Submitted to Bernoulli
arXiv: 0000.0000

Bayesian Linear Regression for Multivariate Responses Under Group Sparsity

BO NING[1], SEONGHYUN JEONG[2], and SUBHASHIS GHOSAL[2†]

[1]*Department of Statistics and Data Science, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA*

[2]*Department of Statistics, North Carolina State University, 4276 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695, USA*
E-mail: [*]bo.ning@yale.edu; [**]sjeong4@ncsu.edu; [†]sghosal@ncsu.edu

We study frequentist properties of a Bayesian high-dimensional multivariate linear regression model with correlated responses. The predictors are separated into many groups and the group structure is pre-determined. Two features of the model are unique: (i) group sparsity is imposed on the predictors. (ii) the covariance matrix is unknown and its dimensions can also be high. We choose a product of independent spike-and-slab priors on the regression coefficients and a new prior on the covariance matrix based on its eigendecomposition. Each spike-and-slab prior is a mixture of a point mass at zero and a multivariate density involving a $\ell_{2,1}$-norm. We first obtain the posterior contraction rate, the bounds on the effective dimension of the model with high posterior probabilities. We then show that the multivariate regression coefficients can be recovered under certain compatibility conditions. Finally, we quantify the uncertainty for the regression coefficients with frequentist validity through a Bernstein-von Mises type theorem. The result leads to selection consistency for the Bayesian method. We derive the posterior contraction rate using the general theory by constructing a suitable test from the first principle using moment bounds for certain likelihood ratios. This leads to posterior concentration around the truth with respect to the average Rényi divergence of order 1/2. This technique of obtaining the required tests for posterior contraction rate could be useful in many other problems.

Keywords: Rényi divergence, Bayesian variable selection, covariance matrix, group sparsity, multivariate linear regression, posterior contraction rate, spike-and-slab prior.

1. Introduction

Asymptotic behaviors of variable selection methods for linear regression were extensively studied (Bühlmann and van der Geer, 2011). However, theoretical studies on Bayesian variable selection methods were limited to relatively simple settings (Castillo et al., 2015; Chae et al., 2019; Martin et al., 2017; Ročková, 2018; Belitser and Ghosal, 2019; Song and Liang, 2017). For example, Castillo et al. (2015) studied a sparse linear regression model in which the response variable is one-dimensional and the variance is known. However,

[†]Research is partially supported by NSF grant number DMS-1510238.

it is not straightforward to extend those results to multivariate linear regression with unknown covariance matrix (or even the univariate case with unknown variance).

Predictors can often be naturally clustered in groups, as in the following examples.

1. *Cancer genomics study.* The relationship between clinical phenotypes and DNA mutations is an important issue for biologists. DNA mutations are detected by DNA sequencing. Since these mutations are spaced linearly along the DNA sequence, it is often assumed that the adjacent DNA mutations on the chromosome have a similar genetic effect and should be grouped together (Li and Zhan, 2010).
2. *Multi-task learning.* When information for multiple tasks is shared, solving tasks simultaneously is desirable to improve learning efficiency and prediction accuracy. Relevant information is preserved across different equations by grouping them together (Lounici et al., 2009).
3. *Causal inference in advertising.* When measuring the effectiveness of an advertising campaign running on stores, counterfactuals need to be constructed using the sales data at some control stores chosen by a variable selection method (Ning et al., 2018). Stores within the same geographical region share the same demographic information, and so can be grouped together before selection.

Driven by those applications, new variable selection methods designed to select or not select variables as groups were developed by imposing *group-sparsity* on the regression coefficients as in the group-lasso (Yuan and Lin, 2006). This method replaces the ℓ_1-norm in the penalty term of the lasso with the $\ell_{2,1}$-norm, which comprises of the ℓ_2-norm put on the predictors within each group and the ℓ_1-norm is put across the groups. Theoretical properties of the group-lasso were studied, and its benefits over the lasso in the group selection problem were demonstrated (Nardi and Rinaldo, 2008; Lounici et al., 2009, 2011; Huang and Zhang, 2010). Recently, various Bayesian methods for selecting variables as groups were also proposed (Li and Zhan, 2010; Curtis et al., 2014; Ročková and Lesaffre, 2014; Xu and Ghosh, 2015; Chen et al., 2016; Greenlaw et al., 2017; Liquet et al., 2017). However, their large-sample frequentist properties are largely unknown.

In this paper, we study a Bayesian method for the multivariate linear regression model with two distinct features: group-sparsity imposed on the regression coefficients and an unknown covariance matrix. To the best of our knowledge, even in a simpler setting without the group-sparsity structure, convergence and selection properties of methods for high-dimensional regression with a multivariate response having an unknown covariance matrix have not been studied in either the frequentist or the Bayesian literature. However, it is important to understand the theoretical properties of these methods because correlated responses arise in many applications. For example, in the study of the causal effect of an advertising campaign, sales in different stores are usually spatially correlated (Ning et al., 2018). Furthermore, when the dimension of the covariance matrix is large, it would affect the quality of the estimation of the regression coefficients.

When the covariance matrix is unknown and high-dimensional, standard techniques for posterior concentration rates (Castillo et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2019) cannot be applied. Also, the general theory of posterior contraction under the average squared Hellinger distance (Ghosal and van der Vaart, 2017) is not sufficient to obtain the rate in terms of the Euclidean metric on the regression parameter. In order to obtain that rate through the general theory, we shall construct certain required tests directly by controlling the moments of likelihood ratios with the parameter space broken up in small pieces. This leads to the posterior contraction rate with respect to the negative average log-affinity, which can be subsequently converted to the rate with respect to the Euclidean metric on the regression parameter. The technique of controlling error probabilities by a moment bound on likelihood ratios appears to be new in the Bayesian literature and may be useful to study rates in other problems.

In this paper, we consider a multivariate linear regression model

$$Y_i = \sum_{j=1}^{G} X_{ij}\boldsymbol{\beta}_j + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where Y_i is a $1 \times d$ response variable, $i = 1, \ldots, n$, X_{ij} is a $1 \times p_j$ predictor variable, $j = 1, \ldots, G$, $\boldsymbol{\beta}_j$ is a $p_j \times d$ matrix containing the regression coefficients, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed (i.i.d) as $\mathcal{N}(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being a $d \times d$ unknown covariance matrix. In other words, in the regression model, there are $G > 1$ non-overlapping groups of predictor variables with the group structure being predetermined. We denote the groups which contain at least a non-zero coordinate as *non-zero groups* and the remaining groups as *zero groups*. The number of total groups G is clearly bounded by p. When $G = p$, it reduces to the setting that the sparsity is imposed on the individual coordinates. Thus the results derived in our paper are also applicable to the ungrouped setting.

