insensitive selection results. On the other hand, we also show that our method is insensitive to the choice of the variance-control

parameters c_j's. This is both theoretically and practically useful since it allows us to place an additional prior, such as the g-priors, over the c_j's while preserving the desired posterior model consistency; see Section 3.4. To the best of our knowledge, Theorem 3.2 is the first theoretical result showing the validity of the Bayesian methods in function component selection in ultrahigh-dimensional settings.

The proof of Theorem 3.2 relies on a uniform approximation of the posterior odds ratios. Accurate probabilistic upper bounds are built uniformly for all these ratios so that the aggregation of the ratios tends to zero. Then the posterior probability of the true model is proven to approach one based on these upper bounds.

Frequentist approaches such as penalized least squares typically require selection of smoothing parameters for functional estimation and selection ([25, 38, 40]), or require certain procedures to improve performance, such as finding the initial estimators ([25]) and thresholding ([40]). In our Bayesian approach, the parameters c_j play similar role as frequentist smoothing parameter. We avoid manual selection of c_j by placing a prior on it. Selection of c_j is automatically done in MCMC sampling. Besides, our approach does not require initial estimators of the functional components, and does not need thresholding procedures. Relevant papers in (generalized) linear regressions also include [28, 29] who obtained posterior contraction results and [35] who applied a Laplace approximation approach. Instead, our interest is a fully Bayesian method with desired posterior selection consistency.

3.2. Misspecified target model space

In this section, we investigate the case $0 < t_n < s_n$, that is, t_n is misspecified as being smaller than the size of the true model. Therefore, the true model is outside the target model space, for which we say that the target model space is misspecified. We conclude that in this false setting the selected model is still not "bad" because it can be asymptotically nested in the true model, uniformly for the choice of m and c_j's.

Define $T_0(t_n) = \{\boldsymbol{\gamma}|0 \le |\boldsymbol{\gamma}| \le t_n, \boldsymbol{\gamma} \subset \boldsymbol{\gamma}^0\}$, $T_1(t_n) = \{\boldsymbol{\gamma}|0 < |\boldsymbol{\gamma}| \le t_n, \boldsymbol{\gamma} \cap \boldsymbol{\gamma}^0 \neq \emptyset,\ \boldsymbol{\gamma}$ is not nested in $\boldsymbol{\gamma}^0\}$, and $T_2(t_n) = \{\boldsymbol{\gamma}|0 < |\boldsymbol{\gamma}| \le t_n, \boldsymbol{\gamma} \cap \boldsymbol{\gamma}^0 = \emptyset\}$. It is easy to see that $T_0(t_n), T_1(t_n), T_2(t_n)$ are disjoint and $T(t_n) = T_0(t_n) \cup T_1(t_n) \cup T_2(t_n)$ is exactly the target model space, i.e., the class of $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| \le t_n$. Throughout this section, we make the following assumptions.

Assumption B.1. *There exist a positive constant d_0 and a positive sequence ρ_n such that, when $n \to \infty$, with probability approaching one,*

$$\begin{aligned} d_0^{-1} &\le \min_{m \in [m_1, m_2]} \min_{0 < |\boldsymbol{\gamma}| \le s_n} \lambda_- \left(\frac{1}{n} \boldsymbol{Z}_{\boldsymbol{\gamma}}^T \boldsymbol{Z}_{\boldsymbol{\gamma}} \right) \\ &\le \max_{m \in [m_1, m_2]} \max_{0 < |\boldsymbol{\gamma}| \le s_n} \lambda_+ \left(\frac{1}{n} \boldsymbol{Z}_{\boldsymbol{\gamma}}^T \boldsymbol{Z}_{\boldsymbol{\gamma}} \right) \le d_0, \ \textit{and} \end{aligned} \tag{3.5}$$

$$\max_{m\in[m_1,m_2]}\max_{\boldsymbol{\gamma}\in T(s_n-1)}\lambda_+\left(\boldsymbol{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}^T\boldsymbol{P}_{\boldsymbol{\gamma}}\boldsymbol{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\right)\le\rho_n. \tag{3.6}$$

Assumption B.2. $\sup_n \max_{\boldsymbol{\gamma},\boldsymbol{\gamma}'\in T(t_n)}\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}')}<\infty$.

It can be shown that the prior defined as in (3.2) satisfies Assumption B.2. To see this, if $t_n < s_n$, then it can be shown that for any $\boldsymbol{\gamma},\boldsymbol{\gamma}'$ with $|\boldsymbol{\gamma}|\le t_n$ and $|\boldsymbol{\gamma}'|\le t_n$,

$$\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}')}=\left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|-|\boldsymbol{\gamma}'|}\le(1+1/a_n)^{t_n}=O(1),$$

where the $O(1)$ term in the final equation is not depending on $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$. This shows that Assumption B.2 holds.

Assumption B.3. *There exists a positive sequence $\{h_m, m\ge 1\}$ such that, as $m, m_1, m_2\to\infty$, $h_m\to\infty$, $m^{-a}h_m$ decreasingly converges to zero, mh_m increasingly converges to ∞, and $\sum_{m_1\le m\le m_2}1/h_m=o(1)$. Furthermore, the sequences $m_1, m_2, h_m, s_n, \theta_n, l_n, \underline{\phi}_n$ satisfy*

(1). $m_2h_{m_2}s_n=o(n\min\{1,\theta_n^2\})$ *and* $m_1^{-a}h_{m_1}s_n^2=o(\min\{1,n^{-1}m_1\log(\underline{\phi}_n),\theta_n^2\})$;
(2). $l_n=O(\underline{\phi}_n\tau_{m_2}^2)$;
(3). $\max\{\rho_n,s_n^2\log p\}=o(\min\{n,m_1\log(n\underline{\phi}_n\tau_{m_2}^2)\})$.

The following result presents a situation in which Assumption B.1 holds. For technical convenience, we require the predictors to be independent. It is conjectured that this result may hold in more general settings.

Proposition 3.3. *Suppose that the predictors $X_1,\dots,X_p$ are iid random variables taking values in $[0,1]$, $s_n^2m_2^2\log p=o(n)$, and $\max_{1\le j\le p}\sup_{l\ge1}\|\varphi_{jl}\|_{\sup}<\infty$. Then Assumption B.1 holds with $\rho_n\propto m_2s_n^2\log p$.*

Assumption B.2 holds when we place indifference prior over the models with size not exceeding t_n. To examine Assumption B.3, we again look at a special case. For simplicity, we suppose the setting of Proposition 3.3 holds. Choose $\tau_l^2=l^{-5}$ for $l\ge1$. Suppose $\log p=n^k$ for $k\in(0,4/5)$, $\theta_n\propto1$, $l_n\propto1$, and $a=4$. Let $m_1=\zeta n^{1/5}+c_{1n}$ and $m_2=\zeta n^{1/5}+c_{2n}$, where $\zeta>0$ is constant, $c_{1n}=o((\log n)^r)$, $c_{2n}=o((\log n)^r)$, $c_{1n}\le c_{2n}$, and $r>0$ is a constant. Let $h_m=(\log m)^r$. Choose $\log(\underline{\phi}_n)=n^{k_1}$ with $k_1>k$. It can be shown in this special situation that Assumption B.3 holds if $s_n=o(n^{\min\{1-k,k_1/2\}}(\log n)^{-r})$. Furthermore, if $s_n=O(1)$, then the condition $s_n^2m_2^2\log p=o(n)$ (see Proposition 3.3) implies $k<3/5$. So the growth rate of p is again not exceeding $\exp(O(n^{3/5}))$.

Theorem 3.4. *Suppose $0<t_n<s_n$ and Assumptions B.1–B.3 are satisfied.*

(i). As $n\to\infty$, $\max_{m\in[m_1,m_2]}\sup_{\underline{\phi}_n\le c_1,\dots,c_p\le\bar{\phi}_n}\frac{\max_{\boldsymbol{\gamma}\in T_1(t_n)\cup T_2(t_n)}p(\boldsymbol{\gamma}|\boldsymbol{D}_n)}{\max_{\boldsymbol{\gamma}\in T_0(t_n)}p(\boldsymbol{\gamma}|\boldsymbol{D}_n)}\to0$, in probability.

(ii). Furthermore, suppose Assumption A.3 (4) is satisfied, and there is $\boldsymbol{\gamma} \in T_0(t_n)\backslash\{\emptyset\}$ *and a constant* $b_0 > 0$ *such that for all* $m \in [m_1, m_2]$,

$$\sum_{j\in\boldsymbol{\gamma}^0\backslash\boldsymbol{\gamma}} \|f_j^0\|_j^2 \le b_0 \sum_{j\in\boldsymbol{\gamma}} \|f_j^0\|_j^2. \tag{3.7}$$

Then, as $n \to \infty$, $\max_{m\in[m_1,m_2]} \sup_{\underline{\phi}_n \le c_1,\ldots,c_p \le \bar{\phi}_n} \frac{p(\emptyset|\boldsymbol{D}_n)}{p(\boldsymbol{\gamma}|\boldsymbol{D}_n)} \to 0$, *in probability.*

When the hyperparameter t_n is incorrectly specified as being smaller than the size of the true model, the selected model $\widehat{\boldsymbol{\gamma}}$ cannot be the true model since necessarily $|\widehat{\boldsymbol{\gamma}}| < s_n$. Theorem 3.4 (i) shows that in this false setting, $\widehat{\boldsymbol{\gamma}}$ can be asymptotically nested to the true model with probability approaching one. This means, as n approaches infinity, all the selected components are the significant ones which ought to be included in the model. Here, the result holds uniformly for m and c_j' s within certain ranges, showing insensitivity of the choice of these hyperparameters. Model selection in misspecified models is an important problem which has been considered by [33] in the parametric framework, who proposed novel generalized versions of Akaike information criterion and Bayesian information criterion, and studied their asymptotic property. To the best of our knowledge, Theorem 3.4 is the first theoretical examination of the (nonparametric) function selection approach when the model space is misspecified.

We should mention that in Theorem 3.4 (i), it is possible that $\widehat{\boldsymbol{\gamma}} = \emptyset$ since $\emptyset$ is a natural subset of $\boldsymbol{\gamma}^0$. When $\boldsymbol{\gamma}^0$ is non-null, we expect $\widehat{\boldsymbol{\gamma}}$ to include some significant variables. Theorem 3.4 (ii) says that this is possible if there exists a non-null model that can be separated from the null model. Explicitly, the condition (3.7) says that the functions $\{f_j^0, j \in \boldsymbol{\gamma}\}$ dominate the functions $\{f_j^0, j \in \boldsymbol{\gamma}^0\backslash\boldsymbol{\gamma}\}$, in terms of the corresponding norms $\|\cdot\|_j$'s. This can be interpreted as that the model $\boldsymbol{\gamma}$ includes a larger amount of the information from the true model than its completion $\boldsymbol{\gamma}^0\backslash\boldsymbol{\gamma}$. Theorem 3.4 (ii) says that under this condition, with probability approaching one, $\boldsymbol{\gamma}$ is more preferred than the null. Therefore, $\widehat{\boldsymbol{\gamma}}$ is asymptotically non-null.

3.3. Basis functions

The proposed approach relies on a proper set of orthonormal basis functions $\{\varphi_{jl}, l \ge 1\}$ in $\mathcal{H}_j$ under the inner product $\langle\cdot,\cdot\rangle_j$. In this section we briefly describe how to empirically construct such functions.

Suppose for each $j = 1,\ldots,p$, $\{B_{jl}, l \ge 0\}$ form a set of basis functions in $L^2[0,1]$. Without loss of generality, assume B_{j0} to be the constant function. For example, in empirical study we can choose the trigonometric polynomial basis, i.e., $B_{j0} = 1$, $B_{jl}(x) = \sqrt{2}\cos(2\pi kx)$ if $l = 2k-1$, and $B_{jl}(x) = \sqrt{2}\sin(2\pi kx)$ if $l = 2k$, for integer $k \ge 1$. Other choices such as Legendre's polynomial basis (see [6]) is also examined. One may even choose $\{B_{jl}, l \ge 0\}$ to be non-orthogonal such as B-spline basis. However, B-spline basis requires selection of

knots, while both trigonometric basis and Legendre basis do not have such issue. In numerical study we only consider trigonometric basis and Legendre basis for convenience.

We may choose a sufficiently large integer M with $M < n$. For $j = 1, \ldots, p$ and $1 \le l \le M$, define $\tilde{B}_{jl}$ to be a real-valued function whose value at X_{ji} is $\tilde{B}_{jl}(X_{ji}) = B_{jl}(X_{ji}) - \frac{1}{n}\sum_{i=1}^{n} B_{jl}(X_{ji})$. Define $\mathbf{W}_{ji} = (\tilde{B}_{j1}(X_{ji}), \ldots, \tilde{B}_{jM}(X_{ji}))^T$, and $\hat{\mathbf{\Sigma}}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{W}_{ji}\mathbf{W}_{ji}^T$. Let $\mathbf{A}_j$ be an M by M invertible matrix such that $\mathbf{A}_j^T \hat{\mathbf{\Sigma}}_j \mathbf{A}_j = \mathbf{I}_M$. Write $\mathbf{A}_j = (\mathbf{a}_{j1}, \ldots, \mathbf{a}_{jM})$, where $\mathbf{a}_{jl}$ is the l-th column, an M-vector. Then define φ_{jl} as a real-valued function whose value at X_{ji} is $\varphi_{jl}(X_{ji}) = \mathbf{a}_{jl}^T \mathbf{W}_{ji}$, for $j = 1, \ldots, p$ and $l = 1, \ldots, M$. In the simplest situation where X_{ji}'s are *iid* uniform in [0,1], for $j = 1, \ldots, p$, it can be seen that $\hat{\mathbf{\Sigma}}_j \approx \mathbf{I}_M$, for which we can choose $\mathbf{A}_j = \mathbf{I}_M$, leading to $\varphi_{jl} = \tilde{B}_{jl}$ for $l = 1, \ldots, M$.

Next we heuristically show that the functions φ_{jl}'s approximately form an orthonormal basis. By the law of large numbers, $E\{\varphi_{jl}(X_j)\} \approx \frac{1}{n}\sum_{i=1}^{n} \varphi_{jl}(X_{ji}) = \frac{1}{n}\mathbf{a}_{jl}^T \sum_{i=1}^{n} \mathbf{W}_{ji} = 0$, and $E\{\varphi_{jl}(X_j)\varphi_{jl'}(X_j)\} \approx \frac{1}{n}\sum_{i=1}^{n} \varphi_{jl}(X_{ji})\varphi_{jl'}(X_{ji}) = \mathbf{a}_{jl}^T \hat{\mathbf{\Sigma}} \mathbf{a}_{jl'} = \delta_{ll'}$, for $l, l' = 1, \ldots, M$, where $\delta_{ll'} = 1$ if $l = l'$, and zero otherwise. Thus, $\{\varphi_{jl}, l = 1, \ldots, M\}$ approximately form an orthonormal system. Furthermore, any $f_j \in \mathcal{H}_j$ admits the approximate expansion $f_j(X_{ji}) \approx \sum_{l=0}^{M} \tilde{\beta}_{jl} B_{jl}(X_{ji})$ for some real sequence $\tilde{\beta}_{jl}$. So

$$0 = E\{f_j(X_j)\} \approx \frac{1}{n}\sum_{i=1}^{n} f_j(X_{ji}) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{l=0}^{M} \tilde{\beta}_{jl} B_{jl}(X_{ji}).$$

Therefore, we obtain that

$$\begin{aligned} f_j(X_{ji}) &\approx \sum_{l=0}^{M} \tilde{\beta}_{jl} B_{jl}(X_{ji}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=0}^{M} \tilde{\beta}_{jl} B_{jl}(X_{ji}) \\ &= \sum_{l=1}^{M} \tilde{\beta}_{jl} \tilde{B}_{jl}(X_{ji}) = \tilde{\boldsymbol{\beta}}_j^T \mathbf{W}_{ji} = (\mathbf{A}_j^{-1}\tilde{\boldsymbol{\beta}}_j)^T (\varphi_{j1}(X_{ji}), \ldots, \varphi_{jM}(X_{ji}))^T, \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \ldots, \tilde{\beta}_{jM})^T$. This means that the function f_j can be approximately represented by the φ_{jl}'s for $l = 1, \ldots, M$. Consequently, $\{\varphi_{jl}, l = 1, \ldots, M\}$ approximately form an orthonormal basis in $\mathcal{H}_j$ given that M is large enough.

3.4. Mixtures of g-prior

The results in Sections 3.1 and 3.2 can also be extended to the g-prior setting. Suppose $c_j = c$ for $j = 1, \ldots, p$. We assume c to have prior density $g(c)$, a function of positive values over $(0, \infty)$ satisfying $\int_0^\infty g(c)dc = 1$, i.e., g is a proper prior. Then (2.7) is actually $p(\boldsymbol{\gamma}|c, \mathbf{D}_n)$. The posterior distribution of $\boldsymbol{\gamma}$ is therefore $p_g(\boldsymbol{\gamma}|\mathbf{D}_n) = \int_0^\infty p(\boldsymbol{\gamma}|c, \mathbf{D}_n)g(c)dc$, with the subscript g emphasizing

the g-prior situation. Then we have the following results parallel to Theorems 3.2 and 3.4. The interpretations are similar to those for Theorems 3.2 and 3.4. Their proofs are similar to those in [44], and thus are omitted.