The above model can be rewritten in the vector form as

$$Y_i = X_i\boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_G')'$ is a $p \times d$ matrix, where $p = \sum_{j=1}^{G} p_j$, and $X_i = (X_{i1}, \ldots, X_{iG})$ is a $1 \times p$ vector. The dimension p can be very large and the dimension d can be large as well, but to a lesser extent.

To allow derivation of asymptotic properties of estimation and selection, certain conditions on the growth of p, G, d and $p_1, \ldots, p_G$ need to be imposed. The dimension p can grow at a rate faster than the sample size n, but we require that the total number of the coefficients in all non-zero groups together are less than n in order. We further assume that the number of coordinates in any single group must be of the order less than n, $G \geq n^c$, for some positive constant c, and $\log G$ grows slower than n. Finally, to make the covariance matrix consistently estimable, we assume that for the dimension d of the covariance matrix, $d^2 \log n$ grows at a rate slower than n.

As for the priors, we choose the product of d independent spike-and-slab priors for $\boldsymbol{\beta}$ and a prior on $\boldsymbol{\Sigma}$ through its eigendecomposition. The latter seems to be a new addition to the literature. The spike-and-slab prior is a mixture of point mass for the zero coordinates and a density for non-zero coordinates. In the ungrouped setting, commonly used densities for non-zero coordinates are a Laplace density (Castillo et al., 2015), a Cauchy density (Castillo and Mismer, 2018) and a normal density with mean chosen by empirical Bayes methods (Martin et al., 2017; Belitser and Ghosal, 2019). In this paper, we choose a density for the non-zero coordinates involving the $\ell_{2,1}$-norm (see (2.1)), which corresponds to the penalty function of the group-lasso. We derive an explicit expression for the normalizing constant of this density.

We shall use the following notations. We assume that $\mathcal{G}_1, \dots, \mathcal{G}_G$ are G disjoint groups such that $\cup_{j=1}^G \mathcal{G}_j = \{1, \dots, p\}$. Since these groups are given and will be kept the same throughout, their notations will be dropped from subscription notations. Each p_j is the number of elements in $\mathcal{G}_j$. Let $p_{\max} = \max\{p_j : 1 \le j \le G\}$. For each $k = 1, \dots, d$, let $S_k \subseteq \{1, \dots, G\}$ stand for the collection of indices of non-zero groups for the kth component and $s_k = |S_k|$ stand its cardinality. Let $S_{0,k}$ be the set consisting of the indices of the true non-zero groups. Let $S = \{S_1, \dots, S_d\}$ be the d-tuple of the model indices, and define $s = \sum_{k=1}^d s_k$, $p_{S_k} = \sum_{j \in S_k} p_j$, and $p_S = \sum_{k=1}^d p_{S_k}$. Similar notations are used for the corresponding true values $S_{0,k}, s_{0,k}, S_0, s_0, p_{S_{0,k}}$ and p_{S_0}. We also define $S_{\boldsymbol{\beta},k}$, $s_{\boldsymbol{\beta},k}$, $S_{\boldsymbol{\beta}}$, and $s_{\boldsymbol{\beta}}$ for an arbitrary $p \times d$ matrix $\boldsymbol{\beta}$.

For a vector A, let $\|A\|_1$, $\|A\|_{2,1}$ and $\|A\|$ be the ℓ_1-, $\ell_{2,1}$- and ℓ_2-norm of A, respectively, where $\|A\|_{2,1} = \sum_{j=1}^G \|A_j\|$ with A_j being the subvector of A consisting of $k \in \mathcal{G}_j$ coordinates. For a matrix $\boldsymbol{B}$, let modB_k be the kth column of $\boldsymbol{B}$, by $\|\boldsymbol{B}\|_F = \sqrt{\mathrm{Tr}(\boldsymbol{B}^T\boldsymbol{B})}$ as the Frobenius norm, and $\|\boldsymbol{B}\|$ as the spectral norm. In particular, for an $n \times p$ matrix $\boldsymbol{C}$, we define the matrix norm $\|\boldsymbol{C}\|_\circ = \max\{\|\boldsymbol{C}_j\| : 1 \le j \le G\}$, where $\boldsymbol{C}_j$ is the submatrix of $\boldsymbol{C}$ consisting of columns C_k with $k \in \mathcal{G}_j$ coordinates. For a $d \times d$ symmetric positive definite matrix $\boldsymbol{D}$, let $\mathrm{eig}_1(\boldsymbol{D}), \dots, \mathrm{eig}_d(\boldsymbol{D})$ denote the eigenvalues of $\boldsymbol{C}$ ordered from the smallest to the largest and $\det(\boldsymbol{D})$ stands for the determinant of $\boldsymbol{D}$. For a scalar c, we denote $|c|$ to be the absolute value of c.

Let $\rho(f,g) = -\log(\int f^{1/2} g^{1/2} d\nu)$ be the Rényi divergence of order $1/2$ between densities f and g and $h^2(f,g) = \int (f^{1/2} - g^{1/2})^2 d\nu$ be their squared Hellinger distance. The Kullback-Leibler divergence and the Kullback-Leibler variation between f and g are respectively given by $K(f,g) = \int f \log(f/g)$ and $V(f,g) = \int f(\log(f/g) - K(f,g))^2$. The notation $\|\mu - \nu\|_{TV}$ denotes the total variation distance between two probability measures μ and ν.

We let $N(\epsilon, \mathcal{F}, \rho)$ stand for the ϵ-covering number of a set $\mathcal{F}$ with respect to a metric ρ, which is the minimal number of ϵ-balls in ρ-metric needed to cover the set $\mathcal{F}$. Let $\boldsymbol{I}_d$ stand for the d dimensional identity matrix and $\mathbb{1}$ stand for the indicator function.

The symbols $\lesssim$ and $\gtrsim$ will be used to denote inequality up and down to a constant while $a \asymp b$ stand for $C_1 a \le b \le C_2 a$ for two constants C_1 and C_2. The notations $a \ll b$ and $a \vee b$ stand for $a/b \to 0$ and $\max\{a,b\}$ respectively. The symbol $\delta_0(\cdot)$ stands for the probability measure with all its mass at 0.

The remainder of the paper is organized as follows. Section 2 describes the priors, along with the necessary assumptions. Section 3 provides the main results. Section 4 discusses algorithms for computation. The proofs of two main theorems are given in Section 5. The supplementary material gives an auxiliary result and presents analogous but slightly weaker results on posterior contraction and selection using a conjugate inverse-Wishart prior on the covariance matrix.

2. Prior specifications

In this section, we introduce the priors used in this paper. We let $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ be independently distributed in the prior. The prior for $\boldsymbol{\beta}$ is mixed over several dimensions and each component of the prior density depends on the $\ell_{2,1}$-norm of $\boldsymbol{\beta}$, while a spike-and-slab prior is put on the group dimension. We put a prior on the covariance matrix through its eigendecomposition $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'$, with independent inverse Gaussian priors for each diagonal entry of $\boldsymbol{D}$ and the uniform prior for $\boldsymbol{P}$ on the group of orthogonal matrices.

2.1. Prior for regression coefficients

We denote the kth column of $\boldsymbol{\beta}$ by β_k and let the notations β_{k,S_k} and β_{k,S_k^c} stand for the collection of regression coefficients in the kth column of the non-zero groups and the zero groups respectively. A spike-and-slab prior is constructed as follows. First, the dimension s is chosen from a prior π on the set $\{0, 1, \ldots, Gd\}$. Next, a d-tuple S of subsets is randomly chosen from the set $\{1, \ldots, G\}^d$ such that $\sum_{k=1}^d s_k = s$. Finally, for each k, a vector β_{k,S_k} is independently chosen from a probability density g_{S_k} on $\mathbb{R}^{p_{S_k}}$ given by (2.2), and the remaining coordinates β_{k,S_k^c} set to 0. To summarize, the prior for $\boldsymbol{\beta}$ is

$$(S, \boldsymbol{\beta}) \to \pi(s) \frac{1}{\binom{Gd}{s}} \prod_{k=1}^{d} g_{S_k}(\beta_{k,S_k}) \delta_0(\beta_{k,S_k^c}), \tag{2.1}$$

where the density $\pi(s)$ is the prior for the dimension s.

Assumption 1 (Prior on dimension). *For some constants $A_1, A_2, A_3, A_4 > 0$,*

$$\frac{A_1}{(G \vee n^{p_{\max}})^{A_3}} \le \frac{\pi(s)}{\pi(s-1)} \le \frac{A_2}{(G \vee n^{p_{\max}})^{A_4}}, \qquad s = 1, \ldots, Gd.$$

If sparsity is imposed at the individual level, i.e. $p_{\max} = 1$, then the assumption is identical to the one given in Castillo et al. (2015). Prior distributions satisfying the assumption can easily be constructed. For example, the complexity prior given by Castillo et al. (2015) satisfies the above assumption if $p_{\max} = 1$, and it can also be easily modified to consider the case when $p_{\max} > 1$.

When sparsity is at the individual level, the Laplace density (Castillo et al., 2015) or the Cauchy density (Castillo and Mismer, 2018) is generally chosen for g, since the normal density has a too sharp tail that overshrinks the non-zero coefficients, although some empirical Bayes modifications of the mean can overcome the issue (see Martin et al., 2017; Belitser and Ghosal, 2019). However, in our setting, as sparsity is imposed at the group level, like the group lasso, we consider the following density using the $\ell_{2,1}$-norm:

$$g_{S_k}(\beta_{k,S_k}) = \left(\prod_{j \in S_k} \left(\frac{\lambda_k}{a_j} \right)^{p_j} \right) \exp\big(-\lambda_k \|\beta_{k,S_k}\|_{2,1} \big), \tag{2.2}$$

where $a_j = \sqrt{\pi}(\Gamma(p_j+1)/\Gamma(p_j/2+1))^{1/p_j} \geq 2$ (see Lemma 6.1 in the supplementary materials). This density has its tail lighter than the corresponding Laplace density. From Stirling's approximation, it follows that $a_j = O(p_j^{1/2})$. A relevant elliptical prior distribution is considered in Gao et al. (2015).

A prior of this type involving the $\ell_{2,1}$-norm was also used in the Bayesian literature in group-sparsity problems (Xu and Ghosh, 2015), but an explicit expression of the normalizing constant was not obtained. Since the normalizing constant depends on the dimension, its value will play a role in the posterior contraction rate.

The tuning parameter λ_k in the prior needs to be bounded both from above and below, specified in Assumption 2 below. A value too large will shrink the non-zero coordinates too much towards to 0. A value too small will be unable to prevent many false signals appearing in the model, which can make the posterior to contract slower.

Assumption 2. *For some constants $B_1, B_2, B_3 > 0$ and each $k = 1, \dots, d$, $\underline{\lambda} \leq \lambda_k \leq \overline{\lambda}$, where*

$$\underline{\lambda} = \frac{\|\boldsymbol{X}\|_\circ}{B_1(G^{1/p_{\max}} \vee n)^{B_2}} \quad \overline{\lambda} = B_3 \|\boldsymbol{X}\|_\circ \sqrt{\log G \vee p_{\max} \log n}. \tag{2.3}$$

The constants B_1, B_2, B_3 can be chosen large enough so that the range can be sufficiently wide. In particular, if $p_{\max} = 1$, this above reduces to the one in Castillo et al. (2015).

Assumption 2 will be coupled with Assumption 3 in Section 3.1 on the true parameters. A particularly interesting case is that every λ_k is set to the lower bound $\underline{\lambda}$ for every k. Then the bound requirement on the true signal will be rather mild.

2.2. Prior for the covariance matrix

For a prior on the covariance matrix $\boldsymbol{\Sigma}$, we use its eigendecomposition $\boldsymbol{PDP}'$. We put an inverse Gaussian prior independently on each eigenvalue of $\boldsymbol{\Sigma}$, or equivalently, on each diagonal entry of $\boldsymbol{D}$. This prior is chosen because of its exponentially decaying tail on both sides. The orthogonal matrix $\boldsymbol{P}$ is given the Haar measure on the compact Lie group of $d \times d$ orthogonal matrices, which is a Riemannian manifold of dimension $d(d-1)/2$ embedded in $\mathbb{R}^{d\times d}$.