Theorem 3.5. *Suppose Assumptions A.1–A.3 are satisfied. Furthermore, g is proper and, as $n \to \infty$, $\int_0^{\underline{\phi}_n} g(c)dc = o(1)$ and $\int_{\bar{\phi}_n}^{\infty} g(c)dc = o(1)$. Then as $n \to \infty$, $\min_{m\in[m_1,m_2]} p_g(\boldsymbol{\gamma}^0|\boldsymbol{D}_n) \to 1$, in probability.*

Theorem 3.6. *Suppose $0 < t_n < s_n$. Let Assumptions B.1–B.3 be satisfied, and g be proper and supported in $[\underline{\phi}_n, \bar{\phi}_n]$, i.e., $g(c) = 0$ if $c \notin [\underline{\phi}_n, \bar{\phi}_n]$.*

(i). As $n \to \infty$, $\max_{m\in[m_1,m_2]} \frac{\max_{\boldsymbol{\gamma}\in T_1(t_n)\cup T_2(t_n)} p_g(\boldsymbol{\gamma}|\boldsymbol{D}_n)}{\max_{\boldsymbol{\gamma}\in T_0(t_n)} p_g(\boldsymbol{\gamma}|\boldsymbol{D}_n)} \to 0$, in probability.

(ii). If, in addition, Assumption A.3 (4) holds, and there exist a $\boldsymbol{\gamma} \in T_0(t_n)\backslash\{\emptyset\}$ and a constant $b_0 > 0$ such that for all $m \in [m_1, m_2]$, $\sum_{j\in\boldsymbol{\gamma}^0\backslash\boldsymbol{\gamma}} \|f_j^0\|_j^2 \le b_0 \sum_{j\in\boldsymbol{\gamma}} \|f_j^0\|_j^2$. Then as $n \to \infty$, $\max_{m\in[m_1,m_2]} \frac{p_g(\emptyset|\boldsymbol{D}_n)}{p_g(\boldsymbol{\gamma}|\boldsymbol{D}_n)} \to 0$, in probability.

We propose two types of g-priors that generalize the Zellner-Siow prior by [55] and generalize the hyper-g prior by [34]. We name them as the *generalized Zellner-Siow* (GZS) prior and the *generalized hyper-g* (GHG) prior respectively. Let $b, \mu > 0$ be fixed hyperparameters. The GZS prior is defined to have the form

$$g(c) = \frac{p^b}{\Gamma(b)} c^{-b-1} \exp(-p^\mu/c), \tag{3.8}$$

and the GHG prior is defined to have the form

$$g(c) = \frac{\Gamma(p^\mu + 1 + b)}{\Gamma(p^\mu + 1)\Gamma(b)} \cdot \frac{c^{p^\mu}}{(1+c)^{p^\mu+1+b}}. \tag{3.9}$$

We conclude that both GZS and GHG priors can yield posterior consistency. To see this, since we assume $p \gg n$, we have $p^\mu/\sqrt{\log n} \to \infty$ as $n \to \infty$. Let $\underline{\phi}_n = p^\mu/\sqrt{\log n}$ and $\bar{\phi}_n = p^{\mu(\log n)^2}$. It can be directly examined that, as $n \to \infty$, the GZS prior satisfies $\int_0^{\underline{\phi}_n} g(c)dc = (\Gamma(b))^{-1} \int_{p^\mu \underline{\phi}_n^{-1}}^{\infty} c^{a-1} \exp(-c)dc = o(1)$, and $\int_{\bar{\phi}_n}^{\infty} g(c)dc = (\Gamma(b))^{-1} \int_0^{p^\mu \bar{\phi}_n^{-1}} c^{a-1} \exp(-c)dc = o(1)$; the GHG prior satisfies $\int_0^{\underline{\phi}_n} g(c)dc = O((p^\mu+1)^{b-1} \exp(-(p^\mu+1)/(1+\underline{\phi}_n))) = o(1)$ and $\int_{\bar{\phi}_n}^{\infty} g(c)dc = O((p^\mu+1)^b/(1+\bar{\phi}_n)) = o(1)$. Furthermore, suppose $\tau_l^2 = l^{-5}$, $a = 4$, $\log p = n^k$ with $k \in (0, 3/5)$, $\theta_n \propto 1$, $s_n \propto 1$, $l_n \propto 1$, $t_n \propto 1$ with $t_n \ge s_n$, $h_m = (\log n)^r$, and $m_1 = \zeta n^{1/5} + o((\log n)^r)$, $m_2 = \zeta n^{1/5} + o((\log n)^r)$, where ζ and r are positive constants and $r \in (0, 1/2)$. It can be examined directly that the above $\underline{\phi}_n$ and $\bar{\phi}_n$ satisfy the assumptions of Theorem 3.5, implying posterior consistency of the g-prior methods. Clearly, the modes of the GZS and GHG priors are both $p^\mu/(b+1)$ which converges to infinity as n inflates, yielding consistent selection results.

4. Computational details

In this section, sampling details will be provided. Instead of fixing t_n and m, we may place priors over them to make the procedure more flexible. Let $c_j = c$ for $j = 1, \ldots, p$. Assume a g-prior (either GZS or GHG) $g(c)$ for c, and denote the priors for t_n and m by $p(t_n)$ and $p(m)$ respectively. For convenience, we consider the priors $p(\boldsymbol{\gamma}|t_n) \propto w_n^{|\boldsymbol{\gamma}|} I(|\boldsymbol{\gamma}| \leq t_n)$ for some prefixed $w_n > 0$, $p(t_n) = I(1 \leq t_n \leq T_n)$ for some prefixed positive integer T_n, and $p(m) = I(m_1 \leq m \leq m_2)$ for some fixed positive integers m_1 and m_2. In particular, $w_n = 1$ yields flat prior on $\boldsymbol{\gamma}$. For other choices of $p(t_n)$ and $p(m)$, the computational details in this section require corresponding modifications. In practice we suggest $m_2 T_n = n/2$ or $n/3$, which as demonstrated in the simulation studies can yield satisfactory selection results. The specification of T_n may also depend on the prior knowledge on model sparsity. Generally speaking, $m_2 T_n$ cannot exceed $n/2$ to yield unique solution; see [11, 32] as examples in sparse recovery.

It follows by (2.6) that the joint posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2, c, m, t_n)$ is

$$\begin{aligned} & p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2, c, m, t_n | \mathbf{D}_n) \\ \propto\ & \sigma^{-(n+\nu+2)} \exp\left(-\frac{\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|^2 + 1}{2\sigma^2}\right) (\sqrt{2\pi c}\sigma)^{-m|\boldsymbol{\gamma}|} \det(\mathbf{T}_m)^{-|\boldsymbol{\gamma}|/2} \\ & \exp\left(-\frac{\sum_{j \in \boldsymbol{\gamma}} \boldsymbol{\beta}_j^T \mathbf{T}_m^{-1} \boldsymbol{\beta}_j}{2c\sigma^2}\right) \prod_{j \in -\boldsymbol{\gamma}} \delta_{\mathbf{0}}(\boldsymbol{\beta}_j) \cdot p(\boldsymbol{\gamma}|t_n) g(c) p(m) p(t_n), \end{aligned} \tag{4.1}$$

where $\delta_{\mathbf{0}}$ denotes the probability measure concentrating on the m-dimensional zero vector. The MCMC sampling procedure is described as follows. For initial values, let $\boldsymbol{\gamma}^{(0)} = \emptyset$, $\boldsymbol{\beta}^{(0)} = \mathbf{0}$. Let $\sigma^2_{(0)}$ and $c^{(0)}$ be uniformly drawn from some compact subsets of $(0, \infty)$, and $m^{(0)}$ and $t_n^{(0)}$ be drawn from $p(m)$ and $p(t_n)$ respectively. Suppose at the q-th iteration, we have obtained samples $(\boldsymbol{\gamma}^{(q)}, \boldsymbol{\beta}^{(q)}, \sigma^2_{(q)}, c^{(q)}, m^{(q)}, t_n^{(q)})$.

Sampling $(\boldsymbol{\beta}, m, \boldsymbol{\gamma})$. The sampling procedure proceeds in two steps. First, one draws $m^{(q+1)}$ given $c^{(q)}, \sigma^2_{(q)}, t_n^{(q)}$. Second, one draws $(\boldsymbol{\beta}^{(q+1)}, \boldsymbol{\gamma}^{(q+1)})$ given $m^{(q+1)}, c^{(q)}, \sigma^2_{(q)}, t_n^{(q)}$. To complete the first step, by integrating out $\boldsymbol{\beta}$ in (4.1), the conditional distribution of m given $c^{(q)}, \sigma^2_{(q)}, t_n^{(q)}$ is found by

$$\begin{aligned} & p(m | c^{(q)}, \sigma^2_{(q)}, t_n^{(q)}, \mathbf{D}_n) \\ \propto\ & p(m) \sum_{|\boldsymbol{\gamma}| \leq t_n^{(q)}} (c^{(q)})^{-m|\boldsymbol{\gamma}|/2} w_n^{|\boldsymbol{\gamma}|} \det(\mathbf{S}^{(q)}_{\boldsymbol{\gamma},m})^{-1/2} \\ & \exp\left(-\frac{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}}(\mathbf{U}^{(q)}_{\boldsymbol{\gamma},m})^{-1}\mathbf{Z}_{\boldsymbol{\gamma}}^T)\mathbf{Y}}{2\sigma^2_{(q)}}\right), \end{aligned} \tag{4.2}$$

where $\mathbf{U}^{(q)}_{\boldsymbol{\gamma},m} = (c^{(q)})^{-1}\boldsymbol{\Lambda}^{-1}_{\boldsymbol{\gamma},m} + \mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}$, $\mathbf{S}^{(q)}_{\boldsymbol{\gamma},m} = \boldsymbol{\Lambda}_{\boldsymbol{\gamma},m}\mathbf{U}^{(q)}_{\boldsymbol{\gamma},m}$, and $\boldsymbol{\Lambda}_{\boldsymbol{\gamma},m} = \mathrm{diag}(\mathbf{T}_m, \ldots, \mathbf{T}_m)$ with $\mathbf{T}_m$ therein appearing $|\boldsymbol{\gamma}|$ times. In principle, one can draw $m^{(q+1)}$

based on (4.2). However, (4.2) involves a computationally expensive sum which is hard to handle in practice. To overcome this difficulty, we propose an alternative (approximate) way of sampling m. When $\boldsymbol{\gamma}^{(q)} = \emptyset$, draw $m^{(q+1)}$ randomly from $[m_1, m_2]$. When $\boldsymbol{\gamma}^{(q)} \neq \emptyset$, from (4.1) the conditional distribution of m given $\boldsymbol{\gamma}^{(q)}, c^{(q)}, \sigma^2_{(q)}, t_n^{(q)}$ is found by

$$\begin{aligned} & p(m|\boldsymbol{\gamma}^{(q)}, c^{(q)}, \sigma^2_{(q)}, t_n^{(q)}, \mathbf{D}_n) \\ \propto \quad & p(m)(c^{(q)})^{-m|\boldsymbol{\gamma}^{(q)}|/2} \det(\mathbf{S}^{(q)}_{\boldsymbol{\gamma}^{(q)},m})^{-1/2} \\ & \exp\left(-\frac{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}^{(q)}}(\mathbf{U}^{(q)}_{\boldsymbol{\gamma}^{(q)},m})^{-1}\mathbf{Z}^T_{\boldsymbol{\gamma}^{(q)}})\mathbf{Y}}{2\sigma^2_{(q)}}\right). \end{aligned} \tag{4.3}$$

In practice, one can draw m from (4.3) by an Metropolis-Hastings step given the current value $\boldsymbol{\gamma}^{(q)}$, which avoids computing the expensive sum and thus is more efficient. Explicitly, given the current value m_{old}, one draws m_{new} from some proposal distribution $Q(m_{new}|m_{old})$. Then accept m_{new} with probability $\frac{p(m_{new}|\boldsymbol{\gamma},c,\sigma^2,t_n,\mathbf{D}_n)}{p(m_{old}|\boldsymbol{\gamma},c,\sigma^2,t_n,\mathbf{D}_n)} \cdot \frac{Q(m_{old}|m_{new})}{Q(m_{new}|m_{old})}$. The choice of the proposal distribution $Q(m'|m)$ is not unique, but can be made very simple. For instance, when $m = m_1$ (or m_2), one draws m' randomly from $\{m, m+1\}$ (or $\{m, m-1\}$); when $m_1 < m < m_2$, one draws m' randomly from $\{m-1, m, m+1\}$.

To complete the second step, we apply a nontrivial variation of the conventional blockwise technique (see [22, 49]) to sample $\boldsymbol{\beta}_j$'s and γ_j's, given an updated sample $m^{(q+1)}$. Note that the sample $\boldsymbol{\beta}_j^{(q)}$'s from the previous q-th iteration have dimension $m^{(q)}$ which might be different from $m^{(q+1)}$. This phenomenon of different dimensions makes the conventional blockwise sampling approach fail since there is an underlying conflict between the current state of m and the (conditioning) blocks from the previous iteration. Motivated from the "dimension-matching" technique in the reversible jump MCMC approach (see [19]), we propose to modify the $\boldsymbol{\beta}_j^{(q)}$'s to be of dimension $m^{(q+1)}$ to match the current state of m. Specifically, if $m^{(q+1)} < m^{(q)}$, define $\tilde{\boldsymbol{\beta}}_j^{(q)}$ to be an $m^{(q+1)}$-dimensional vector which consists of the first $m^{(q+1)}$ elements of $\boldsymbol{\beta}_j^{(q)}$. If $m^{(q+1)} > m^{(q)}$, then define $\tilde{\boldsymbol{\beta}}_j^{(q)} = ((\boldsymbol{\beta}_j^{(q)})^T, \mathbf{0}^T_{m^{(q+1)}-m^{(q)}})^T$, where $\mathbf{0}_h$ denotes the h-dimensional zero vector. That is, $\tilde{\boldsymbol{\beta}}_j^{(q)}$ is $m^{(q+1)}$-dimensional with the first $m^{(q)}$ elements being exactly the ones of $\boldsymbol{\beta}_j^{(q)}$, and the remaining $m^{(q+1)} - m^{(q)}$ elements being zero. If $m^{(q+1)} = m^{(q)}$, then set $\tilde{\boldsymbol{\beta}}_j^{(q)} = \boldsymbol{\beta}_j^{(q)}$. Repeating the above procedure for all $j = 1, \ldots, p$, one obtains $\tilde{\boldsymbol{\beta}}_j^{(q)}$'s, a "modified" set of samples from the previous stage.

Suppose we have updated samples $(\boldsymbol{\beta}_1^{(q+1)}, \gamma_1^{(q+1)}), \ldots, (\boldsymbol{\beta}_{j-1}^{(q+1)}, \gamma_{j-1}^{(q+1)})$, in which all $\boldsymbol{\beta}_{j'}^{(q+1)}$, for $j' = 1, \ldots, j-1$, are $m^{(q+1)}$-dimensional. Define $\mathbf{b}_{j'} = (\boldsymbol{\beta}_{j'}^{(q+1)}, \gamma_{j'}^{(q+1)})$ for $j' = 1, \ldots, j-1$, $\mathbf{b}_{j'} = (\tilde{\boldsymbol{\beta}}_{j'}^{(q)}, \gamma_{j'}^{(q)})$ for $j' = j+1, \ldots, p$, and

$\mathbf{b}_j = (\boldsymbol{\beta}_j, \gamma_j)$, where $\boldsymbol{\beta}_j$ is an $m^{(q+1)}$-dimensional nominal vector and both $\boldsymbol{\beta}_j$ and γ_j will be updated. For convenience, define $\mathbf{b}_{-j} = \{\mathbf{b}_{j'}, j' = 1, \ldots, p, j' \neq j\}$ to be the conditioning blocks. The full conditional of $\mathbf{b}_j$ given $\mathbf{b}_{-j}$ and other variables highly depends on the size of $\boldsymbol{\gamma}_{-j} = (\gamma_1^{(q+1)}, \ldots, \gamma_{j-1}^{(q+1)}, \gamma_{j+1}^{(q)}, \ldots, \gamma_p^{(q)})$. Specifically, for effective sampling, $|\boldsymbol{\gamma}_{-j}|$ cannot exceed $t_n^{(q)}$ since otherwise the block $\mathbf{b}_j$ will have zero posterior probability. When $|\boldsymbol{\gamma}_{-j}| = t_n^{(q)}$, $\gamma_j^{(q+1)}$ has to be zero since otherwise the conditional probability becomes zero. In this case, one simply sets $\boldsymbol{\beta}_j^{(q+1)} = \mathbf{0}_{m^{(q+1)}}$.