We found that the naturally conjugate inverse Wishart prior on $\boldsymbol{\Sigma}$ may induce a suboptimal posterior contraction rate due to its weaker tail property when d increases to infinity. Nevertheless, because of the practical importance of this prior, we present the contraction rate for this prior in the supplementary material. When d is fixed, the rate is the same as in the main theorem in this paper using the above stated prior on $\boldsymbol{\Sigma}$. When additional structure like sparsity are assumed on large covariance or precision (inverse covariance) matrices, prior distributions can be assigned by respecting such structure (Banerjee and Ghosal, 2014, 2015; Pati et al., 2014). In such a situation, an improved rate may be possible; see the remark at the end of Section 3.1. Other significant priors used in the literature, such as reference priors (Yang and Berger, 1994; Sun and Berger, 2007), are harder to handle because the general theory of posterior contraction does not apply to these improper priors, and moreover, tail bounds for the corresponding eigenvalue distribution need to be available.

3. Main results

3.1. Posterior contraction rate

We study the posterior contraction rate for the model and the priors given in Section 2. We denote $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ as the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, respectively. Recall the notations $s_{0,k} = |S_{0,k}|$, $S_0 = \{S_{0,1}, \dots, S_{0,d}\}$, and $s_0 = \sum_{k=1}^{d} s_{0,k}$.

The general theory of posterior contraction for independent non-identically distributed observations (see Theorem 8.23 of Ghosal and van der Vaart, 2017) is often used to derive a posterior contraction rate. The general theory characterizes the contraction rate in terms of the average squared Hellinger distance by default, unless an additional testing property in the model is established. However, closeness in terms of the average squared Hellinger distance between multivariate normal densities with varying mean and an unknown covariance does not necessarily imply that the mean parameters in the two densities are also close on average in terms of the Euclidean distance. To alleviate the problem, we work directly with the average Rényi divergence of order $1/2$, which is still very tractable in the multivariate normal setting, and gives rise to closeness in terms of the desirable Euclidean distance. To this end, we directly construct a suitable test using the likelihood ratio for the null against some representative points in the alternative described by the complement of a Rényi ball around the null intersected with a sieve, and then showing that such a test also works well for testing the null value against a neighborhood of the representative point, by controlling the moments of the likelihood ratio of the representative point and the points in the neighborhood. Finally, by controlling the number of pieces needed to cover the sieve, we construct a single test with required control over the error probabilities for testing the null value against the whole of the alternative intersected with the sieve, which can then be used in the general theory of posterior contraction.

The general theory also requires lower bounds for prior concentration near the true parameter value, which is possible provided that we require the true values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ to be restricted into certain regions (see Assumption 3 below). This is unlike Castillo et al. (2015), who obtained results uniformly over the whole space as their case (univariate with known variance and Laplace prior) allows explicit expressions for direct treatment.

Assumption 3. *The true values satisfy* $\boldsymbol{\beta}_0 \in \mathcal{B}_0$ *and* $\boldsymbol{\Sigma}_0 \in \mathcal{H}_0$, *for*

$$\mathcal{B}_0 = \left\{ \boldsymbol{\beta} : \sum_{k=1}^{d} \|\beta_k\|_{2,1} \le \overline{\beta} \right\}, \quad \mathcal{H}_0 = \{ \boldsymbol{\Sigma} : b_1 \boldsymbol{I}_d \le \boldsymbol{\Sigma} \le b_2 \boldsymbol{I}_d \}, \tag{3.1}$$

where $b_1, b_2 > 0$ *are fixed values and* $\overline{\beta} = s_0(\log G \vee p_{\max} \log n)/\max\{\lambda_k : 1 \le k \le d\}$.

The largest value of $\overline{\beta}$ is obtained by taking $\lambda_k = \underline{\lambda}$ for all k. In this case, the upper bound becomes $\overline{\beta} = B_1 s_0(\log G \vee p_{\max} \log n)(G^{1/p_{\max}} \vee n)^{B_2}/\|\boldsymbol{X}\|_\circ$, which is a very mild restriction if B_2 is chosen large enough.

Theorem 3.1. *For the model (1.1) and the priors given in Section 2, we have that for a sufficiently large $M_1 > 0$,*

$$\sup_{\boldsymbol{\beta}_0 \in \mathcal{B}_0, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi\Big(\boldsymbol{\beta} : \|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_F^2 \geq M_1 n \epsilon_n^2 \Big| Y_1, \dots, Y_n\Big) \to 0, \tag{3.2}$$

$$\sup_{\boldsymbol{\beta}_0 \in \mathcal{B}_0, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi\Big(\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_F^2 \geq M_1 \epsilon_n^2 \Big| Y_1, \dots, Y_n\Big) \to 0, \tag{3.3}$$

where

$$\epsilon_n = \max\left\{\sqrt{\frac{s_0 \log G}{n}}, \sqrt{\frac{s_0 p_{\max} \log n}{n}}, \sqrt{\frac{d^2 \log n}{n}}\right\} \to 0. \tag{3.4}$$

Remark 1. Unlike in the classical approach where variable selection is regulated by a penalty function that corresponds to a prior density on the regression coefficients, in the Bayesian approach, sparsity is imposed by the prior on the dimension. The prior density on the regression coefficients still plays a significant role in controlling the prior concentration and the tail behavior, but to a lesser extent. Thus, instead of using the prior given in (2.2), one can also choose a Laplace density for the coordinates in the non-zero groups. Then the $\ell_{2,1}$-norm of $\beta_{0,k}$, $\|\beta_{0,k}\|_{2,1}$, in the set $\mathcal{B}_0$ should be replaced by $\|\beta_{0,k}\|_1$. Clearly, $\|\beta_{0,k}\|_{2,1} \leq \|\beta_{0,k}\|_1$, and hence in the latter case, the set $\mathcal{B}_0$ will be smaller.

Remark 2. When $G = p$, and hence $p_{\max} = 1$, the posterior contraction rate simplifies to $\epsilon_n = \max\{\sqrt{(s_0 \log p)/n}, \sqrt{(d^2 \log n)/n}\}$. The first term in the rate is the same as the rate obtained when the sparsity is imposed at the individual level, such as in Bühlmann and van der Geer (2011) and Castillo et al. (2015). When $G \ll p$, the same rate can be obtained if $p_{\max} \log n \lesssim \log G$.

The first term of the rate in Theorem 3.1 coincides with the rate obtained for a group-lasso estimator of the multi-task learning problem studied by Lounici et al. (2011). Their setup is not directly comparable with ours but their analogous rate coincides with ours up to a logarithmic factor and they showed its optimality in a minimax sense. Uder the setting $d = 1$, the rate obtained in Huang and Zhang (2010) is $(p_{S_0} + s_0 \log G)/n$, which is only slightly faster than our rate, and will coincide with ours up to the logarithmic factor whenever $p_{S_0} \asymp s_0 p_{\max}$. This can often happen provided that the non-zero groups are not consisting of a few large and the rest small groups.

If there is additional lower-dimensional structure in the orthogonal matrix $\boldsymbol{P}$, the last term in (3.4) may be improved, because in a lower-dimensional manifold, the prior concentration rate will be higher and the entropy estimates will be lower. The simplest such structure is the trivial situation $\boldsymbol{P} = \boldsymbol{I}$, which leads to diagonal covariance matrix and the reduction of d^2 to d. More generally, a block-diagonal structure with L non-overlapping blocks of size $d_1, \dots, d_L$, $\sum_{l=1}^{L} d_l = d$, will reduce d^2 to $\sum_{l=1}^{L} d_l^2$.

From Theorem 3.1, the posterior contraction rate slows down significantly if the dimension of the covariance is too high, but a better rate may be possible if a lower dimensional structures is present in the covariance of the precision matrix. For instance, if the responses are independent across components, then the model (1.1) can be written as d independent model with each one is

$$\sigma_k{}^{-1} Y_{ik} = \sigma_k{}^{-1} X_i \beta_k + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim \mathcal{N}(0,1).$$

Then one can estimate the parameters in the d models separately. The posterior concentration rate for each corresponding posterior becomes $\epsilon_n = (\sum_{k=1}^{d} \epsilon_{n,k}^2)^{1/2}$, where $\epsilon_{n,k} = \max\{\sqrt{(s_{0,k} \log G)/n}, \sqrt{(s_{0,k} p_{\max} \log n)/n}\}$ is the individual rates for the kth component, $k = 1, \dots, d$.

3.2. Dimensionality and recovery

In this section, we show dimensionality control and recovery properties of the the marginal posterior of $\boldsymbol{\beta}$.

Lemma 3.2 (Dimension). *For the model (1.1) and the priors given in Section 2, we have that for a sufficiently large number $M_2 > 0$,*

$$\sup_{\boldsymbol{\beta}_0 \in \mathcal{B}_0, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi\Big(\boldsymbol{\beta} : s_{\boldsymbol{\beta}} \geq M_2 s^\star \Big| Y_1, \dots, Y_n\Big) \to 0,$$

where $s^\star = s_0 \vee \{d^2 \log n / (\log G \vee p_{\max} \log n)\}$.

From Lemma 3.2, $s^\star > s_0$ if $d^2 \log n \gg s_0(\log G \vee p_{\max} \log n)$. This means that the support of the posterior can substantially overshoot the true dimension s_0. In the next corollary, we show that even when $s^\star > s_0$, the posterior is still able to recover $\boldsymbol{\beta}_0$ in terms of the distance to the truth.

Corollary 3.3 (Recovery). *For the model (1.1) and the priors given in Section 2, we have that for a sufficiently large constant $M_3 > 0$,*

$$\sup_{\boldsymbol{\beta}_0 \in \mathcal{B}_0, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi\left(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_F^2 \geq \frac{M_3 n \epsilon_n^2}{\|\boldsymbol{X}\|_\circ^2 \phi_{\ell_2}^2 (s_0 + M_2 s^\star)} \middle| Y_1, \dots, Y_n \right) \to 0, \tag{3.5}$$

where $\phi_{\ell_2}^2$ is the restricted eigenvalue (see Definition 3.4 below).

Definition 3.4 (Restricted eigenvalue). *The smallest scaled singular value of dimension $\tilde{s}$ is defined as*

$$\phi_{\ell_2}^2(\tilde{s}) = \inf\left\{ \frac{\|\boldsymbol{X}\boldsymbol{\beta}\|_F^2}{\|\boldsymbol{X}\|_\circ^2 \|\boldsymbol{\beta}\|_F^2}, \ 0 \leq s_{\boldsymbol{\beta}} \leq \tilde{s} \right\}. \tag{3.6}$$

As $p \gg n$, the smallest eigenvalue of the design matrix must be 0. The restricted eigenvalue condition keeps the smallest eigenvalue for the sub-matrix of the design matrix, corresponding to the coefficients within non-zero groups, bounded away from 0.

The results in terms of other norms for the difference between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ can be also derived by assuming different assumptions on the smallest eigenvalue for the sub-matrix of the design matrix. For example, by using the uniform compatibility condition (in Definition 3.5 below), we can conclude that for a sufficiently large number $M_4 > 0$,

$$\sup_{\boldsymbol{\beta}_0 \in \mathcal{B}_0, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi \left(\left(\sum_{k=1}^{d} \|\beta_k - \beta_{0,k}\|_{2,1} \right)^2 \geq \frac{M_4 s^\star n \epsilon_n^2}{\|\boldsymbol{X}\|_\circ^2 \phi_{\ell_{2,1}}^2 (s_0 + M_2 s^\star)} \middle| Y_1, \dots, Y_n \right) \to 0. \tag{3.7}$$

We omit the proof since it is almost identical to that of Corollary 3.3.

Definition 3.5 (Uniform compatibility, $\ell_{2,1}$-norm). *The $\ell_{2,1}$-compatibility number in vectors of dimension $\tilde{s}$ is defined as*

$$\phi_{\ell_{2,1}}^2(\tilde{s}) = \inf \left\{ \frac{s_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta}\|_F^2}{\|\boldsymbol{X}\|_\circ^2 (\sum_{k=1}^d \|\beta_k\|_{2,1})^2}, \; 0 \leq s_{\boldsymbol{\beta}} \leq \tilde{s} \right\}.$$

By the Cauchy-Schwarz inequality, $\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\beta}\|_F \geq \sum_{k=1}^d \|\beta_k\|_{2,1}$, and it follows that $\phi_{\ell_2}(\tilde{s}) \leq \phi_{\ell_{2,1}}(\tilde{s})$ for any $\tilde{s} \ll Gd$.