Next we suppose $|\boldsymbol{\gamma}_{-j}| < t_n^{(q)}$. For $j', j = 1, \ldots, p$, define $\mathbf{Z}_{j'}^{(q+1)} = (\boldsymbol{\Phi}_{j'1}, \ldots, \boldsymbol{\Phi}_{j'm^{(q+1)}})$, an n by $m^{(q+1)}$ matrix, and define $\mathbf{Z}_{-j}^{(q+1)} = (\mathbf{Z}_{j'}^{(q+1)}, j' \neq j)$, an n by $m^{(q+1)}(p-1)$ matrix. Similarly, define $\tilde{\boldsymbol{\beta}}_{-j}^{(q)}$ to be the $m^{(q+1)}(p-1)$-vector formed by $\boldsymbol{\beta}_1^{(q+1)}, \ldots, \boldsymbol{\beta}_{j-1}^{(q+1)}, \tilde{\boldsymbol{\beta}}_{j+1}^{(q)}, \ldots, \tilde{\boldsymbol{\beta}}_p^{(q)}$. Let $\mathbf{u}_j = \mathbf{Y} - \mathbf{Z}_{-j}^{(q+1)} \tilde{\boldsymbol{\beta}}_{-j}^{(q)}$. Note $p(\gamma_j, \boldsymbol{\gamma}_{-j}) = 1$. Then we have from (4.1) that

$$\begin{aligned} & p(\boldsymbol{\beta}_j, \gamma_j = 1 | \mathbf{b}_{-j}, \sigma_{(q)}^2, c^{(q)}, t_n^{(q)}, m^{(q+1)}, \mathbf{D}_n) \\ \propto \quad & \exp\left(-\frac{\|\mathbf{u}_j - \mathbf{Z}_j^{(q+1)} \boldsymbol{\beta}_j\|^2}{2\sigma_{(q)}^2}\right) (\sqrt{2\pi c^{(q)}} \sigma_{(q)})^{-m^{(q+1)}} \det(\mathbf{T}_{m^{(q+1)}})^{-1/2} \\ & \exp\left(-\frac{\boldsymbol{\beta}_j^T \mathbf{T}_{m^{(q+1)}}^{-1} \boldsymbol{\beta}_j}{2c^{(q)} \sigma_{(q)}^2}\right) w_n. \end{aligned} \tag{4.4}$$

Integrating out $\boldsymbol{\beta}_j$ in (4.4), one obtains that

$$\begin{aligned} & p(\gamma_j = 1 | \mathbf{b}_{-j}, \sigma_{(q)}^2, c^{(q)}, t_n^{(q)}, m^{(q+1)}, \mathbf{D}_n) \\ \propto \quad & \det(\mathbf{Q}_j^{(q)} \mathbf{T}_{m^{(q+1)}})^{-1/2} (c^{(q)})^{-m^{(q+1)}/2} \\ & \exp\left(-\frac{\|\mathbf{u}_j\|^2 - \mathbf{u}_j^T \mathbf{Z}_j^{(q+1)} (\mathbf{Q}_j^{(q)})^{-1} (\mathbf{Z}_j^{(q+1)})^T \mathbf{u}_j}{2\sigma_{(q)}^2}\right) w_n, \end{aligned} \tag{4.5}$$

where $\mathbf{Q}_j^{(q)} = (c^{(q)})^{-1} \mathbf{T}_{m^{(q+1)}}^{-1} + (\mathbf{Z}_j^{(q+1)})^T \mathbf{Z}_j^{(q+1)}$. Similarly, one obtains from (4.1) that

$$p(\boldsymbol{\beta}_j, \gamma_j = 0 | \mathbf{b}_{-j}, \sigma_{(q)}^2, c^{(q)}, t_n^{(q)}, m^{(q+1)}, \mathbf{D}_n) \propto \exp\left(-\frac{\|\mathbf{u}_j - \mathbf{Z}_j^{(q+1)} \boldsymbol{\beta}_j\|^2}{2\sigma_{(q)}^2}\right) \delta_{\mathbf{0}}(\boldsymbol{\beta}_j). \tag{4.6}$$

Integrating out $\boldsymbol{\beta}_j$ in (4.6) one obtains that

$$p(\gamma_j = 0 | \mathbf{b}_{-j}, \sigma_{(q)}^2, c^{(q)}, t_n^{(q)}, m^{(q+1)}, \mathbf{D}_n) \propto \exp\left(-\frac{\|\mathbf{u}_j\|^2}{2\sigma_{(q)}^2}\right). \tag{4.7}$$

Consequently, from (4.5) and (4.7) we draw $\gamma_j^{(q+1)}$ from

$$p(\gamma_j = 1 | \mathbf{b}_{-j}, \sigma_{(q)}^2, c^{(q)}, t_n^{(q)}, m^{(q+1)}, \mathbf{D}_n) = \frac{1}{1 + \theta_j}, \tag{4.8}$$

where

$$\theta_j = \det(\mathbf{Q}_j^{(q)}\mathbf{T}_{m^{(q+1)}})^{1/2}(c^{(q)})^{m^{(q+1)}/2} \exp\left(-\frac{\mathbf{u}_j^T\mathbf{Z}_j^{(q+1)}(\mathbf{Q}_j^{(q)})^{-1}(\mathbf{Z}_j^{(q+1)})^T\mathbf{u}_j}{2\sigma_{(q)}^2}\right) w_n^{-1}.$$

It can be shown from (4.4) and (4.6) that

$$\begin{aligned}
&\boldsymbol{\beta}_j|\gamma_j^{(q+1)} = 1, \mathbf{b}_{-j}, \sigma_{(q)}^2, c^{(q)}, t_n^{(q)}, m^{(q+1)}, \mathbf{D}_n \\
\sim\ & N\left((\mathbf{Q}_j^{(q)})^{-1}(\mathbf{Z}_j^{(q+1)})^T\mathbf{u}_j, \sigma_{(q)}^2(\mathbf{Q}_j^{(q)})^{-1}\right), \\
& p(\boldsymbol{\beta}_j = \mathbf{0}|\gamma_j^{(q+1)} = 0, \mathbf{b}_{-j}, \sigma_{(q)}^2, c^{(q)}, t_n^{(q)}, m^{(q+1)}, \mathbf{D}_n) = 1, \qquad (4.9)
\end{aligned}$$

from which $\boldsymbol{\beta}_j^{(q+1)}$ is drawn. In the above procedure, finding the matrix product $\mathbf{Z}_{-j}^{(q+1)}\tilde{\boldsymbol{\beta}}_{-j}^{(q)}$ is a time-consuming step. It is possible to avoid computing this matrix product by iteratively using the following relation

$$\mathbf{Z}_{-(j+1)}^{(q+1)}\tilde{\boldsymbol{\beta}}_{-(j+1)}^{(q)} = \mathbf{Z}_{-j}^{(q+1)}\tilde{\boldsymbol{\beta}}_{-j}^{(q)} + \mathbf{Z}_j^{(q+1)}\boldsymbol{\beta}_j^{(q+1)} - \mathbf{Z}_{j+1}^{(q+1)}\tilde{\boldsymbol{\beta}}_{j+1}^{(q)}. \qquad (4.10)$$

In practice, one only needs to compute $\mathbf{Z}_{-1}^{(q+1)}\tilde{\boldsymbol{\beta}}_{-1}^{(q)}$ since the subsequent products can be iteratively updated through (4.10).

The proposed blockwise sampling scheme (4.8) and (4.9) can be viewed as a generalization of [22] from $m = 1$ (without group structure) to general m (with group structure). This generalization is nontrivial because we allow m, the dimension of $\boldsymbol{\beta}_j$, to change across the consecutive iterations. When updating $\mathbf{b}_j$ given the blocks from the previous iteration whose dimensions might be different from the current value of m, we have to apply the "dimension-matching" technique to the previous samples of $\boldsymbol{\beta}_j$'s so that they have the same dimension as the current m. By doing so, one can apply the conventional blockwise techniques to update the blocks consecutively. Note that when m does not change across the iterations, there is no need to use such "dimension-matching" procedure. Furthermore, the proposed blockwise technique can only be used for the constrained situation, i.e., when $|\boldsymbol{\gamma}_{-j}| \le t_n^{(q)}$, which is essentially a constrained version (with group structure) of the conventional blockwise sampling approaches.

Sampling σ^2. From (4.1), it can be easily seen that the full conditional of σ^2 is

$$\begin{aligned}
&\sigma^2|\boldsymbol{\gamma}^{(q+1)}, \boldsymbol{\beta}^{(q+1)}, m^{(q+1)}, c^{(q)}, t_n^{(q)}, \mathbf{D}_n \\
&\sim IG\left(\frac{n+\nu+m^{(q+1)}|\boldsymbol{\gamma}^{(q+1)}|}{2}, \right. \\
&\left.\frac{\|\mathbf{Y}-\mathbf{Z}^{(q+1)}\boldsymbol{\beta}^{(q+1)}\|^2 + (\boldsymbol{\beta}_{\boldsymbol{\gamma}^{(q+1)}}^{(q+1)})^T\boldsymbol{\Lambda}_{\boldsymbol{\gamma}^{(q+1)},m^{(q+1)}}^{-1}\boldsymbol{\beta}_{\boldsymbol{\gamma}^{(q+1)}}^{(q+1)}(c^{(q)})^{-1}+1}{2}\right),
\end{aligned}$$

where $IG(a,b)$ denotes the inverse Gamma distribution. Denote $\sigma^2_{(q+1)}$ as the updated sample.

Sampling c. When $g(c)$ is chosen to be the GZS prior specified as (3.8), we can use a Gibbs sampling step to draw $c^{(q+1)}$. Indeed, the full conditional of c is found to be

$$IG\left(m^{(q+1)}|\boldsymbol{\gamma}^{(q+1)}|/2+b, p^{\mu}+(\boldsymbol{\beta}^{(q+1)}_{\boldsymbol{\gamma}^{(q+1)}})^T\boldsymbol{\Lambda}^{-1}_{\boldsymbol{\gamma}^{(q+1)},m^{(q+1)}}\boldsymbol{\beta}^{(q+1)}_{\boldsymbol{\gamma}^{(q+1)}}/(2\sigma^2_{(q+1)})\right).$$

When $g(c)$ is chosen to be the GHG prior specified as (3.9), we need an Metropolis-Hastings step. Explicitly, the full conditional of c is

$$\begin{aligned} & p(c|\boldsymbol{\gamma}^{(q+1)},\boldsymbol{\beta}^{(q+1)},\sigma^2_{(q+1)},m^{(q+1)},t_n^{(q)},\mathbf{D}_n) \\ \propto\ & c^{-m^{(q+1)}|\boldsymbol{\gamma}^{(q+1)}|/2}\exp\left(-(\boldsymbol{\beta}^{(q+1)}_{\boldsymbol{\gamma}^{(q+1)}})^T\boldsymbol{\Lambda}^{-1}_{\boldsymbol{\gamma}^{(q+1)},m^{(q+1)}}\boldsymbol{\beta}^{(q+1)}_{\boldsymbol{\gamma}^{(q+1)}}/(2c\sigma^2_{(q+1)})\right)g(c). \end{aligned}$$

Write $c=\exp(\kappa)$, then the full conditional of κ is

$$\begin{aligned} & p(\kappa|\boldsymbol{\gamma}^{(q+1)},\boldsymbol{\beta}^{(q+1)},\sigma^2_{(q+1)},m^{(q+1)},t_n^{(q)},\mathbf{D}_n) \\ \propto\ & \exp\Bigg(-\left(m^{(q+1)}|\boldsymbol{\gamma}^{(q+1)}|/2-1\right)\kappa \\ & -(\boldsymbol{\beta}^{(q+1)}_{\boldsymbol{\gamma}^{(q+1)}})^T\boldsymbol{\Lambda}^{-1}_{\boldsymbol{\gamma}^{(q+1)},m^{(q+1)}}\boldsymbol{\beta}^{(q+1)}_{\boldsymbol{\gamma}^{(q+1)}}/(2\exp(\kappa)\sigma^2_{(q+1)})\Bigg)g(\exp(\kappa)). \end{aligned}$$

Given an old value κ_{old}, draw $\kappa_{new}\sim N(\kappa_{old},\sigma^2_{\kappa})$ for some fixed σ^2_{κ}. Then accept κ_{new} with probability

$$\frac{p(\kappa_{new}|\boldsymbol{\gamma}^{(q+1)},\boldsymbol{\beta}^{(q+1)},\sigma^2_{(q+1)},m^{(q+1)},t_n^{(q)},\mathbf{D}_n)}{p(\kappa_{old}|\boldsymbol{\gamma}^{(q+1)},\boldsymbol{\beta}^{(q+1)},\sigma^2_{(q+1)},m^{(q+1)},t_n^{(q)},\mathbf{D}_n)}.$$

Sampling t_n. It is easy to see that the full conditional of t_n is uniform over $[|\boldsymbol{\gamma}^{(q+1)}|,T_n]$, from which we obtain $t_n^{(q+1)}$.

5. Simulation study

In this section we demonstrate the performance of the proposed method through simulated data. Specifically, we compare our Bayesian method based on GZS and GHG priors, denoted as BGZS and BGHG respectively, with the iterative nonparametric independence screening combined with penGAM, denoted as INIS-penGAM, and its greedy modification, denoted as g-INIS-penGAM, both proposed by [12]. Other well-known approaches include the penalized method for additive model (penGAM) proposed by [38], and the iterative sure independence screening (ISIS) combined with SCAD proposed by [13, 16]; see [12] for numerical details.

We adopted two simulation settings considered by [12, 25] in the following examples in which $p=1000$ and $n=400$. We chose somewhat arbitrarily the

hyperparameter $\nu = 6$ in the prior (2.4). In both the GZS and GHG priors defined by (3.8) and (3.9), we chose $b = 0$. To see how sensitive the results are with respect to the choice of μ, we considered difference values of μ. The test functions are defined by

$$f_1(x) = x, \quad f_2(x) = (2x-1)^2, \quad f_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)}, \quad \text{and}$$

$$f_4(x) = 0.1\sin(2\pi x)+0.2\cos(2\pi x)+0.3\sin(2\pi x)^2+0.4\cos(2\pi x)^3+0.5\sin(2\pi x)^3.$$

Example 5.1. We adopted the simulation setting of Example 3 in [12]. Specifically, the data were generated from the additive model $Y = 5f_1(X_1)+3f_2(X_2)+4f_3(X_3)+6f_4(X_4)+\sqrt{1.74}\epsilon$, where $\epsilon \sim N(0,1)$. The covariates were simulated by $X_j = (W_j + \rho U)/(1+\rho)$, $j = 1, \ldots, p$, where W_j's and U are *iid* draws from uniform distribution over $[0,1]$. $\rho = 0$ yields independent X_j's and $\rho = 1$ yields dependent covariates with pairwise correlation 0.5.

Example 5.2. We adopted the simulation setting of Example 4 in [12]. This example is more challenging in that it contains more true functions than Example 5.1. Specifically, the data were generated from the following model

$$\begin{aligned} Y \quad = \quad & f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) \\ & + 1.5f_1(X_5) + 1.5f_2(X_6) + 1.5f_3(X_7) + 1.5f_4(X_8) \\ & + 2f_1(X_9) + 2f_2(X_{10}) + 2f_3(X_{11}) + 2f_4(X_{12}) + \sqrt{0.5184}\epsilon, \end{aligned}$$

where $\epsilon \sim N(0,1)$. The covariates X_j's were generated according to Example 5.1.

Example 5.3. We reconsidered the model of Example 5.1 by adopting a different signal-to-noise ratio. Specifically, the data were generated from $Y = 5f_1(X_1) + 3f_2(X_2) + 4f_3(X_3) + 6f_4(X_4) + \sqrt{3}\epsilon$, where X_j's and ϵ were generated based on the same scheme as Example 5.1.

In Examples 5.1 and 5.2, [12] used five spline basis functions to represent the nonparametric functions. In the present paper we considered both Legendre polynomial basis and trigonometric polynomial basis. In both cases, we chose $m_1 = 4$ and $m_2 = 6$ so that the number of basis functions m varies around 5 to enhance flexibility. We used $\mu = 0.5, 0.6, 0.8$ and $0.8, 0.9, 1.1$ for the above two bases, respectively, to demonstrate the insensitivity of the results. The MCMC algorithm introduced in Section 4 was implemented for posterior sampling. Results were based on 100 replicated data sets. Based on each data, we generated Markov chains with length 4000 for each model parameter. The prior for t_n was chosen as uniform in $\{1, \ldots, T_n\}$. Note in model (2.2) there are at most $m|\boldsymbol{\gamma}|$ nonzero Fourier coefficients. In the present setup, this quantity is upper bounded by mT_n. We chose $T_n = [n/(3m)]$ so that the maximum number of nonzero coefficients does not exceed $n/3$. In [11, 32] it was shown that the number of nonzero coefficients cannot exceed $n/2$ for uniqueness of the solution in sparse recovery. Here we reduced the upper bound to $n/3$ to gain more sparse solutions. For GHG prior, we chose $\sigma_\kappa^2 = 0.2$ for the MH update of κ in sampling c; see Section 4 for detailed description.

Recall that the Fourier coefficient vector $\boldsymbol{\beta}_j$ may change dimension across iterations, i.e., the so-called trans-dimensional problem. The resulting chains may include varying-dimension components. It is well known in the literature that the classic approaches for convergence diagnostics may fail. Following [20], we used the chains of MSE, a natural scalar statistics, to monitor MCMC convergence of the Fourier coefficients, which successfully resolves the trans-dimensional problem. Although we are aware that such scalar statistics cannot guarantee convergence of the full chains, its computational convenience is attractive. Moreover, the scope of the current paper focuses more about the selection and estimation issues, for which monitoring convergence of the MSE chains is believed to be a reasonable strategy. In our study we used Gelman-Rubin's statistics (see [18]) to monitor convergence of the chains relating to MSE and the remaining parameters. Confirming chain convergence, we dropped the first half of the posterior samples as burnings and only used the second half to conduct statistical procedures.

We reported the average number of true positives (TP), the average number of false positives (FP), the prediction errors (PE) based on BGZS and BGHG, and compared them with INIS and g-INIS. Marginal inclusion rule is adopted to select the model. That is, the jth variable is selected if its posterior exclusion probability $P_j = 1 - p(\gamma_j = 1|\mathbf{D}_n) \leq \widehat{p}$ for some quantity $\widehat{p} \in (0,1)$. We chose $\widehat{p} = 0.5$ to yield median probability models; see [1]. The TP/FP is the number of true/false inclusions in the selected model. The PE was calculated as $\sum_{q=1}^{Q} \|\mathbf{Y} - \widehat{\mathbf{Y}}^{(q)}\|^2/(nQ)$, where $\widehat{\mathbf{Y}}^{(q)} = \mathbf{Z}^{(q)}\boldsymbol{\beta}^{(q)}$ is the fitted response value obtained from the qth iteration. In other words, PE is the average value of the mean square errors (MSE) along with the iterations.

Results on TP, FP and PE using BGZS and BGHG were summarized in Tables 1–2 and Tables 3–4, based on Legendre polynomial basis and trigonometric polynomial basis, respectively. Results on INIS and g-INIS were directly summarized from [12]. In Example 5.1, we observed that, for both bases, BGZS and BGHG perform equally well as INIS and g-INIS in terms of TP, but perform better in terms of FP and PE.

In Example 5.2 where Legendre polynomial basis was used, both Bayesian approaches perform better than INIS and g-INIS. Specifically, when $\rho = 1$ and $\mu = 0.6$, both BGZS and BGHG yield larger TP, smaller PE, and comparable FP; when $\mu = 0.8$, both BGZS and BGHG yield smaller FP and PE, and comparable TP.

In Example 5.2 where trigonometric basis was used, the performance is not as good as using Legendre polynomial basis, but is still satisfactory. Specifically, when $\rho = 1$ and $\mu = 0.8$, both BGZS and BGHG yield slightly larger TP and FP than INIS and g-INIS (implying less conservative selection results), and when $\mu = 1.1$, both methods yield slightly smaller TP and FP (implying more conservative selection results); when $\rho = 0$, $\mu = 0.8$ or 0.9, both BGZS and BGHG can select all the significant variables though they yield slightly larger FP. In all the cases, the proposed Bayesian methods yield smaller PE.

TABLE 1
Simulation results of Example 5.1 using Legendre polynomial basis

ρ	Method		TP	FP	PE
0	INIS		4.00 (0.00)	2.58 (2.24)	3.02 (0.34)
	g-INIS		4.00 (0.00)	0.67 (0.75)	2.92 (0.30)
	BGZS	$\mu = 0.5$	4.00 (0.00)	0.03 (0.17)	2.25 (0.20)
		$\mu = 0.6$	4.00 (0.00)	0.02 (0.14)	2.25 (0.16)
		$\mu = 0.8$	4.00 (0.00)	0.03 (0.17)	2.23 (0.17)
	BGHG	$\mu = 0.5$	4.00 (0.00)	0.03 (0.17)	2.25 (0.20)
		$\mu = 0.6$	4.00 (0.00)	0.02 (0.14)	2.25 (0.16)
		$\mu = 0.8$	4.00 (0.00)	0.03 (0.17)	2.24 (0.17)
1	INIS		3.98 (0.00)	15.76 (6.72)	2.97 (0.39)
	g-INIS		4.00 (0.00)	0.98 (1.49)	2.61 (0.26)
	BGZS	$\mu = 0.5$	3.99 (0.10)	0.06 (0.28)	2.02 (0.16)
		$\mu = 0.6$	3.99 (0.10)	0.05 (0.22)	2.00 (0.15)
		$\mu = 0.8$	3.99 (0.10)	0.05 (0.22)	2.04 (0.15)
	BGHG	$\mu = 0.5$	3.98 (0.14)	0.08 (0.30)	2.02 (0.16)
		$\mu = 0.6$	3.99 (0.10)	0.06 (0.24)	2.00 (0.15)
		$\mu = 0.8$	3.99 (0.10)	0.05 (0.22)	2.04 (0.15)

TABLE 2
Simulation results of Example 5.2 using Legendre polynomial basis

ρ	Method		TP	FP	PE
0	INIS		11.97 (0.00)	3.22 (1.49)	0.97 (0.11)
	g-INIS		12.00 (0.00)	0.73 (0.75)	0.91 (0.10)
	BGZS	$\mu = 0.5$	11.98 (0.14)	0.74 (1.00)	0.60 (0.05)
		$\mu = 0.6$	11.98 (0.14)	0.54 (0.86)	0.59 (0.05)
		$\mu = 0.8$	11.98 (0.14)	0.41 (0.65)	0.60 (0.05)
	BGHG	$\mu = 0.5$	11.98 (0.14)	0.70 (0.93)	0.60 (0.05)
		$\mu = 0.6$	11.98 (0.14)	0.58 (0.90)	0.59 (0.05)
		$\mu = 0.8$	11.98 (0.14)	0.44 (0.67)	0.60 (0.05)
1	INIS		10.01 (1.49)	15.56 (0.93)	1.03 (0.13)
	g-INIS		10.78 (0.75)	1.08 (1.49)	0.87 (0.11)
	BGZS	$\mu = 0.5$	10.75 (0.80)	1.25 (1.30)	0.54 (0.05)
		$\mu = 0.6$	10.92 (0.69)	1.08 (1.29)	0.54 (0.05)
		$\mu = 0.8$	10.76 (0.79)	0.88 (1.27)	0.54 (0.05)
	BGHG	$\mu = 0.5$	10.74 (0.75)	1.13 (1.20)	0.54 (0.05)
		$\mu = 0.6$	10.86 (0.72)	1.10 (1.18)	0.54 (0.05)
		$\mu = 0.8$	10.72 (0.80)	0.82 (1.13)	0.54 (0.05)

The above results are not sensitive to the choice of μ, though certain μ may yield slightly better performance. Due to the essentially different basis structures, the feasible ranges of μ should be slightly different. We found that, at least in the above examples, $\mu \in [0.5, 0.8]$ and $\mu \in [0.8, 1.1]$ are feasible ranges for Legendre polynomial basis and trigonometric polynomial basis. Any choice of μ within these ranges can provide satisfactory results. Values outside the ranges may slightly lower the level of accuracy.

In Example 5.3, the values of m_1, m_2, μ and the prior for t_n were chosen to be the same as in Example 5.1. Results are summarized in Tables 5 and 6, showing that when signal-to-noise ratio becomes smaller, our method still yields satisfactory performance in terms of model selection and estimation, although the results are slightly worse than the stronger signal-to-noise ratio situation as summarized in Tables 1 and 3.

TABLE 3
Simulation results of Example 5.1 using trigonometric polynomial basis

ρ	Method		TP	FP	PE
0	INIS		4.00 (0.00)	2.58 (2.24)	3.02 (0.34)
	g-INIS		4.00 (0.00)	0.67 (0.75)	2.92 (0.30)
	BGZS	$\mu = 0.8$	4.00 (0.00)	0.04 (0.19)	2.07 (0.14)
		$\mu = 0.9$	4.00 (0.00)	0.06 (0.24)	2.07 (0.15)
		$\mu = 1.1$	4.00 (0.00)	0.02 (0.14)	2.09 (0.17)
	BGHG	$\mu = 0.8$	4.00 (0.00)	0.04 (0.19)	2.07 (0.14)
		$\mu = 0.9$	4.00 (0.00)	0.06 (0.24)	2.07 (0.15)
		$\mu = 1.1$	4.00 (0.00)	0.02 (0.14)	2.09 (0.17)
1	INIS		3.98 (0.00)	15.76 (6.72)	2.97 (0.39)
	g-INIS		4.00 (0.00)	0.98 (1.49)	2.61 (0.26)
	BGZS	$\mu = 0.8$	4.00 (0.00)	0.08 (0.44)	1.76 (0.15)
		$\mu = 0.9$	4.00 (0.00)	0.04 (0.20)	1.78 (0.12)
		$\mu = 1.1$	4.00 (0.00)	0.00 (0.00)	1.76 (0.13)
	BGHG	$\mu = 0.8$	4.00 (0.00)	0.10 (0.46)	1.76 (0.15)
		$\mu = 0.9$	4.00 (0.00)	0.04 (0.20)	1.78 (0.12)
		$\mu = 1.1$	4.00 (0.00)	0.00 (0.00)	1.76 (0.13)

TABLE 4
Simulation results of Example 5.2 using trigonometric polynomial basis

ρ	Method		TP	FP	PE
0	INIS		11.97 (0.00)	3.22 (1.49)	0.97 (0.11)
	g-INIS		12.00 (0.00)	0.73 (0.75)	0.91 (0.10)
	BGZS	$\mu = 0.8$	12.00 (0.00)	1.22 (1.34)	0.54 (0.05)
		$\mu = 0.9$	12.00 (0.00)	1.24 (1.27)	0.54 (0.06)
		$\mu = 1.1$	11.88 (0.32)	0.34 (0.77)	0.58 (0.05)
	BGHG	$\mu = 0.8$	12.00 (0.00)	1.16 (1.40)	0.54 (0.05)
		$\mu = 0.9$	12.00 (0.00)	1.10 (1.01)	0.54 (0.05)
		$\mu = 1.1$	11.88 (0.33)	0.30 (0.68)	0.58 (0.05)
1	INIS		10.01 (1.49)	15.56 (0.93)	1.03 (0.13)
	g-INIS		10.78 (0.75)	1.08 (1.49)	0.87 (0.11)
	BGZS	$\mu = 0.8$	10.86 (0.67)	2.18 (1.81)	0.44 (0.05)
		$\mu = 0.9$	10.76 (0.82)	1.34 (1.56)	0.47 (0.05)
		$\mu = 1.1$	10.46 (0.86)	0.50 (0.81)	0.53 (0.05)
	BGHG	$\mu = 0.8$	10.88 (0.69)	2.06 (1.81)	0.44 (0.05)
		$\mu = 0.9$	10.68 (0.82)	1.58 (1.75)	0.47 (0.05)
		$\mu = 1.1$	10.44 (0.84)	0.48 (0.76)	0.53 (0.05)

6. A real data example

We apply our method to Near Infrared (NIR) data from the R package "chemometrics." The NIR data contains $n = 166$ alcoholic fermentation mashes of different feedstock (rye, wheat and corn). There are $p = 235$ predictor variables containing the first derivatives of near infrared spectroscopy (NIR) absorbance values at 1115–2285 nm and two response variables containing glucose and ethanol concentrations (in g/L) respectively. This data was recently considered by [10] who used an approximated posterior likelihood approach. Each of the 235 variables is believed to be nonlinearly associated with the response variable, and hence, a nonparametric model with additive components was fitted. In our Bayesian selection method we chose $m_1 = 4$ and $m_2 = 6$, $T_n = [n/(3 * m_2)] = 9$, i.e., the dimension of the target models does not ex-

TABLE 5
Simulation results of Example 5.3 using Legendre polynomial basis

ρ	Method		TP	FP	PE
0	BGZS	$\mu = 0.5$	4.00 (0.00)	0.10 (0.36)	3.50 (0.31)
		$\mu = 0.6$	4.00 (0.00)	0.16 (0.42)	3.47 (0.23)
		$\mu = 0.8$	4.00 (0.00)	0.16 (0.55)	3.49 (0.31)
	BGHG	$\mu = 0.5$	4.00 (0.00)	0.10 (0.36)	3.50 (0.31)
		$\mu = 0.6$	4.00 (0.00)	0.16 (0.42)	3.47 (0.23)
		$\mu = 0.8$	4.00 (0.00)	0.14 (0.54)	3.49 (0.31)
1	BGZS	$\mu = 0.5$	3.84 (0.37)	0.92 (1.96)	3.16 (0.36)
		$\mu = 0.6$	3.94 (0.24)	0.34 (1.04)	3.15 (0.27)
		$\mu = 0.8$	3.78 (0.42)	0.06 (0.31)	3.31 (0.28)
	BGHG	$\mu = 0.5$	3.84 (0.37)	0.98 (2.16)	3.17 (0.35)
		$\mu = 0.6$	3.94 (0.24)	0.46 (1.64)	3.14 (0.29)
		$\mu = 0.8$	3.78 (0.42)	0.06 (0.24)	3.31 (0.28)

TABLE 6
Simulation results of Example 5.3 using Fourier basis

ρ	Method		TP	FP	PE
0	BGZS	$\mu = 0.5$	4.00 (0.00)	0.96 (1.49)	3.12 (0.32)
		$\mu = 0.6$	4.00 (0.00)	0.70 (1.56)	3.20 (0.34)
		$\mu = 0.8$	4.00 (0.00)	0.10 (0.30)	3.30 (0.23)
	BGHG	$\mu = 0.5$	4.00 (0.00)	0.96 (1.60)	3.11 (0.34)
		$\mu = 0.6$	4.00 (0.00)	1.00 (2.10)	3.17 (0.34)
		$\mu = 0.8$	4.00 (0.00)	0.10 (0.30)	3.32 (0.25)
1	BGZS	$\mu = 0.5$	3.96 (0.20)	2.66 (2.70)	2.49 (0.32)
		$\mu = 0.6$	3.92 (0.27)	1.16 (2.02)	2.66 (0.31)
		$\mu = 0.8$	3.96 (0.20)	0.16 (0.54)	2.99 (0.28)
	BGHG	$\mu = 0.5$	3.94 (0.24)	3.04 (2.79)	2.45 (0.31)
		$\mu = 0.6$	3.92 (0.27)	1.24 (2.08)	2.66 (0.33)
		$\mu = 0.8$	3.96 (0.20)	0.14 (0.45)	2.99 (0.27)

ceed 9. All the hyperparameters were chosen to be the same as the simulation setup in Section 5. Our BGHG and BGZS approaches both select five variables. Interestingly, we found that our selected model includes the model selected by [10] using an approximation method. Thus, our method yields a bit more conservative selection results.

7. Conclusions

A fully Bayesian approach is proposed to handle the ultrahigh-dimensional nonparametric additive models, and the theoretical properties are carefully studied. The numerical results demonstrate satisfactory performance of the method, in terms of selection and estimation accuracy. The method can achieve high level accuracy in both Legendre polynomial basis and trigonometric polynomial basis. Therefore, basis selection is not a critically important issue for the proposed approach, though, to make the approach highly accurate, the choice of the hyperparameter μ in the proposed g-priors should be slightly different in using different bases. The numerical findings suggest us to use $\mu \in [0.5, 0.8]$ and $\mu \in [0.8, 1.1]$ for Legendre polynomial basis and trigonometric polynomial basis, respectively. The values outside these ranges are found to merely slightly lower

the accuracy within an acceptable range. Besides, the choice of m may affect selection accuracy which needs to be done carefully in practical applications.

The theory and algorithm proposed in this paper can be adapted for the usual group selection problems where the model is linear and the model coefficients are grouped. The proposed method in this paper can be adapted without much difficulty to this setting.

Based on Laplace type approximations, [10] proposed an MCMC free selection approach in high-dimensional additive models. The method of [10] approximates the posterior likelihood while our approach directly handles the posterior likelihood. And hence the difference between the two approaches is mainly due to the approximation error, though it is believed that the approximation error asymptotically vanishes.

As an alternative method, [1] proposed median probability selection method. The median probability selection procedure is essentially a marginal selection method in the sense that it marginally determines whether or not a variable is included based on its marginal posterior probability. The variable is included if its marginal posterior probability is greater than half. Posterior model consistency says that the true model has posterior probability tending to one. Therefore, the variables included in the true model all have posterior probabilities greater than half. This means that posterior model selection is a stronger result than median probability selection. In many situations, as a stronger result, posterior consistency requires stronger conditions than median probability selection. And hence, it might be more realistic to explore the latter, which is another future extension of the current paper.

Acknowledgement

Zuofeng Shang was a postdoctorate researcher supported by NSF-DMS 0808864, NSF-EAGER 1249316, a gift from Microsoft, and a gift from Google. The research of Ping Li is also partially supported by NSF-III-1360971, NSF-Bigdata-1419210, ONR-N00014-13-1-0764, and AFOSR-FA9550-13-1-0137.

Appendix: Proofs

To prove Theorem 3.2, we need the following preliminary lemma. The proof is similar to that of Lemma 1 in [44] and thus is omitted.

Lemma 1. *Suppose* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$ *is independent of* $\mathbf{Z}_j$*'s. Furthermore,* $m_2 \leq n = o(p)$.

(i). Let $\boldsymbol{\nu}_{\boldsymbol{\gamma},m}$ *be an* n*-dimensional vector indexed by* $\boldsymbol{\gamma} \in \mathcal{S}$*, a subset of the model space, and integer* $1 \leq m \leq m_2$*. Adopt the convention that* $\boldsymbol{\nu}_{\boldsymbol{\gamma},m}^T \boldsymbol{\epsilon} / \|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\| = 0$ *when* $\boldsymbol{\nu}_{\boldsymbol{\gamma},m} = 0$*. Let* $\#\mathcal{S}$ *denote the cardinality of* $\mathcal{S}$ *with* $\#\mathcal{S} \geq 2$*. Then*

$$\max_{1 \leq m \leq m_2} \max_{\boldsymbol{\gamma} \in \mathcal{S}} \frac{|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}^T \boldsymbol{\epsilon}|}{\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|} = O_P\left(\sqrt{\log(m_2 \#\mathcal{S})}\right). \tag{8.1}$$

In particular, let $\boldsymbol{\nu}_{\boldsymbol{\gamma},m} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{\gamma}})\boldsymbol{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}$ *for* $\boldsymbol{\gamma} \in S_2(t_n)$*, we have*

$$\max_{1\le m\le m_2}\max_{\boldsymbol{\gamma}\in S_2(t_n)}\frac{|\boldsymbol{\nu}^T_{\boldsymbol{\gamma},m}\boldsymbol{\epsilon}|}{\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|} = O_P(\sqrt{\log m_2 + t_n\log p}) = O_P(\sqrt{t_n\log p}). \tag{8.2}$$

(ii). For any fixed $\alpha > 4$*,*

$$\lim_{n\to\infty} P\left(\max_{1\le m\le m_2}\max_{\boldsymbol{\gamma}\in S_1(t_n)}\boldsymbol{\epsilon}^T(\boldsymbol{P}_{\boldsymbol{\gamma}} - \boldsymbol{P}_{\boldsymbol{\gamma}^0})\boldsymbol{\epsilon}/(|\boldsymbol{\gamma}| - s_n) \le \alpha\sigma_0^2\log p\right) = 1.$$

(iii). Adopt the convention that $\boldsymbol{\epsilon}^T\boldsymbol{P}_{\boldsymbol{\gamma}}\boldsymbol{\epsilon}/|\boldsymbol{\gamma}| = 0$ *when* γ *is null. Then for any fixed* $\alpha > 4$*,*

$$\lim_{n\to\infty} P\left(\max_{1\le m\le m_2}\max_{\boldsymbol{\gamma}\in S_2(t_n)}\boldsymbol{\epsilon}^T\boldsymbol{P}_{\boldsymbol{\gamma}}\boldsymbol{\epsilon}/|\boldsymbol{\gamma}| \le \alpha\sigma_0^2\log p\right) = 1.$$

Proof of Proposition 3.1

Let $C_\varphi = \max_{1\le j\le p}\sup_{l\ge 1}\|\varphi_{jl}\|_{\sup}$. We first show that (3.3) holds with $\frac{1}{n}\mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}$ therein replaced with $E\{\frac{1}{n}\mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}\}$. Then we show (3.3) by using concentration inequalities which establish sharp approximations between $\frac{1}{n}\mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}$ and $E\{\frac{1}{n}\mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}\}$.