3.3. Distributional approximation

To establish selection consistency, Castillo et al. (2015) devised a key technique through a distributional approximation for the posterior distribution. As in a Bernstein-von Mises (BvM) theorem, the posterior distribution of the regression parameter is approximated by a relatively simpler distribution, but unlike in a traditional BvM theorem for increasing dimensional parameters (Ghosal, 1999, 2000; Bontemps, 2011) or low-dimensional functionals (de Jonge and van Zanten, 2013; Gao and Zhou, 2016), the approximating distribution is a mixture of multivariate normal instead of a single one.

To derive an appropriate distributional approximation, we rewrite the model (1.1) as

$$Y_i = \mathrm{Vec}(\boldsymbol{\beta})\tilde{\boldsymbol{X}}_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\mathrm{Vec}(\boldsymbol{\beta})$ is obtained by stacking all the columns of $\boldsymbol{\beta}$ into a pd-dimensional row vector, $\tilde{\boldsymbol{X}}_i = \boldsymbol{I}_d \otimes X_i'$ is a $pd \times d$ block-diagonal matrix. The log-likelihood function is given by

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\left(\det(\boldsymbol{\Sigma})\right) - \frac{1}{2}\sum_{i=1}^{n} \|\boldsymbol{\Sigma}^{-1/2}\left(Y_i - \mathrm{Vec}(\boldsymbol{\beta})\tilde{\boldsymbol{X}}_i\right)'\|^2. \tag{3.8}$$

For any measurable subset $\mathcal{B}$ of $\mathbb{R}^{p\times d}$, the marginal posterior distribution of $\boldsymbol{\beta}$ is

$$\Pi(\boldsymbol{\beta}\in\mathcal{B}|Y_1,\dots,Y_n)=\frac{\int\int_{\mathcal{B}}\exp\left(\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma})-\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0)\right)d\Pi(\boldsymbol{\beta})d\Pi(\boldsymbol{\Sigma})}{\int\int\exp\left(\ell(\boldsymbol{\beta},\boldsymbol{\Sigma})-\ell(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0)\right)d\Pi(\boldsymbol{\beta})d\Pi(\boldsymbol{\Sigma})},\tag{3.9}$$

with

$$d\Pi(\boldsymbol{\beta})=\sum_{S:s\le Gd}\frac{\pi(s)}{\binom{Gd}{s}}\prod_{k=1}^{d}\left\{\left(\prod_{j\in S_k}\left(\frac{\lambda_k}{a_j}\right)^{p_j}\right)\exp(-\lambda_k\|\beta_{k,S_k}\|_{2,1})d\beta_{k,S_k}\otimes\delta_{S_k^c}\right\}.$$

In the next theorem, we shall show that under certain conditions, the posterior probability $\Pi(\boldsymbol{\beta}\in\mathcal{B}|Y_1,\dots,Y_n)$ can be approximated by

$$\Pi^{\infty}(\boldsymbol{\beta}\in\mathcal{B}|Y_1,\dots,Y_n)=\frac{\int_{\mathcal{B}}\exp\{\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}_0)-\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0)\}dU(\boldsymbol{\beta})}{\int\exp\{\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}_0)-\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0)\}dU(\boldsymbol{\beta})},$$

where

$$dU(\boldsymbol{\beta})=\sum_{S:s\le M_2s^\star}\frac{\pi(s)}{\binom{Gd}{s}}\prod_{k=1}^{d}\left\{\left(\prod_{j\in S_k}\left(\frac{\lambda_k}{a_j}\right)^{p_j}\right)d\beta_{k,S_k}\otimes\delta_{S_k^c}\right\}.\tag{3.10}$$

This means that $\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma})$ can be replaced by $\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}_0)$ with the true $\boldsymbol{\Sigma}_0$ and the impact of the $\ell_{2,1}$-term in the prior density vanishes. Let $\tilde{\boldsymbol{X}}_{i,S}$ be the submatrix of $\tilde{\boldsymbol{X}}_i$ chosen by S, with its dimension $p_S\times d$. If $p_S\le n$ for a given S, the maximum likelihood estimator (MLE) for $\beta_S^\star=(\beta'_{1,S_1},\dots,\beta'_{d,S_d})'$ given the true covariance matrix $\boldsymbol{\Sigma}_0$ is unique. We denote the MLE and the information matrix as

$$\hat{\beta}_S^\star=\left(\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_{i,S}\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}'_{i,S}\right)^{-1}\left(\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_{i,S}\boldsymbol{\Sigma}_0^{-1}Y_i'\right),\quad\hat{\mathbb{I}}_S=\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_{i,S}\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}'_{i,S}.$$

Then we can also write

$$\Pi^{\infty}(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)\propto\sum_{S:s\le M_2s^\star}w_S^{\infty}\mathcal{N}\left(\hat{\beta}_S^\star,n^{-1}\hat{\mathbb{I}}_S^{-1}\right)\otimes\delta_{S^c},\tag{3.11}$$

where

$$w_S^{\infty}\propto\frac{\pi(s)}{\binom{Gd}{s}}\left(\prod_{k=1}^{d}\prod_{j\in S_k}\left(\frac{\lambda_k\sqrt{2\pi}}{a_j}\right)^{p_j}\right)\det\left(n\hat{\mathbb{I}}_S\right)^{-1/2}\exp\left\{\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{X}'_{i,S}\hat{\beta}_S^\star\|^2\right\},$$

with $\sum_S w_S^{\infty}=1$.

Before we formally state the theorem, we recall the notion of the *small* λ *regime* (see Castillo et al., 2015). Clearly, bounded λ-values belong to the small λ regime. In our setting, we say λ_k belongs to the *small* λ *regime* if $\max\{\lambda_k\epsilon_n\sqrt{s^\star n}/\|\boldsymbol{X}\|_\circ : 1\le k\le$

$d\} \to 0$. In this regime, the impact of the $\ell_{2,1}$-penalty vanishes, and hence the MLE $\hat{\beta}_S^\star$ is asymptotically unbiased and does not depend on the choice of different values of λ_k. When choosing the value of λ_k outside the small λ regime, this MLE is no longer asymptotically unbiased (see Theorem 11 of the supplementary material of Castillo et al. (2015)). In order to make the remainder of the approximation tend to zero, we also assume that $\epsilon_n^2\sqrt{s^\star n}\big(\sqrt{s^\star n\epsilon_n^2} \vee \sqrt{p_{\max} d^3 \log G}\big) \to 0$ for $\ell_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ to be replaced by $\ell_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}_0)$.

Theorem 3.6 (Distributional approximation)**.** *For the model (1.1), the priors given in Section 2 with λ in the small λ regime, and the sequence*

$$\delta_n(s_0) = \epsilon_n\sqrt{s^\star n}\max\left(\max\{\lambda_k : 1 \le k \le d\}/\|\boldsymbol{X}\|_\circ, \epsilon_n^2\sqrt{s^\star n}, \epsilon_n\sqrt{p_{\max} d^3 \log G}\right),$$

we have that any positive sequence $\eta_n \to 0$ and some positive constant $c > 0$,

$$\sup_{\substack{\boldsymbol{\beta}_0 \in \{\mathcal{B}_0 : \delta_n(s_0) < \eta_n, \\ \phi_{\ell_{2,1}}(s_0 + M_2 s^\star) > c\},\, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0}} \mathbb{E}_0\|\Pi(\boldsymbol{\beta} \in \cdot|Y_1, \ldots, Y_n) - \Pi^\infty(\boldsymbol{\beta} \in \cdot|Y_1, \ldots, Y_n)\|_{TV} \to 0.$$

3.4. Selection

In this section, we establish selection consistency using Bernstein-von Mises theorem of the previous section. We assume the dimension of the covariance and the coordinates in the non-zero groups are sufficiently small. We also assume the smallest signal cannot be too small, which is

$$\tilde{\mathcal{B}} = \left\{\boldsymbol{\beta} : \min\{\|\beta_{jk}\|^2 : j \in S_{0,k}, k = 1, \ldots, d\} \ge \frac{M_3 n\epsilon_n^2}{\|\boldsymbol{X}\|_\circ^2 \phi_{\ell_2}^2(s_0 + M_2 s^\star)}\right\}. \tag{3.12}$$

This condition can be viewed as the *Beta-min condition* under the group sparsity setting. The lower bound displayed in the condition is derived from (3.5). Unlike the *Beta-min condition* in Castillo et al. (2015) which requires each individual coordinate is bounded away from 0, our condition allows zero coordinates to be included in a non-zero group.

The Beta-min condition is not vacuous, in that the lower bound in (3.1) is smaller than the upper bound in (3.12). To see this, note that under the small λ regime, $(\max_k \lambda_k)^{-1} \gg \sqrt{s^\star n\epsilon_n^2}/\|\boldsymbol{X}\|_\circ$. Therefore, $\bar{\beta} \gg \sqrt{n\epsilon_n^2}/\|\boldsymbol{X}\|_\circ$, and the right side coincides with the lower bound up to a constant, establishing the claim.

We now complete this section by stating the following theorem.

Theorem 3.7 (Selection consistency)**.** *For the model (1.1), the priors given in Section 2, some positive constant $c > 0$, and some sequences $\eta_n \to 0$ and $s_n \le G^a$ with $a < A_4 - 3/2$, we have that*

$$\sup_{\substack{\boldsymbol{\beta}_0 \in \{\mathcal{B}_0 \cap \tilde{\mathcal{B}} : s_0 \le s_n, \delta_n(s_0) \le \eta_n, \\ \phi_{\ell_{2,1}}(s_0 + M_2 s^\star) > c\},\, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0}} \mathbb{E}_0\Pi(\boldsymbol{\beta} : S_{\boldsymbol{\beta}} = S_0|Y_1, \ldots, Y_n) \to 1.$$

Under the conditions in Theorem 3.7, the marginal posterior distribution of $\boldsymbol{\beta}$ in nonzero groups can be further approximated by a multivariate normal distribution with mean $\mathrm{Vec}(\hat{\beta}^{\star}_{S_0})$ and the covariance matrix $\hat{\mathbb{I}}^{-1}_{S_0} = n\big(\sum_{i=1}^{n} \boldsymbol{X}_{i,S_0}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X}'_{i,S_0}\big)^{-1}$. Therefore, credible sets for $\boldsymbol{\beta}$ can be obtained directly from the approximating multivariate normal density. It may be noted that under the setting of the theorem, the lower bound in the Beta-min condition goes to zero, implying that the condition becomes milder with increasing sample size.

4. Computational algorithms

Various sampling-based computation algorithms have been developed to compute the posterior distribution in the sparse linear regression model with a spike-and-slab prior under the setting that the covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_d$ and sparsity is imposed on individual coefficients. A summary of those algorithms is provided in Section 5 of Castillo et al. (2015). Recently, Xu and Ghosh (2015) developed an MCMC algorithm using a spike-and-slab prior for group variable selection. They placed a beta-binomial prior on the dimension and a prior involves $\ell_{2,1}$ norm, similar to ours, on the regression coefficients.

Since priors used in this paper are new, we outline an MCMC algorithm to compute the posterior distribution. For each iteration of the algorithm, one can start with sampling S from the marginal posterior distribution with $\boldsymbol{\beta}$ integrated out. Next, conditioning on the current S, draw $\boldsymbol{\beta}$ from the corresponding conditional posterior distribution. Since the prior for $\boldsymbol{\beta}$ is not a conjugate prior, the Metropolis-Hasting algorithm can be used with the proposal density chosen as a multivariate normal distribution centered at its current value. Last, sample $\boldsymbol{\Sigma}$ through sampling $\boldsymbol{P}$ and $\boldsymbol{D}$, and then calculating $\boldsymbol{PDP'}$. To sample the diagonal elements of $\boldsymbol{D}$, one can convert them to log scale and then for each element, choose the proposal density as a normal distribution centered at its current value in log scale. To sample $\boldsymbol{P}$, one can draw a new value $\boldsymbol{P}^{\star}$ uniformly from the group of orthogonal matrices. Then the acceptance ratio equals to the likelihood ratio. When d is large, in order to increase acceptance rate of the Metropolis-Hasting algorithm, one can restrict the proposal density to local moves through multiplying by a random orthogonal matrix with some ϵ of the identity matrix. If the conjugate inverse Wishart prior is used instead, then the conditional posterior distribution of $\boldsymbol{\Sigma}$ is also an inverse Wishart distribution. One can sample $\boldsymbol{\Sigma}$ from that distribution directly.