For any $\mathbf{a}_j = (a_{j1},\ldots,a_{jm})^T$, $j = 1,\ldots,p$, note $\mathbf{Z}_j\mathbf{a}_j = \sum_{l=1}^m a_{jl}\boldsymbol{\Phi}_{jl}$. Define $\mathbf{a}_{\boldsymbol{\gamma}}$ to be the $m|\boldsymbol{\gamma}|$-vector formed by $\mathbf{a}_j$'s with $j\in\boldsymbol{\gamma}$. Therefore, we obtain that

$$\begin{aligned}\mathbf{a}^T_{\boldsymbol{\gamma}}E\{\mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}\}\mathbf{a}_{\boldsymbol{\gamma}} &= E\left\{\left(\sum_{j\in\boldsymbol{\gamma}}\mathbf{Z}_j\mathbf{a}_j\right)^T\left(\sum_{j\in\boldsymbol{\gamma}}\mathbf{Z}_j\mathbf{a}_j\right)\right\}\\ &= \sum_{j\in\boldsymbol{\gamma}}E\{\mathbf{a}_j^T\mathbf{Z}_j^T\mathbf{Z}_j\mathbf{a}_j\} + \sum_{\substack{j,j'\in\boldsymbol{\gamma}\\ j\ne j'}}E\{\mathbf{a}_j^T\mathbf{Z}_j^T\mathbf{Z}_{j'}\mathbf{a}_{j'}\}.\end{aligned}$$

Since φ_{jl}'s are orthonormal in $\mathcal{H}_j$, $E\{\mathbf{a}_j^T\mathbf{Z}_j^T\mathbf{Z}_j\mathbf{a}_j\} = nE\{(\sum_{l=1}^m a_{jl}\varphi_{jl}(X_{ji}))^2\} = n\sum_{l=1}^m a_{jl}^2$. On the other hand, for any $j,j'\in\boldsymbol{\gamma}$, $j\ne j'$, $|E\{\mathbf{a}_j^T\mathbf{Z}_j^T\mathbf{Z}_{j'}\mathbf{a}_{j'}\}| = n|E\{\sum_{l=1}^m a_{jl}\varphi_{jl}(X_{ji})\sum_{l=1}^m a_{j'l}\varphi_{j'l}(X_{j'i})\}| \le n\rho(|j-j'|)\sqrt{\sum_{l=1}^m a_{jl}^2}\sqrt{\sum_{l=1}^m a_{j'l}^2}$. Therefore, by Cauchy's inequality

$$\begin{aligned}&\left|\sum_{\substack{j,j'\in\boldsymbol{\gamma}\\ j\ne j'}}E\{\mathbf{a}_j^T\mathbf{Z}_j^T\mathbf{Z}_{j'}\mathbf{a}_{j'}\}\right|\\ \le\ & n\sum_{\substack{j,j'\in\boldsymbol{\gamma}\\ j\ne j'}}\rho(|j-j'|)\sqrt{\sum_{l=1}^m a_{jl}^2}\sqrt{\sum_{l=1}^m a_{j'l}^2}\end{aligned}$$

$$
\begin{aligned}
&= n\sum_{r=1}^{\infty}\rho(r)\sum_{j\in\boldsymbol{\gamma}}\sqrt{\sum_{l=1}^{m}a_{jl}^2}\sum_{j'\in\boldsymbol{\gamma},|j-j'|=r}\sqrt{\sum_{l=1}^{m}a_{j'l}^2}\\
&\le n\sum_{r=1}^{\infty}\rho(r)\sqrt{\sum_{j\in\boldsymbol{\gamma}}\sum_{l=1}^{m}a_{jl}^2}\sqrt{\sum_{j\in\boldsymbol{\gamma}}\left(\sum_{j'\in\boldsymbol{\gamma},|j-j'|=r}\sqrt{\sum_{l=1}^{m}a_{j'l}^2}\right)^2}\\
&\le n\sum_{r=1}^{\infty}\rho(r)\sqrt{\sum_{j\in\boldsymbol{\gamma}}\sum_{l=1}^{m}a_{jl}^2}\sqrt{2\sum_{j\in\boldsymbol{\gamma}}\sum_{j'\in\boldsymbol{\gamma},|j-j'|=r}\sum_{l=1}^{m}a_{j'l}^2}\\
&\le 2n\sum_{r=1}^{\infty}\rho(r)\sum_{j\in\boldsymbol{\gamma}}\sum_{l=1}^{m}a_{jl}^2.
\end{aligned}
$$

Therefore, for any $m\in[m_1,m_2]$ and $\boldsymbol{\gamma}\neq\emptyset$,

$$
1-2\sum_{r=1}^{\infty}\rho(r)\le\lambda_-\left(E\{\frac{1}{n}\mathbf{Z}_{\boldsymbol{\gamma}}^T\mathbf{Z}_{\boldsymbol{\gamma}}\}\right)\le\lambda_+\left(E\{\frac{1}{n}\mathbf{Z}_{\boldsymbol{\gamma}}^T\mathbf{Z}_{\boldsymbol{\gamma}}\}\right)\le 1+2\sum_{r=1}^{\infty}\rho(r). \quad (8.3)
$$

Next we look at the difference $\boldsymbol{\Delta}=\frac{1}{n}(\mathbf{Z}_{\boldsymbol{\gamma}}^T\mathbf{Z}_{\boldsymbol{\gamma}}-E\{\mathbf{Z}_{\boldsymbol{\gamma}}^T\mathbf{Z}_{\boldsymbol{\gamma}}\})$. The representative entry is

$$
\frac{1}{n}\sum_{i=1}^{n}[\varphi_{jl}(X_{ji})\varphi_{j'l'}(X_{j'i})-E\{\varphi_{jl}(X_{ji})\varphi_{j'l'}(X_{j'i})\}],
$$

for $j,j'\in\boldsymbol{\gamma}$, and $l,l'=1,\dots,m$. Since φ_{jl}'s are uniformly bounded by C_φ, fixing $C>0$ such that $C^2>8C_\varphi^4$, by Hoeffding's inequality,

$$
\begin{aligned}
&P\left(\max_{\substack{j,j'=1,\dots,p\\ l,l'=1,\dots,m_2}}\left|\sum_{i=1}^{n}[\varphi_{jl}(X_{ji})\varphi_{j'l'}(X_{j'i})-E\{\varphi_{jl}(X_{ji})\varphi_{j'l'}(X_{j'i})\}]\right|\right.\\
&\left.\ge C\sqrt{n\log p}\right)\\
\le\ & 2\sum_{j,j'=1}^{p}\sum_{l,l'=1}^{m_2}2\exp\left(-\frac{2C^2n\log p}{4nC_\varphi^4}\right)\le 2p^{4-C^2/(2C_\varphi^4)}\to 0,\ \text{as } n\to\infty.
\end{aligned}
$$

Therefore,

$$
\max_{\substack{j,j'=1,\dots,p\\ l,l'=1,\dots,m_2}}\left|\sum_{i=1}^{n}[\varphi_{jl}(X_{ji})\varphi_{j'l'}(X_{j'i})-E\{\varphi_{jl}(X_{ji})\varphi_{j'l'}(X_{j'i})\}]\right|=O_P(\sqrt{n\log p}).
$$

Denote $\Delta_{j,l;j',l'}$ to be the $(j,l;j',l')$-th entry of $\boldsymbol{\Delta}$. By [23], with probability approaching one, for any $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}|\le 2t_n$, and $m\in[m_1,m_2]$, the spectral norm of $\boldsymbol{\Delta}$ is upper bounded by

$$
\|\boldsymbol{\Delta}\|_{\text{spectral}}\le\max_{j',l'}\sum_{j\in\boldsymbol{\gamma},1\le l\le m}|\Delta_{j,l;j',l'}|\le C'\frac{t_n^2m_2^2\log p}{n},
$$

for some fixed large $C' > 0$. That is, when $n, p \to \infty$,

$$\max_{|\boldsymbol{\gamma}| \le 2t_n} \max_{m \in [m_1, m_2]} \|\boldsymbol{\Delta}\|_{\text{spectral}} \le C' \frac{t_n^2 m_2^2 \log p}{n} = o(1).$$

By Weyl's inequality on eigenvalues (see [23]) and by (8.3), one can properly choose a small $c_0 > 0$ to satisfy (3.3), which completes the proof. Using similar proofs of Proposition 2.1 in [42], it can be shown that (3.3) implies Assumption A.1. The details are straightforward and thus are omitted.

Proof of Theorem 3.2

Denote $\boldsymbol{\beta}_j^0 = (\beta_{j1}^0, \ldots, \beta_{jm}^0)^T$ for $j = 1, \ldots, p$. Define $k_n = \sum_{j \in \boldsymbol{\gamma}^0} \|\boldsymbol{\beta}_j^0\|^2$ and $\psi_n = \min_{j \in \boldsymbol{\gamma}^0} \|\boldsymbol{\beta}_j^0\|$. Before giving the proof of Theorem 3.2, we should mention that Assumption A.3 is actually equivalent to the following Assumption A.4 which assumes the growing rates on terms involving the Fourier coefficients of the partial Fourier series, i.e., k_n and ψ_n. The difference between Assumptions A.3 and A.4 is that l_n and θ_n in the former are replaced with k_n and ψ_n in the latter, respectively. This modified assumption is easier to use in technical proofs.

Assumption A.4. *There exists a positive sequence* $\{h_m, m \ge 1\}$ *such that, as* $m, m_1, m_2 \to \infty$, $h_m \to \infty$, $m^{-a} h_m$ *decreasingly converges to zero,* $m h_m$ *increasingly converges to* ∞, *and* $\sum_{m_1 \le m \le m_2} 1/h_m = o(1)$. *Furthermore, the sequences* $m_1, m_2, h_m, s_n, t_n, \psi_n, k_n, \underline{\phi}_n, \bar{\phi}_n$ *satisfy*

(1). $m_2 h_{m_2} s_n = o(n \min\{1, \psi_n^2\})$ *and* $m_1^{-a} h_{m_1} s_n^2 = o(\min\{1, n^{-1} m_1 \log(\underline{\phi}_n), \psi_n^2, \psi_n^4\})$;
(2). $t_n \ge s_n$ *and* $t_n \log p = o(n \log(1 + \min\{1, \psi_n^2\}))$;
(3). $k_n = O(\underline{\phi}_n \tau_{m_2}^2)$ *and* $\log p = o(m_1 \log(n \underline{\phi}_n \tau_{m_2}^2))$;
(4). $m_2 s_n \log(1 + n\bar{\phi}_n) = o(n \log(1 + \min\{1, \psi_n^2\}))$.

To see the equivalence, it can be directly shown by (3.1) that uniformly for $m \in [m_1, m_2]$

$$l_n - k_n = \sum_{j \in \boldsymbol{\gamma}^0} \sum_{l \ge m+1} |\beta_{jl}^0|^2 \le C_\beta s_n m_1^{-a}. \tag{8.4}$$

On the other hand, for any $j \in \boldsymbol{\gamma}^0$ and any $m \in [m_1, m_2]$, we have $\|f_j^0\|_j^2 = \sum_{l=1}^m |\beta_{jl}^0|^2 + \sum_{l=m+1}^\infty |\beta_{jl}^0|^2 \le \sum_{l=1}^m |\beta_{jl}^0|^2 + C_\beta m_1^{-a}$ and, obviously, $\|f_j^0\|_j^2 \ge \sum_{l=1}^m |\beta_{jl}^0|^2$, which lead to $\psi_n^2 \le \theta_n^2 \le \psi_n^2 + C_\beta m_1^{-a}$. Therefore,

$$0 \le \theta_n^2 - \psi_n^2 \le C_\beta m_1^{-a}. \tag{8.5}$$

By (8.4) and (8.5) and direct examinations, it can be verified that Assumption A.4 is equivalent to Assumption A.3. We will prove the desired theorem based on the equivalent Assumptions A.1, A.2 and A.4.

Throughout the entire section of proof, we use "w.p.a.1" to mean "with probability approaching one". Using the trivial fact

$$p(\boldsymbol{\gamma}^0|\mathbf{D}_n) = \frac{1}{1+\sum_{\boldsymbol{\gamma}\neq\boldsymbol{\gamma}^0}\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\boldsymbol{\gamma}^0|\mathbf{D}_n)}},$$

to obtain the desired result it is sufficient to show $\sum_{\boldsymbol{\gamma}\neq\boldsymbol{\gamma}^0}\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\boldsymbol{\gamma}^0|\mathbf{D}_n)}$ approaches zero in probability. For any $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| \leq t_n$, consider the following decomposition

$$\begin{aligned}
& -\log\left(\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\boldsymbol{\gamma}^0|\mathbf{D}_n)}\right) \\
= \;& \log\left(\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}^0)}\right) + \frac{1}{2}\log\left(\frac{\det(\mathbf{W}_{\boldsymbol{\gamma}})}{\det(\mathbf{W}_{\boldsymbol{\gamma}^0})}\right) \\
& + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}}\mathbf{U}_{\boldsymbol{\gamma}}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}}^T)\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})}\right) \\
& - \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}^0}\mathbf{U}_{\boldsymbol{\gamma}^0}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T)\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Y}}\right) \\
& + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Y}}\right).
\end{aligned}$$

Denote the five terms by J_1, J_2, J_3, J_4, J_5. It follows by Assumption A.2 that J_1 is bounded below uniformly for $\boldsymbol{\gamma} \in S(t_n)$. It is also easy to see that $J_3 \geq 0$ almost surely. To prove J_4 is lower bounded, by Sherman-Morrison-Woodbury (see [47]),

$$(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}+\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0}^{-1})^{-1} = (\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-1} - (\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0}+(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-1})^{-1}(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-1},$$

and by $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0} \geq \underline{\phi}_n\tau_m^2\mathbf{I}_{ms_n}$ and similar calculations in the proof of Theorem 2.2 in [42], it can be shown that

$$\frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}^0}\mathbf{U}_{\boldsymbol{\gamma}^0}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T)\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Y}} \leq 1 + \underline{\phi}_n^{-1}\tau_m^{-2}\frac{\mathbf{Y}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Y}}.$$

Note $\mathbf{Y} = \mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0 + \tilde{\boldsymbol{\eta}}$, where $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$, $\boldsymbol{\eta} = \sum_{j\in\boldsymbol{\gamma}^0}\sum_{l=m+1}^{\infty}\beta_{jl}^0\boldsymbol{\Phi}_{jl}$, $\boldsymbol{\Phi}_{jl} = (\varphi_{jl}(X_{j1}),\ldots,\varphi_{jl}(X_{jn}))^T$, and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^T$. Since for any m,

$$E\{\boldsymbol{\epsilon}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\epsilon}\} = ms_n\sigma_0^2, \text{ and}$$

$$\begin{aligned}
E\{\|\boldsymbol{\eta}\|^2\} &= nE\{(\sum_{j\in\boldsymbol{\gamma}^0}\sum_{l=m+1}^{\infty}\beta_{jl}^0\varphi_{jl}(X_{ji}))^2\} \\
&\leq ns_n\sum_{j\in\boldsymbol{\gamma}^0}E\{(\sum_{l=m+1}^{\infty}\beta_{jl}^0\varphi_{jl}(X_{ji}))^2\}
\end{aligned}$$

$$= \quad ns_n \sum_{j\in\boldsymbol{\gamma}^0} \sum_{l=m+1}^{\infty} |\beta_{jl}^0|^2 \le C_\beta n s_n^2 m^{-a},$$

where the last inequality follows by assumption (3.1), it can be shown by Bonferroni inequality that as $n \to \infty$,

$$P\left(\max_{m_1\le m\le m_2} m^{-1}h_m^{-1}\boldsymbol{\epsilon}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\epsilon} \le s_n\sigma_0^2\right) \to 1, \quad \text{and}$$
$$P\left(\max_{m_1\le m\le m_2} m^{a}h_m^{-1}\|\boldsymbol{\eta}\|^2 \le C_\beta n s_n^2\right) \to 1. \tag{8.6}$$

(8.6) will be frequently used in the proof of the main results in this paper. Since $\boldsymbol{\eta}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\eta} \le \|\boldsymbol{\eta}\|^2$, we have, w.p.a.1, for $m \in [m_1, m_2]$,

$$\begin{aligned}
& \mathbf{Y}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Y} \\
\le\ & 2\left(\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0\|^2 + \tilde{\boldsymbol{\eta}}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\tilde{\boldsymbol{\eta}}\right) \\
\le\ & 2\left(\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0\|^2 + c_0 n^{-1}\tilde{\boldsymbol{\eta}}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\tilde{\boldsymbol{\eta}}\right) \\
\le\ & 2\left(\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0\|^2 + 2c_0 n^{-1}\boldsymbol{\eta}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\eta} + 2c_0 n^{-1}\boldsymbol{\epsilon}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\epsilon}\right) \\
\le\ & 2\left(\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0\|^2 + 2c_0 C_\beta s_n^2 m^{-a}h_m + 2c_0\sigma_0^2 n^{-1} m h_m s_n\right) \\
\le\ & 2\left(\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0\|^2 + 2c_0 C_\beta s_n^2 m_1^{-a}h_{m_1} + 2c_0\sigma_0^2 n^{-1} m_2 h_{m_2} s_n\right).
\end{aligned}$$

Since $k_n \ge s_n\psi_n^2 \gg s_n^2 m_1^{-a}h_{m_1} + n^{-1}m_2 h_{m_2}s_n$, w.p.a.1, for $m \in [m_1, m_2]$, $\mathbf{Y}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Y} \le 2k_n(1+o(1))$. On the other hand, w.p.a.1, for $m \in [m_1, m_2]$,

$$\begin{aligned}
& \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Y} \\
=\ & \tilde{\boldsymbol{\eta}}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\boldsymbol{\eta} + 2\boldsymbol{\eta}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\
=\ & O\left(n s_n^2 m_1^{-a}h_{m_1} + n\sqrt{s_n^2 m_1^{-a}h_{m_1}} + m_2 h_{m_2}s_n\right) + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\
=\ & \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + O\left(n\sqrt{s_n^2 m_1^{-a}h_{m_1}} + m_2 h_{m_2}s_n\right).
\end{aligned} \tag{8.7}$$

By (1) in Assumption A.4, (8.7) implies $\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Y} = n\sigma_0^2(1+o_P(1))$. Therefore, w.p.a.1., for $m \in [m_1, m_2]$,

$$-J_4 \le \frac{n+\nu}{2}\log\left(1 + \frac{2k_n(1+o(1))}{n\underline{\phi}_n\tau_{m_2}^2\sigma_0^2}\right) = O(1),$$

where the last upper bound follows by $k_n = O(\underline{\phi}_n\tau_{m_2}^2)$, i.e., Assumption A.4 (3). This shows that, w.p.a.1, J_4 is lower bounded uniformly for $m \in [m_1, m_2]$ and c_j's $\in [\underline{\phi}_n, \bar{\phi}_n]$.

Next we approximate J_5 in two situations. First, for $\boldsymbol{\gamma} \in S_2(t_n)$, a direct calculation leads to

$$\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y} = \|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|^2 + 2\boldsymbol{\nu}_{\boldsymbol{\gamma},m}^T\tilde{\boldsymbol{\eta}} + \tilde{\boldsymbol{\eta}}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\tilde{\boldsymbol{\eta}},$$

where $\boldsymbol{\nu}_{\boldsymbol{\gamma},m} = (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}$. Since w.p.a.1., for $m \in [m_1, m_2]$, $\boldsymbol{\eta}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\boldsymbol{\eta} \le \|\boldsymbol{\eta}\|^2 \le C_\beta n s_n^2 m_1^{-a} h_{m_1}$, and $\boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\boldsymbol{\epsilon} \le \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \le 2n\sigma_0^2$, by Lemma 1 (iii), for a prefixed $\alpha > 4$

$$\tilde{\boldsymbol{\eta}}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\tilde{\boldsymbol{\eta}} \ge \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} - \alpha\sigma_0^2 t_n \log p - \sqrt{2C_\beta\sigma_0^2 n^2 s_n^2 m_1^{-a} h_{m_1}}.$$

Meanwhile, by Lemma 1 (i), for some large constant $C' > 0$ and w.p.a.1., uniformly for $m \in [m_1, m_2]$, $|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}^T\boldsymbol{\epsilon}| \le C'\sqrt{t_n \log p}\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|$ and $|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}^T\boldsymbol{\eta}| \le \sqrt{C_\beta n s_n^2 m_1^{-a} h_{m_1}}\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|$. By Assumption A.1, $\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|^2 \ge c_0^{-1} n\psi_n^2$, therefore we obtain that

$$\begin{aligned}
& \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y} \\
\ge\ & c_0^{-1} n\psi_n^2 \left(1 + O_P\left(\sqrt{\frac{t_n \log p}{n\psi_n^2}} + \sqrt{\frac{n s_n^2 m_1^{-a} h_{m_1}}{n\psi_n^2}}\right)\right. \\
& \left. + O_P\left(\frac{t_n \log p + n s_n^2 m_1^{-a} h_{m_1}}{n\psi_n^2}\right)\right) + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\
=\ & c_0^{-1} n\psi_n^2 (1 + o_P(1)) + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}.
\end{aligned}$$

Note Assumption A.4 (1) leads to $n\sqrt{s_n^2 m_1^{-a} h_{m_1}} + m_2 h_{m_2} s_n = o(n\psi_n^2)$ and $n\sqrt{s_n^2 m_1^{-a} h_{m_1}} + m_2 h_{m_2} s_n = o(n)$. By (8.7), we have, w.p.a.1., uniformly for $m \in [m_1, m_2]$,

$$J_5 \ge \frac{n+\nu}{2} \log\left(1 + \frac{c_0^{-1}\psi_n^2(1 + o(1))}{\sigma_0^2}\right) \ge \frac{n+\nu}{2} \log\left(1 + C'\psi_n^2\right),$$

for some large constant $C' > 0$.

Next we consider $\boldsymbol{\gamma} \in S_1(t_n)$. It can be checked by (8.7), Lemma 1 and straightforward calculations that for a fixed $\alpha > 4$, w.p.a.1., uniformly for $m \in [m_1, m_2]$,

$$\begin{aligned}
J_5 &= \frac{n+\nu}{2} \log\left(1 - \frac{\tilde{\boldsymbol{\eta}}^T(\mathbf{P}_{\boldsymbol{\gamma}} - \mathbf{P}_{\boldsymbol{\gamma}^0})\tilde{\boldsymbol{\eta}}}{1 + \tilde{\boldsymbol{\eta}}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\tilde{\boldsymbol{\eta}}}\right) \\
&\ge \frac{n+\nu}{2} \log\left(1 - \frac{2\|\boldsymbol{\eta}\|^2 + 2\boldsymbol{\epsilon}^T(\mathbf{P}_{\boldsymbol{\gamma}} - \mathbf{P}_{\boldsymbol{\gamma}^0})\boldsymbol{\epsilon}}{1 + \tilde{\boldsymbol{\eta}}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\tilde{\boldsymbol{\eta}}}\right) \\
&\ge \frac{n+\nu}{2} \log\left(1 - \frac{2C_\beta n s_n^2 m_1^{-a} h_{m_1} + 2(|\boldsymbol{\gamma}| - s_n)\alpha\sigma_0^2 \log p}{1 + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + O\left(n\sqrt{s_n^2 m_1^{-a} h_{m_1}} + m_2 h_{m_2} s_n\right)}\right)
\end{aligned}$$

$$\begin{aligned} &\geq \frac{n+\nu}{2}\log\left(1-\frac{2C_\beta n s_n^2 m_1^{-a}h_{m_1}+2(|\boldsymbol{\gamma}|-s_n)\alpha\sigma_0^2\log p}{n\sigma_0^2(1+o(1))}\right)\\ &\geq -(3C_\beta\sigma_0^{-2}ns_n^2m_1^{-a}h_{m_1}+2(|\boldsymbol{\gamma}|-s_n)\alpha_0\log p),\end{aligned}$$

where the last inequality follows by $t_n\log p = o(n)$, i.e., Assumption A.4 (2), the inequality that $\log(1-x)\geq -2x$ when $x\in(0,1/2)$, and a suitably fixed $\alpha_0\in(4,\alpha)$.

In the end we analyze the term J_2. Using the proof of Lemma A.2 in [42], it can be shown that for any c_j's $\in[\underline{\phi}_n,\bar{\phi}_n]$ and $m\in[m_1,m_2]$,

$$\begin{aligned} &J_2\geq\frac{1}{2}m_1(|\boldsymbol{\gamma}|-s_n)\log\left(1+c_0^{-1}n\underline{\phi}_n\tau_{m_2}^2\right)\ \text{ for any } \boldsymbol{\gamma}\in S_1(t_n),\text{ and}\\ &J_2\geq-\frac{m_2s_n}{2}\log\left(1+c_0n\bar{\phi}_n\tau_1^2\right)\ \text{ for any } \boldsymbol{\gamma}\in S_2(t_n).\end{aligned}\tag{8.8}$$

To make the proofs more readable, we give the brief proof of (8.8). When $\boldsymbol{\gamma}\in S_1(t_n)$, by Sylvester's determinant formula (see [47]), Assumption A.1 and straightforward calculations we have

$$\begin{aligned}\det(\mathbf{U}_{\boldsymbol{\gamma}}) &= \det(\mathbf{U}_{\boldsymbol{\gamma}^0})\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}^{-1}+\mathbf{Z}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}^T(\mathbf{I}_n-\mathbf{Z}_{\boldsymbol{\gamma}^0}\mathbf{U}_{\boldsymbol{\gamma}^0}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T)\mathbf{Z}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}\right)\\ &\geq \det(\mathbf{U}_{\boldsymbol{\gamma}^0})\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}^{-1}+\mathbf{Z}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}^T(\mathbf{I}_n-\mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Z}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}\right)\\ &\geq \det(\mathbf{U}_{\boldsymbol{\gamma}^0})\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0}^{-1}+c_0^{-1}n\mathbf{I}_{m|\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}^0|}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\det(\mathbf{W}_{\boldsymbol{\gamma}})}{\det(\mathbf{W}_{\boldsymbol{\gamma}^0})} &= \frac{\det(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}})}{\det(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0})}\cdot\frac{\det(\mathbf{U}_{\boldsymbol{\gamma}})}{\det(\mathbf{U}_{\boldsymbol{\gamma}^0})}\\ &\geq \left(1+c_0^{-1}n\underline{\phi}_n\tau_m^2\right)^{m(|\boldsymbol{\gamma}|-s_n)}\geq\left(1+c_0^{-1}n\underline{\phi}_n\tau_{m_2}^2\right)^{m_1(|\boldsymbol{\gamma}|-s_n)}.\end{aligned}$$

Taking logarithm on both sides, we obtain the first inequality in (8.8). When $\boldsymbol{\gamma}\in S_2(t_n)$, since $\det(\mathbf{W}_{\boldsymbol{\gamma}})\geq 1$, the second inequality in (8.8) follows by

$$\begin{aligned}J_2 &\geq -\frac{1}{2}\log(\det(\mathbf{W}_{\boldsymbol{\gamma}^0}))=-\frac{1}{2}\log\left(\det\left(\mathbf{I}_{ms_n}+\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0}^{1/2}\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^0}^{1/2}\right)\right)\\ &\geq -\frac{ms_n}{2}\log\left(1+c_0n\bar{\phi}_n\tau_1^2\right)\geq-\frac{m_2s_n}{2}\log\left(1+c_0n\bar{\phi}_n\tau_1^2\right).\end{aligned}$$

To the end of the proof, we notice that based on the above approximations of J_1 to J_5, there exist some large positive constants $\tilde{C}$ and N such that when $n\geq N$, w.p.a.1., for any c_j's $\in[\underline{\phi}_n,\bar{\phi}_n]$ and $m\in[m_1,m_2]$,

$$\begin{aligned}&\sum_{\boldsymbol{\gamma}\in S_1(t_n)}\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\boldsymbol{\gamma}^0|\mathbf{D}_n)}\\ \leq\ &\tilde{C}\sum_{\boldsymbol{\gamma}\in S_1(t_n)}\exp\left(3C_\beta\sigma_0^{-2}ns_n^2m_1^{-a}h_{m_1}+2\alpha_0(|\boldsymbol{\gamma}|-s_n)\log p\right.\end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{m_1(|\boldsymbol{\gamma}| - s_n)}{2} \log(1 + c_0^{-1} n \underline{\phi}_n \tau_{m_2}^2) \Bigg) \\
&= \tilde{C} \sum_{v=s_n+1}^{t_n} \binom{p - s_n}{v - s_n} \left(\frac{p^{2\alpha_0} \exp(3C_\beta \sigma_0^{-2} n s_n^2 m_1^{-a} h_{m_1})}{(1 + c_0^{-1} n \underline{\phi}_n \tau_{m_2}^2)^{m_1/2}} \right)^{v - s_n} \\
&= \tilde{C} \sum_{v=1}^{t_n - s_n} \binom{p - s_n}{v} \left(\frac{p^{2\alpha_0} \exp(3C_\beta \sigma_0^{-2} n s_n^2 m_1^{-a} h_{m_1})}{(1 + c_0^{-1} n \underline{\phi}_n \tau_{m_2}^2)^{m_1/2}} \right)^{v} \\
&\le \tilde{C} \sum_{v=1}^{t_n - s_n} \frac{p^v}{v!} \left(\frac{p^{2\alpha_0} \exp(3C_\beta \sigma_0^{-2} n s_n^2 m_1^{-a} h_{m_1})}{(1 + c_0^{-1} n \underline{\phi}_n \tau_{m_2}^2)^{m_1/2}} \right)^{v} \\
&\le \tilde{C} \left(\exp\left(\frac{p^{2\alpha_0 + 1} \exp(3C_\beta \sigma_0^{-2} n s_n^2 m_1^{-a} h_{m_1})}{(1 + c_0^{-1} n \underline{\phi}_n \tau_{m_2}^2)^{m_1/2}} \right) - 1 \right) \to 0, \text{ as } n \to \infty,
\end{aligned}
$$

where the last limit follows by Assumption A.4 (1) & (3), and by Assumption A.4 (4) we can make N large enough so that $m_2 s_n \log(1 + c_0 n \bar{\phi}_n \tau_1^2) \le \frac{n+\nu}{2} \log(1 + C' \psi_n^2)$ for $n \ge N$, which leads to

$$
\begin{aligned}
&\sum_{\boldsymbol{\gamma} \in S_2(t_n)} \frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\boldsymbol{\gamma}^0|\mathbf{D}_n)} \\
&\le \tilde{C} \sum_{\boldsymbol{\gamma} \in S_2(t_n)} \exp\left(\frac{1}{2} m_2 s_n \log(1 + c_0 n \bar{\phi}_n \tau_1^2) - \frac{n+\nu}{2} \log(1 + C' \psi_n^2) \right) \\
&\le \tilde{C} \sum_{\boldsymbol{\gamma} \in S_2(t_n)} \exp\left(-\frac{n+\nu}{4} \log(1 + C' \psi_n^2) \right) \\
&\le \tilde{C} \cdot \# S_2(t_n) \cdot (1 + C' \psi_n^2)^{-(n+\nu)/4} \\
&\le \tilde{C} \cdot p^{t_n} \cdot (1 + C' \psi_n^2)^{-(n+\nu)/4} \to 0, \text{ as } n \to \infty,
\end{aligned}
$$

where the last limit follows by Assumption A.4 (2). This completes the proof of Theorem 3.2.

Before proving Theorem 3.4, we need the following lemma. The proof is similar to that of Lemma 2 in [44] and thus is omitted.

Lemma 2. *Suppose* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_0^2 \boldsymbol{I}_n)$. *Adopt the convention that* $\boldsymbol{\nu}_{\boldsymbol{\gamma}}^T \boldsymbol{\epsilon} / \|\boldsymbol{\nu}_{\boldsymbol{\gamma}}\| = 0$ *when* $\boldsymbol{\nu}_{\boldsymbol{\gamma}} = 0$, *and* $\boldsymbol{\epsilon}^T \boldsymbol{P}_{\boldsymbol{\gamma}} \boldsymbol{\epsilon} / |\boldsymbol{\gamma}| = 0$ *when* $\boldsymbol{\gamma}$ *is null. Furthermore,* $m_2 \le n = o(p)$.

(i). For $\boldsymbol{\gamma} \in T_0(t_n)$, *define* $\boldsymbol{\nu}_{\boldsymbol{\gamma}} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{\gamma}}) \boldsymbol{Z}_{\boldsymbol{\gamma}^0 \setminus \boldsymbol{\gamma}} \boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0 \setminus \boldsymbol{\gamma}}$. *Then*

$$
\max_{1 \le m \le m_2} \max_{\boldsymbol{\gamma} \in T_0(t_n)} \frac{|\boldsymbol{\nu}_{\boldsymbol{\gamma}}^T \boldsymbol{\epsilon}|}{\|\boldsymbol{\nu}_{\boldsymbol{\gamma}}\|} = O_P(\sqrt{s_n + \log m_2}).
$$

(ii). For $\boldsymbol{\gamma} \in T_1(t_n)$, *denote* $\boldsymbol{\gamma}^* = \boldsymbol{\gamma} \cap \boldsymbol{\gamma}^0$ *which is non-null. For any fixed* $\alpha > 6$,

$$
\lim_{n \to \infty} P\left(\max_{1 \le m \le m_2} \max_{\boldsymbol{\gamma} \in T_1(t_n)} \frac{\boldsymbol{\epsilon}^T (\boldsymbol{P}_{\boldsymbol{\gamma}} - \boldsymbol{P}_{\boldsymbol{\gamma}^*}) \boldsymbol{\epsilon}}{|\boldsymbol{\gamma}| - |\boldsymbol{\gamma}^*|} \le \alpha \sigma_0^2 s_n \log p \right) = 1.
$$

(iii). Then for any fixed $\alpha > 4$,

$$\lim_{n\to\infty} P\left(\max_{1\le m\le m_2}\max_{\boldsymbol{\gamma}\in T_2(t_n)} \boldsymbol{\epsilon}^T \boldsymbol{P}_{\boldsymbol{\gamma}}\boldsymbol{\epsilon}/|\boldsymbol{\gamma}| \le \alpha\sigma_0^2 \log p\right) = 1.$$

Proof of Proposition 3.3

Let $C_\varphi = \max_{1\le j\le p}\sup_{l\ge 1}\|\varphi_{jl}\|_{\sup}$. By Proposition 3.1, we obtain that (3.5) holds. Next we show that (3.6) holds with $\rho_n \propto m_2 s_n^2 \log p$. Define $\boldsymbol{\Delta} = \mathbf{Z}^T_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}$. The diagonal entry of $\boldsymbol{\Delta}$ is $\boldsymbol{\Delta}_{j,l} = \boldsymbol{\Phi}^T_{jl}\mathbf{P}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{jl}$ for $j \in \boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}$, and $l = 1,\dots,m$. By [5], any random variable ξ almost surely bounded by a number $b > 0$ satisfies $E\{\exp(a\xi)\} \le \exp(a^2b^2/2)$, i.e., ξ is sub-Gaussian. Since $\varphi_{jl}(X_{ji})$, $i = 1,\dots,n$, are independent and uniformly bounded by C_φ, for any n-vector $\mathbf{a} = (a_1,\dots,a_n)^T$, $E\{\exp(\mathbf{a}^T\boldsymbol{\Phi}_{jl})\} = \prod_{i=1}^n E\{\exp(a_i\varphi_{jl}(X_{ji}))\} \le \prod_{i=1}^n \exp(a_i^2C_\varphi^2/2) = \exp(\|\mathbf{a}\|^2C_\varphi^2/2)$, that is, $\boldsymbol{\Phi}_{jl}$ is sub-Gaussian. By Theorem 2.1 of [27], for some $C > 2$ which implies $5CC_\varphi^2|\boldsymbol{\gamma}|\log p > C_\varphi^2(|\boldsymbol{\gamma}| + 2\sqrt{|\boldsymbol{\gamma}|t} + 2t)$ with $t = C|\boldsymbol{\gamma}|\log p$, we have

$$\begin{aligned}
& P\left(\max_{m\in[m_1,m_2]}\max_{0<|\boldsymbol{\gamma}|<s_n}\max_{\substack{j\in\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}\\ l=1,\dots,m}} \boldsymbol{\Phi}^T_{jl}\mathbf{P}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{jl}/|\boldsymbol{\gamma}| \ge CC_\varphi^2\log p\right)\\
\le\ & \sum_{1\le m\le m_2}\sum_{0<|\boldsymbol{\gamma}|<s_n}\sum_{\substack{j\in\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}\\ l=1,\dots,m}} P\left(\boldsymbol{\Phi}^T_{jl}\mathbf{P}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{jl} \ge CC_\varphi^2|\boldsymbol{\gamma}|\log p\right)\\
\le\ & \sum_{1\le m\le m_2}\sum_{0<|\boldsymbol{\gamma}|<s_n}\sum_{\substack{j\in\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}\\ l=1,\dots,m}} E\left\{P\left(\boldsymbol{\Phi}^T_{jl}\mathbf{P}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{jl} \ge CC_\varphi^2|\boldsymbol{\gamma}|\log p\middle|\mathbf{P}_{\boldsymbol{\gamma}}\right)\right\}\\
\le\ & \sum_{1\le m\le m_2}\sum_{0<|\boldsymbol{\gamma}|<s_n}\sum_{\substack{j\in\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}\\ l=1,\dots,m}} \exp(-C|\boldsymbol{\gamma}|\log p)\\
\le\ & m_2^2 s_n\sum_{r=1}^{s_n-1}\binom{p}{r}p^{-Cr} \le m_2^2 s_n\sum_{r=1}^{s_n-1}\frac{p^r}{r!}p^{-Cr} \le m_2^2 s_n(\exp(p^{1-C})-1)\\
=\ & O(m_2^2 s_n/p) = o(1),
\end{aligned}$$

therefore,

$$\max_{m\in[m_1,m_2]}\max_{0<|\boldsymbol{\gamma}|<s_n}\max_{\substack{j\in\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}\\ l=1,\dots,m}} \boldsymbol{\Phi}^T_{jl}\mathbf{P}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{jl}/|\boldsymbol{\gamma}| = O_P(\log p).$$

So with probability approaching one, for any $m \in [m_1, m_2]$ and $\boldsymbol{\gamma} \in T(s_n - 1)\setminus\{\emptyset\}$, $\lambda_+(\mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}) \le \operatorname{trace}(\mathbf{Z}^T_{\boldsymbol{\gamma}}\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}}) \le C' m_2 s_n^2 \log p$, for some large constant $C' > 0$. This completes the proof.

Proof of Theorem 3.4 (i)

Like in Assumption A.4, one can replace θ_n and l_n in Assumption B.3 by ψ_n and k_n while preserving an equivalent condition. Specifically, by the statements in the beginning of Theorem 3.2, it can be shown that the following assumption is an equivalent version of Assumption B.3.

Assumption B.4. *There exists a positive sequence* $\{h_m, m \geq 1\}$ *such that, as* $m, m_1, m_2 \to \infty$, $h_m \to \infty$, $m^{-a}h_m$ *decreasingly converges to zero,* mh_m *increasingly converges to* ∞, *and* $\sum_{m_1 \leq m \leq m_2} 1/h_m = o(1)$. *Furthermore, the sequences* $m_1, m_2, h_m, s_n, \psi_n, k_n, \underline{\phi}_n$ *satisfy*

(1). $m_2 h_{m_2} s_n = o(n \min\{1, \psi_n^2\})$ *and* $m_1^{-a} h_{m_1} s_n^2 = o(\min\{1, n^{-1} m_1 \log(\underline{\phi}_n), \psi_n^2\})$;
(2). $k_n = O(\underline{\phi}_n \tau_{m_2}^2)$;
(3). $\max\{\rho_n, s_n^2 \log p\} = o(\min\{n, m_1 \log(n\underline{\phi}_n \tau_{m_2}^2)\})$.

Next we will prove the theorem based on Assumptions B.1, B.2 and B.4. We first show that w.p.a.1, for $m \in [m_1, m_2]$, $\max_{\boldsymbol{\gamma} \in T_1(t_n)} p(\boldsymbol{\gamma}|\mathbf{D}_n)/p(\boldsymbol{\gamma} \cap \boldsymbol{\gamma}^0|\mathbf{D}_n)$ converges to zero. Since the denominator is bounded by $\max_{\boldsymbol{\gamma} \in T_0(t_n)} p(\boldsymbol{\gamma}|\mathbf{D}_n)$, it follows that $\max_{\boldsymbol{\gamma} \in T_1(t_n)} p(\boldsymbol{\gamma}|\mathbf{D}_n)/\max_{\boldsymbol{\gamma} \in T_0(t_n)} p(\boldsymbol{\gamma}|\mathbf{D}_n) \to 0$ in probability. Second, we show, w.p.a.1, for $m \in [m_1, m_2]$, $\max_{\boldsymbol{\gamma} \in T_2(t_n)} p(\boldsymbol{\gamma}|\mathbf{D}_n)/p(\emptyset|\mathbf{D}_n) \to 0$. This will complete the proof. Next we proceed in two steps.

Step 1: Consider the following decomposition for $\boldsymbol{\gamma} \in T_1(t_n)$,

$$\begin{aligned}
& -\log\left(\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\boldsymbol{\gamma}^*|\mathbf{D}_n)}\right) \\
= \; & -\log\left(\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}^*)}\right) + \frac{1}{2}\log\left(\frac{\det(\mathbf{W}_{\boldsymbol{\gamma}})}{\det(\mathbf{W}_{\boldsymbol{\gamma}^*})}\right) \\
& + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}}\mathbf{U}_{\boldsymbol{\gamma}}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}}^T)\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}}\right) \\
& - \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}^*}\mathbf{U}_{\boldsymbol{\gamma}^*}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T)\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}\right) \\
& + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}\right),
\end{aligned}$$

where $\boldsymbol{\gamma}^* = \boldsymbol{\gamma} \cap \boldsymbol{\gamma}^0 \neq \emptyset$. Denote the five items by J_1, J_2, J_3, J_4, J_5. We use the methods in the proof of Theorem 3.2 to analyze the five terms. Note that J_1 is bounded below by Assumption B.2, and $J_3 \geq 0$ almost surely. To handle J_4, using Sherman-Morrison-Woodbury matrix identity,

$$\begin{aligned}
& \frac{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}^*}\mathbf{U}_{\boldsymbol{\gamma}^*}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T)\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}} \\
= \; & 1 + \frac{\mathbf{Y}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}((\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-1} - \mathbf{U}_{\boldsymbol{\gamma}^*}^{-1})\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}
\end{aligned}$$

$$= 1 + \frac{\mathbf{Y}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*} + (\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-1})^{-1}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Y}}{1 + \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}$$

$$\le 1 + \underline{\phi}_n^{-1}\tau_m^{-2}\frac{\mathbf{Y}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Y}}{1 + \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}$$

$$\le 1 + 2\underline{\phi}_n^{-1}\tau_m^{-2}\frac{(\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0)^T\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0 + \tilde{\boldsymbol{\eta}}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\tilde{\boldsymbol{\eta}}}{1 + \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}.$$

Without loss of generality, assume $\mathbf{Z}_{\boldsymbol{\gamma}^0} = (\mathbf{Z}_{\boldsymbol{\gamma}^*}, \mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*})$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0 = ((\boldsymbol{\beta}_{\boldsymbol{\gamma}^*}^0)^T, (\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^0)^T)^T$. By a direct calculation it can be examined that

$$\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^0} = \begin{pmatrix} \mathbf{I}_{|\boldsymbol{\gamma}^*|} & (\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-1}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*} \\ \mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-1} & \mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*} \end{pmatrix}.$$

By Assumption B.1, w.p.a.1, for $\boldsymbol{\gamma} \in T_2(t_n)$ and $m \in [m_1, m_2]$,

$$\lambda_+\left(\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\right) \le \frac{d_0}{n}\lambda_+\left(\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\mathbf{P}_{\boldsymbol{\gamma}^*}\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\right) \le \frac{d_0\rho_n}{n},$$

which implies, w.l.p., $\lambda_+(\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}) \le 1 + \frac{d_0\rho_n}{n}$. Therefore, it can be shown that

$$(\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0)^T\mathbf{Z}_{\boldsymbol{\gamma}^0}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0 \le (1 + \frac{d_0\rho_n}{n})k_n.$$

On the other hand, by (8.6) in the proof of Theorem 3.2, it can be shown that, w.p.a.1, for $m \in [m_1, m_2]$, $\tilde{\boldsymbol{\eta}}^T\mathbf{Z}_{\boldsymbol{\gamma}^*}(\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\mathbf{Z}_{\boldsymbol{\gamma}^*})^{-2}\mathbf{Z}_{\boldsymbol{\gamma}^*}^T\tilde{\boldsymbol{\eta}} \le \frac{2d_0}{n}(\|\boldsymbol{\eta}\|^2 + \boldsymbol{\epsilon}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\epsilon}) \le \frac{2d_0}{n}(\sigma_0^2 s_n m_2 h_{m_2} + C_\beta m_1^{-a} h_{m_1} n s_n^2)$. Meanwhile, by (8.7), $\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y} \ge \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^0})\mathbf{Y} = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + O(n\sqrt{s_n^2 m_1^{-a} h_{m_1}} + m_2 h_{m_2} s_n)$. So for $m \in [m_1, m_2]$, and c_j's $\in [\underline{\phi}_n, \bar{\phi}_n]$, we have $0 \le -J_4 \le \frac{n+\nu}{2}\log(1 + \frac{2(1+d_0\rho_n/n)k_n(1+o_P(1))}{n\underline{\phi}_n\tau_{m_2}^2\sigma_0^2}) = O_P(1)$ since $k_n = O(\underline{\phi}_n\tau_{m_2}^2)$ (see Assumption B.4).

To approximate J_5, without loss of generality, we may assume $\mathbf{Z}_{\boldsymbol{\gamma}^0} = (\mathbf{Z}_{\boldsymbol{\gamma}^*}, \mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*})$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}^0}^0 = ((\boldsymbol{\beta}_{\boldsymbol{\gamma}^*}^0)^T, (\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^0)^T)^T$. It can be shown by Assumption B.1, B.4 (1), (8.6), and Lemma 2 (ii) that

$$\begin{aligned}
&\mathbf{Y}^T(\mathbf{P}_{\boldsymbol{\gamma}} - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y} \\
\le\ & 2(\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^0)^T\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^T(\mathbf{P}_{\boldsymbol{\gamma}} - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^0 + 4\boldsymbol{\eta}^T(\mathbf{P}_{\boldsymbol{\gamma}} - \mathbf{P}_{\boldsymbol{\gamma}^*})\boldsymbol{\eta} \\
& + 4\boldsymbol{\epsilon}^T(\mathbf{P}_{\boldsymbol{\gamma}} - \mathbf{P}_{\boldsymbol{\gamma}^*})\boldsymbol{\epsilon} \\
\le\ & 2\rho_n\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^0\|^2 + 4(C_\beta n s_n^2 m_1^{-a} h_{m_1} + \alpha\sigma_0^2 s_n^2 \log p) \\
\le\ & 2g_n(\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^0\|^2 + \alpha_1),
\end{aligned}$$

where $g_n = \max\{\rho_n, n s_n^2 m_1^{-a} h_{m_1}, s_n^2 \log p\}$, $\alpha > 4$ and α_1 are fixed positive constants. On the other hand, define $\boldsymbol{\nu}_{\boldsymbol{\gamma}^*,m} = (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\boldsymbol{\beta}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}^0$. Then

by Assumption B.1, $\|\boldsymbol{\nu}_{\boldsymbol{\gamma}^*,m}\|^2 \geq (d_0^{-1}n-\rho_n)\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\|^2$. By Lemma 2 (i), w.p.a.1, for any $m \in [m_1, m_2]$ and $\boldsymbol{\gamma} \in T_1(t_n)$,

$$\begin{aligned}
& \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y} \\
= \;& \|\boldsymbol{\nu}_{\boldsymbol{\gamma}^*,m}\|^2 + 2\boldsymbol{\nu}^T_{\boldsymbol{\gamma}^*,m}\tilde{\boldsymbol{\eta}} + \tilde{\boldsymbol{\eta}}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\tilde{\boldsymbol{\eta}} \\
\geq \;& \|\boldsymbol{\nu}_{\boldsymbol{\gamma}^*,m}\|^2 \left(1 + O\left(\frac{\sqrt{s_n + \log m_2} + \sqrt{ns_n^2 m_1^{-a} h_{m_1}}}{\sqrt{n\psi_n^2}}\right)\right) \\
& + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{P}_{\boldsymbol{\gamma}^0}\boldsymbol{\epsilon} - 2\|\boldsymbol{\eta}\| \cdot \|\boldsymbol{\epsilon}\| \\
= \;& \|\boldsymbol{\nu}_{\boldsymbol{\gamma}^*,m}\|^2 \left(1 + O\left(\frac{\sqrt{s_n + \log m_2} + \sqrt{ns_n^2 m_1^{-a} h_{m_1}}}{\sqrt{n\psi_n^2}}\right)\right) \\
& + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} - \sigma_0^2 s_n m_2 h_{m_2} - 2C'\sqrt{C_\beta n^2 s_n^2 m_1^{-a} h_{m_1}} \\
= \;& ((d_0^{-1}n - \rho_n)\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\|^2 + n\sigma_0^2)(1 + o(1)),
\end{aligned}$$

for some constant $C' > 0$. Therefore, for some large positive constant C'', w.p.a.1, for any $m \in [m_1, m_2]$ and $\boldsymbol{\gamma} \in T_1(t_n)$,

$$\begin{aligned}
J_5 &= \frac{n+\nu}{2}\log\left(1 - \frac{\mathbf{Y}^T(\mathbf{P}_{\boldsymbol{\gamma}} - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}^*})\mathbf{Y}}\right) \\
&\geq \frac{n+\nu}{2}\log\left(1 - \frac{2g_n(\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\|^2 + \alpha_1)}{((d_0^{-1}n - \rho_n)\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}^*}\|^2 + n\sigma_0^2)(1 + o(1))}\right) \geq -C''g_n.
\end{aligned}$$

By similar arguments in the proof of Theorem 3.2, it can be shown that for any $m \in [m_1, m_2]$, c_j's $\in [\underline{\phi}_n, \bar{\phi}_n]$, and $\boldsymbol{\gamma} \in T_1(t_n)$, $J_2 \geq \frac{m_1}{2}\log(1 + (d_0^{-1}n - \rho_n)\underline{\phi}_n\tau^2_{m_2})$. So, w.p.a.1, for any $m \in [m_1, m_2]$ and $\boldsymbol{\gamma} \in T_1(t_n)$, for some constant $\widetilde{C} > 0$

$$\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\boldsymbol{\gamma}^*|\mathbf{D}_n)} \leq \widetilde{C}\exp\left(-\frac{m_1}{2}\log\left(1 + (nd_0^{-1} - \rho_n)\underline{\phi}_n\tau^2_{m_2}\right) + C''g_n\right) \to 0.$$

Thus, $\max_{m\in[m_1,m_2]}\max_{c_j\in[\underline{\phi}_n,\bar{\phi}_n]}\frac{\max_{\boldsymbol{\gamma}\in T_1(t_n)}p(\boldsymbol{\gamma}|\mathbf{D}_n)}{\max_{\boldsymbol{\gamma}\in T_0(t_n)}p(\boldsymbol{\gamma}|\mathbf{D}_n)} = o_P(1)$.

Step 2: Next we consider the following decomposition for $\boldsymbol{\gamma} \in T_2(t_n)$,

$$\begin{aligned}
& -\log\left(\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\emptyset|\mathbf{D}_n)}\right) \\
= \;& -\log\left(\frac{p(\boldsymbol{\gamma})}{p(\emptyset)}\right) + \frac{1}{2}\log\left(\frac{\det(\mathbf{W}_{\boldsymbol{\gamma}})}{\det(\mathbf{W}_\emptyset)}\right) \\
& + \frac{n+\nu}{2}\log\left(\frac{1 + \mathbf{Y}^T(\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\gamma}}\mathbf{U}_{\boldsymbol{\gamma}}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}}^T)\mathbf{Y}}{1 + \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}}\right) \\
& + \frac{n+\nu}{2}\log\left(\frac{1 + \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}}{1 + \mathbf{Y}^T\mathbf{Y}}\right).
\end{aligned}$$

Denote the above four terms by J_1, J_2, J_3, J_4. It is clear that J_1 is lower bounded, and $J_3 \geq 0$. We approximate J_4. For $\boldsymbol{\gamma} \in T_2(t_n)$, let $\boldsymbol{\nu}_{\boldsymbol{\gamma},m} = \mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}$. By Assumption B.1, $\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|^2 \leq \rho_n k_n$. Thus, by Lemma 2, for some fixed $\alpha > 4$, for any $m \in [m_1, m_2]$,

$$\begin{aligned}\mathbf{Y}^T\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Y} &\leq 2\left(\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|^2 + 2\boldsymbol{\eta}^T\mathbf{P}_{\boldsymbol{\gamma}}\boldsymbol{\eta} + 2\boldsymbol{\epsilon}^T\mathbf{P}_{\boldsymbol{\gamma}}\boldsymbol{\epsilon}\right)\\ &\leq 2(\rho_n k_n + 2C_\beta n s_n^2 m_1^{-a} h_{m_1} + 2\alpha\sigma_0^2 s_n \log p) \leq 2g_n(k_n + \alpha_2),\end{aligned}$$

where $g_n = \max\{\rho_n, n s_n^2 m_1^{-a} h_{m_1}, s_n \log p\}$, and α_2 is some fixed positive constant. On the other hand, since $E\{|(\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0})^T\boldsymbol{\epsilon}|^2/\|\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}\|^2\} = \sigma_0^2$ we have $|(\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0})^T\boldsymbol{\epsilon}|/\|\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}\| = O_P(1)$. Thus, $|(\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0})^T\tilde{\boldsymbol{\eta}}| \leq \|\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}\| \cdot (\|\boldsymbol{\eta}\| + O_P(1))$. Since $\sqrt{s_n^2 m_1^{-a} h_{m_1}} = o(\psi_n^2) = o(k_n)$, we have

$$\begin{aligned}\mathbf{Y}^T\mathbf{Y} &= \|\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}\|^2 + 2(\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0})^T\tilde{\boldsymbol{\eta}} + \tilde{\boldsymbol{\eta}}^T\tilde{\boldsymbol{\eta}}\\ &= \|\mathbf{X}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}\|^2\left(1 + O_P\left(\sqrt{\frac{1 + n s_n^2 m_1^{-a} h_{m_1}}{n k_n}}\right)\right) + n\sigma_0^2(1 + o_P(1))\\ &\quad + O_P\left(\sqrt{n^2 s_n^2 m_1^{-a} h_{m_1}}\right)\\ &= \|\mathbf{X}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}\|^2(1 + o_P(1)) + n\sigma_0^2(1 + o_P(1))\\ &\geq (d_0^{-1} n k_n + n\sigma_0^2) \cdot (1 + o_P(1)).\end{aligned}$$

Therefore, w.p.a.1, for $\boldsymbol{\gamma} \in T_2(t_n)$ and $m \in [m_1, m_2]$,

$$J_4 \geq \frac{n+\nu}{2}\log\left(1 - \frac{2g_n(k_n + \alpha_2)}{(d_0^{-1} n k_n + n\sigma_0^2)}\right) \geq -C' g_n,$$

for some large constant $C' > 0$.

Meanwhile, by similar proof in Step 1, it can be verified that for $\boldsymbol{\gamma} \in T_2(t_n)$ and $m \in [m_1, m_2]$, $J_2 \geq \frac{m_1}{2}\log(1 + n d_0^{-1}\underline{\phi}_n \tau_{m_2}^2)$ which holds for c_j's $\in [\underline{\phi}_n, \bar{\phi}_n]$. Then w.p.a.1, for $\boldsymbol{\gamma} \in T_2(t_n)$, c_j's $\in [\underline{\phi}_n, \bar{\phi}_n]$ and $m \in [m_1, m_2]$,

$$\frac{p(\boldsymbol{\gamma}|\mathbf{D}_n)}{p(\emptyset|\mathbf{D}_n)} \leq \tilde{C}\exp\left(-\frac{m_1}{2}\log(1 + n d_0^{-1}\underline{\phi}_n\tau_{m_2}^2) + C' g_n\right) = o_P(1),$$

where $\tilde{C}$ is some large positive constant. This shows

$$\max_{m\in[m_1,m_2]}\max_{c_j\in[\underline{\phi}_n,\bar{\phi}_n]}\frac{\max_{\boldsymbol{\gamma}\in T_2(t_n)} p(\boldsymbol{\gamma}|\mathbf{D}_n)}{\max_{\boldsymbol{\gamma}\in T_0(t_n)} p(\boldsymbol{\gamma}|\mathbf{D}_n)} = o_P(1).$$

This shows the desired result.

Proof of Theorem 3.4 (ii)

Under Assumption B.4, it can be shown using similar arguments in the beginning of the proof of Theorem 3.2 that Assumption A.3 (4) is equivalent to the

following assumption, i.e., Assumption A.4 (4),

$$m_2 s_n \log(1+n\bar{\phi}_n) = o(n\log(1+\min\{1,\psi_n^2\})). \tag{8.9}$$

Similarly, (3.7) can be shown to be equivalent to

$$\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\|^2 \le b_0'\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2, \tag{8.10}$$

where $b_0' > 0$ is constant. To see this, using (3.1) and $\psi_n^2 \gg m_1^{-a}$ (see Assumption B.4 (1)), it can be shown that $\sum_{j\in\boldsymbol{\gamma}}\|f_j^0\|_j^2 = \|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2(1+o(1))$ and $\sum_{j\in\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\|f_j^0\|_j^2 = \|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\|^2(1+o(1))$, uniformly for $m\in[m_1,m_2]$. Then it can be seen that (8.10) is equivalent to (3.7). Next we will prove the theorem based on Assumptions B.1, B.2, B.4, (8.9) and (8.10).

For the $\boldsymbol{\gamma}$ specified in the theorem, we consider the following decomposition

$$\begin{aligned}
&-\log\left(\frac{p(\emptyset|\mathbf{Z})}{p(\boldsymbol{\gamma}|\mathbf{Z})}\right)\\
=\;&-\log\left(\frac{p(\emptyset)}{p(\boldsymbol{\gamma})}\right)+\frac{1}{2}\log\left(\frac{1}{\det(\mathbf{W}_{\boldsymbol{\gamma}})}\right)\\
&-\frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T(\mathbf{I}_n-\mathbf{Z}_{\boldsymbol{\gamma}}\mathbf{U}_{\boldsymbol{\gamma}}^{-1}\mathbf{Z}_{\boldsymbol{\gamma}}^T)\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n-\mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}}\right)\\
&+\frac{n+\nu}{2}\log\left(\frac{1+\mathbf{Y}^T\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n-\mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}}\right).
\end{aligned}$$

Denote the above four terms by J_1, J_2, J_3, J_4. Again, J_1 has finite lower bound. By similar proof in Step 1 of Theorem 3.4, one can show that w.p.a.1, for $m\in[m_1,m_2]$ and c_j's $\in[\underline{\phi}_n,\bar{\phi}_n]$, $0\le -J_3 = O_P(1)$.

To analyze J_4, note $J_4 = \frac{n+\nu}{2}\log(1+\frac{\mathbf{Y}^T\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Y}}{1+\mathbf{Y}^T(\mathbf{I}_n-\mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}})$. Let $\boldsymbol{\nu}_{\boldsymbol{\gamma},m} = \mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}^0}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0}$. It can be directly examined by property of $\mathbf{P}_{\boldsymbol{\gamma}}$ that $\boldsymbol{\nu}_{\boldsymbol{\gamma},m} = \mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}+\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}$. By Assumption B.1 and $\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\|^2 \le b_0'\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2$, i.e., (8.10), we have

$$|(\boldsymbol{\beta}^0_{\boldsymbol{\gamma}})^T\mathbf{Z}_{\boldsymbol{\gamma}}^T\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}| \le \|\mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|\cdot\sqrt{\rho_n b_0'}\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|.$$

Meanwhile, $\|\mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2 \ge nd_0^{-1}\|\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2$. Therefore, by $\rho_n = o(n)$, it can be shown that

$$\begin{aligned}
\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|^2 &= \|\mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2\left(1+\frac{2(\boldsymbol{\beta}^0_{\boldsymbol{\gamma}})^T\mathbf{Z}_{\boldsymbol{\gamma}}^T\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}}{\|\mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2}+\frac{\|\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{Z}_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}^0\setminus\boldsymbol{\gamma}}\|^2}{\|\mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2}\right)\\
&= \|\mathbf{Z}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}}\|^2(1+o(1)),
\end{aligned}$$

for all $m\in[m_1,m_2]$. Since for each $m\in[m_1,m_2]$, $\boldsymbol{\nu}^T_{\boldsymbol{\gamma},m}\boldsymbol{\epsilon}/\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\| \sim N(0,\sigma_0^2)$, we obtain $\max_{m\in[m_1,m_2]}|\boldsymbol{\nu}^T_{\boldsymbol{\gamma},m}\boldsymbol{\epsilon}|/\|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\| = O_P(\sqrt{\log m_2})$. Also note, w.p.a.1,

for $m \in [m_1, m_2]$, $|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}^T \boldsymbol{\eta}| \le \|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\| \cdot \|\boldsymbol{\eta}\| \le \sqrt{C_\beta n s_n^2 m_1^{-a} h_{m_1}} \|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|$, thus we obtain that

$$\begin{aligned}\mathbf{Y}^T \mathbf{P}_{\boldsymbol{\gamma}} \mathbf{Y} &\ge \|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|^2 + 2\boldsymbol{\nu}_{\boldsymbol{\gamma},m}^T \tilde{\boldsymbol{\eta}} \\ &= \|\boldsymbol{\nu}_{\boldsymbol{\gamma},m}\|^2 \left(1 + O_P\left(\sqrt{\frac{n s_n^2 m_1^{-a} h_{m_1} + \log m_2}{n\psi_n^2}}\right)\right) \\ &\ge n d_0^{-1} \|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^0\|^2 (1 + o_P(1)).\end{aligned}$$

To approximate $\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y}$, let $\tilde{\boldsymbol{\nu}}_{\boldsymbol{\gamma},m} = (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Z}_{\boldsymbol{\gamma}^0 \setminus \boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}^0 \setminus \boldsymbol{\gamma}}^0$. It can be verified that $\max_{m \in [m_1,m_2]} |\tilde{\boldsymbol{\nu}}_{\boldsymbol{\gamma},m}^T \boldsymbol{\epsilon}| / \|\tilde{\boldsymbol{\nu}}_{\boldsymbol{\gamma},m}\| = O_P(\sqrt{\log m_2})$, and, by Assumption B.1, we have $\|\tilde{\boldsymbol{\nu}}_{\boldsymbol{\gamma},m}\|^2 \ge (n d_0^{-1} - \rho_n)\|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0 \setminus \boldsymbol{\gamma}}^0\|^2$, and $\|\tilde{\boldsymbol{\nu}}_{\boldsymbol{\gamma},m}\|^2 \le n d_0 \|\boldsymbol{\beta}_{\boldsymbol{\gamma}^0 \setminus \boldsymbol{\gamma}}^0\|^2 \le n d_0 b_0' \|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^0\|^2$. Therefore, it can be shown by direct calculation that w.p.a.1, for $m \in [m_1, m_2]$,

$$\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{Y} \le (\|\tilde{\boldsymbol{\nu}}_{\boldsymbol{\gamma},m}\|^2 + n\sigma_0^2)(1 + o_P(1)) \le (n d_0 b_0' \|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^0\|^2 + n\sigma_0^2)(1 + o_P(1)).$$

Therefore, w.p.a.1, for $m \in [m_1, m_2]$,

$$\begin{aligned}J_4 &\ge \frac{n+\nu}{2} \log\left(1 + \frac{n d_0^{-1} \|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^0\|^2 (1 + o(1))}{(n d_0 b_0' \|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^0\|^2 + n\sigma_0^2)}\right) \\ &\ge \frac{n+\nu}{2} \log\left(1 + \frac{1 + o(1)}{d_0^2 b_0'} \cdot \frac{\psi_n^2}{\psi_n^2 + \zeta_0}\right),\end{aligned}$$

where $\zeta_0 = \sigma_0^2/(d_0 b_0')$ and the last inequality follows by $\|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^0\|^2 \ge \psi_n^2$. Therefore, we can obtain that $J_4 \ge \frac{n+\nu}{2} \log(1 + \frac{1+o(1)}{d_0^2 b_0'} \cdot \min\{1/2, \psi_n^2/(2\zeta_0)\})$.

Finally, by the proof of (8.8), it can be shown that w.p.a.1, for $m \in [m_1, m_2]$, $J_2 \ge -\frac{m_2 s_n}{2} \log(1 + d_0 n \bar{\phi}_n \tau_1^2)$. So by (8.9), w.p.a.1, for $m \in [m_1, m_2]$, as $n \to \infty$,

$$\begin{aligned}&\frac{p(\emptyset|\mathbf{D}_n)}{p(\boldsymbol{\gamma}|\mathbf{D}_n)} \\ \le\ & \tilde{C} \exp\left(\frac{m_2 s_n}{2} \log(1 + d_0 n \bar{\phi}_n \tau_1^2)\right. \\ & \left. - \frac{n+\nu}{2} \log\left(1 + \frac{1 + o(1)}{d_0^2 b_0'} \cdot \min\{1/2, \psi_n^2/(2\zeta_0)\}\right)\right) \to 0,\end{aligned}$$

where $\tilde{C}$ is a large positive constant. This completes the proof.

References

[1] Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 870–897. MR2065192

[2] Berger, J. O. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122. MR1394065

[3] Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference* **112**, 241–258. MR1961733
[4] Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Annals of Statistics* **31**, 536–559. MR1983541
[5] Buldygin, V. and Kozachenko, Y. (2000). *Metric Characterization of Random Variables and Random Processes*. Providence, RI: American Mathematical Society. MR1743716
[6] Courant, R. and Hilbert, D. (1953). *Methods of Mathematical Physics*, Volume **1**. New York: Interscience Publischer, Inc. MR0065391
[7] Casella, C., Girón, F. J., Martínez, M. L. and Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *Annals of Statistics* **37**, 1207–1228. MR2509072
[8] Chipman, H., George, E. and McCulloch, R. (2010). BART: Bayesian adaptive regression trees. *Annals of Applied Statistics* **4**, 266–298. MR2758172
[9] Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–401. MR1649120
[10] Curtis, M., Banerjee, S. and Ghosal, S. (2014). Fast Bayesian model assessment for nonparametric additive regression. *Computational Statistics & Data Analysis* **71**, 347–358. MR3131975
[11] Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 2197–2202. MR1963681
[12] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of American Statistical Association* **116**, 544–557. MR2847969
[13] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of Royal Statistical Society B* **70**, 849–911. MR2530322
[14] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148. MR2640659
[15] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics* **38**, 3567–3604. MR2766861
[16] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional variable selection: Beyond the lienar model. *Journal of Machine Learning Research* **10**, 1829–1853. MR2550099
[17] Fernández, C., Ley, E. and Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**, 381–427. MR1820410
[18] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis* (2nd ed). Chapman & Hall/CRC. MR1385925
[19] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. MR1380810

[20] Green, P. and Hastie, D. (2009). Reversible jump MCMC. Technical Report, University of Bristol.
[21] Girón, F. J., Moreno, E., Casella, G. and Martínez, M. L. (2010). Consistency of objective Bayes factors for nonnested linear models and increasing model dimension. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* **104**, 57–67. MR2666441
[22] Godsill, J. S. and Rayner, P. J. W. (1998). Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler. *IEEE Trans. Speech Audio Process* **6**, 352–372.
[23] Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations* (2nd ed). John Hopkins Univ. Press, Baltimore. MR1002570
[24] Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* **36**, 587–613. MR2396808
[25] Huang, J., Horowitz, J. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics* **38**, 2282–2313. MR2676890
[26] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. MR1082147
[27] Hsu, D., Kakade, S. M. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communication in Probability* **17**, 1–6. MR2994877
[28] Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Annals of Statistics* **35**, 1487–1511. MR2351094
[29] Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Annals of Statistics* **36**, 2207–2231. MR2458185
[30] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107**, 649–660. MR2980074
[31] Koltchinskii, V. and Yuan, M. (2008). Sparse recovery in large ensembles of kernel machines. 21 st Annual Conference on Learning Theory-COLT 2008, Helsinki, Finland, July 9–12, 2008, R. A. Servedio and T. Zhang (eds.), Omnipress, pp. 229–238.
[32] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics* **37** 3498–3528. MR2549567
[33] Lv, J. and Liu, J. S. (2013). Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B*, to appear.
[34] Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410–423. MR2420243

[35] Liang, F., Song, Q. and Yu, K. (2013). Bayesian subset modeling for high dimensional generalized linear models. *Journal of the American Statistical Association*, in press. MR3174644
[36] Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics* **34**, 2272–2297. MR2291500
[37] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34**, 1436–1462. MR2278363
[38] Meier, L., van de Geer, S. and Buehlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37**, 3779–3821. MR2572443
[39] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37**, 246–270. MR2488351
[40] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B* **71**, 1009–1030. MR2750255
[41] Stone, C. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**, 689–705. MR0790566
[42] Shang, Z. and Clayton, M. K. (2011). Consistency of Bayesian model selection for linear modelswith a growing number of parameters. *Journal of Statistical Planning and Inference* **11**, 3463–3474. MR2817355
[43] Shang, Z. and Clayton, M. K. (2012). An application of Bayesian variable selection to spatial concurrent linear models. *Environmental and Ecological Statistics* **19**, 521–544. MR2999671
[44] Shang, Z. and Li, P. (2014). Bayesian ultrahigh-dimensional screening via MCMC. *Journal of Statistical Planning and Inference*, in press.
[45] Scheipl, F., Fahrmeir, L. and Kneib, T. (2012). Spike-and-slab priors for function selection in structured regression models. *Journal of the American Statistical Association* **107**, 1518–1532. MR3036413
[46] Sabanés Bové, D., Held, L. and Kauermann, G. (2011). Mixtures of g-priors for generalised additive model delection with penalised splines. Technical Report, University of Zurich.
[47] Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis* (2nd ed). Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR1958247
[48] Shen, X., Pan, W., Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association* **107**, 223–232. MR2949354
[49] Wolfe, P. J., Godsill, S. J. and Ng, W.-J. (2004). Bayesian variable selection and regularization for time-frequency surface estimation. *Journal of the Royal Statistical Society, Series B* **66**, 575–589. MR2088291
[50] van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics* **36**, 614–645. MR2396809
[51] Xue, L. and Zou, H. (2011). Sure independence screening and compressed random sensing. *Biometrika* **98**, 371–380. MR2806434

[52] Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics* **40**, 2541–2571.

[53] Yang, Y. and Zou, H. (2013). A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions. *Statistics and Its Interface* **6**, 167–173. MR3066682

[54] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics* **36**, 1567–1594. MR2435448

[55] Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, A. Zellner (ed.), Edward Elgar Publishing Limited, pp. 389–399.

[56] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541–2567. MR2274449