5. Proofs

The lower bound for the denominator in the expression for the posterior probability obtained in the following result relies on sufficient prior concentration near the truth and is instrumental in establishing the posterior contraction rate. Let f stands for the joint density of $(Y_1, \dots, Y_n)$ under a generic value of the parameter $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and f_0 stand for that under the true value $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$.

Lemma 5.1. *For some constant $C_1 > 0$, $\mathcal{B}_0$ and $\mathcal{H}_0$ are defined in* (3)*, we have* $\sup\{\mathbb{P}_0(E_n^c) : \boldsymbol{\beta}_0 \in \mathcal{B}_0, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0\} \to 0$*, where the set* $E_n = \Big\{\int\int \frac{f}{f_0} d\Pi(\boldsymbol{\beta}) d\Pi(\boldsymbol{\Sigma}) \geq e^{-C_1 n \epsilon_n^2}\Big\}$.

Proof. In view of Lemma 8.10 of Ghosal and van der Vaart (2017), it suffices that

$$-\log \Pi\left\{(\boldsymbol{\beta}, \boldsymbol{\Sigma}) : K(f_0, f) \leq n\epsilon_n^2, V(f_0, f) \leq n\epsilon_n^2\right\} \lesssim n\epsilon_n^2, \tag{5.1}$$

where $K(f_0, f)$ and $V(f_0, f)$ respectively stand for the average Kullback-Leibler divergence and average Kullback-Leibler variation between f_0 and f given by

$$\frac{1}{n}K(f_0, f) = \frac{1}{2}\left(\operatorname{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0) - d - \log\det(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0) + \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i'\|^2\right),$$

$$\frac{1}{n}V(f_0, f) = \frac{1}{2}\left(\operatorname{Tr}\left((\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0)^2\right) - 2\operatorname{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0) + d\right) + \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i'\|^2.$$

Define a set of covariance matrices $\mathcal{A}_1$ and $\mathcal{A}_2$ a set of pairs of regression coefficients and covariance matrices by

$$\begin{aligned}\mathcal{A}_1 = \Big\{\boldsymbol{\Sigma} : {} & \operatorname{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0) - d - \log\det(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0) \leq \epsilon_n^2, \\ & \operatorname{Tr}\left((\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0)^2\right) - 2\operatorname{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0) + d \leq \epsilon_n^2\Big\}, \\ \mathcal{A}_2 = \Big\{(\boldsymbol{\beta}, \boldsymbol{\Sigma}) : {} & \sum_{i=1}^{n} \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i'\|^2 \leq n\epsilon_n^2, \ \sum_{i=1}^{n} \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i'\|^2 \leq n\epsilon_n^2/2\Big\}.\end{aligned}$$

Then a lower bound for the prior probability in (5.1) can be obtained by lower bounding $\Pi(\mathcal{A}_1)$ and $\Pi(\mathcal{A}_2|\mathcal{A}_1)$ separately and multiplying.

Writing $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1/2}$, $\mathcal{A}_1$ can be written as

$$\mathcal{A}_1 = \left\{\boldsymbol{\Sigma} : \sum_{k=1}^{d}\left(\operatorname{eig}_k(\boldsymbol{\Sigma}^{*-1}) - 1 - \log\operatorname{eig}_k(\boldsymbol{\Sigma}^{*-1})\right) \leq \epsilon_n^2, \ \sum_{k=1}^{d}\left(\operatorname{eig}_k(\boldsymbol{\Sigma}^{*-1}) - 1\right)^2 \leq \epsilon_n^2\right\}.$$

By Taylor's expansion $\log(x+1) = x - x^2/2 + o(1)$ as $x \to 0$ and since $\epsilon_n \to 0$, it follows that the second condition in $\mathcal{A}_1$ implies the first, and hence $\mathcal{A}_1 = \big\{\boldsymbol{\Sigma} : \sum_{k=1}^{d}(\operatorname{eig}_k(\boldsymbol{\Sigma}^{*-1}) - 1)^2 \leq \epsilon_n^2\big\}$ for sufficiently large n. Since the eigenvalues of $\boldsymbol{\Sigma}_0$ are between b_1 and b_2 by Assumption 3, Lemma A.1 of Banerjee and Ghosal (2015) gives that $\sum_{k=1}^{d}(\operatorname{eig}_k(\boldsymbol{\Sigma}^{*-1}) - 1)^2 \leq b_2^2\|\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}\|_F^2$, and hence $\mathcal{A}_1 \supset \big\{\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}\|_F \leq \epsilon_n/b_2\big\}$ for sufficiently large n. Writing in terms of the eigendecomposition $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'$, the triangle inequality, the norm-inequality $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \min\{\|\boldsymbol{A}\|\|\boldsymbol{B}\|_F, \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|\}$ and the facts that $\|\boldsymbol{P}\| = 1 = \|\boldsymbol{P}_0\|$ and $\|\boldsymbol{D}_0^{-1}\|$ is bounded, we have that

$$\|\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}\|_F \leq \|\boldsymbol{P}_0\|\|\boldsymbol{P}\|\|\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1}\|_F + (\|\boldsymbol{P}_0\|\|\boldsymbol{D}_0^{-1}\| + \|\boldsymbol{P}\|\|\boldsymbol{D}^{-1}\|)\|\boldsymbol{P} - \boldsymbol{P}_0\|_F$$

$$\lesssim \|\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1}\|_F + \|\boldsymbol{P} - \boldsymbol{P}_0\|_F + \|\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1}\|_F \|\boldsymbol{P} - \boldsymbol{P}_0\|_F,$$

since $\|\boldsymbol{D}^{-1}\| \leq \|\boldsymbol{D}_0^{-1}\| + \|\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1}\|$, and the spectral norm is always bounded by the Frobenius norm. Therefore, we have that

$$\mathcal{A}_1 \supset \left\{\boldsymbol{\Sigma} : \|\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1}\|_F \leq \epsilon_n/c_1,\ \|\boldsymbol{P} - \boldsymbol{P}_0\|_F \leq \epsilon_n/c_1\right\},$$

for some $c_1 > 0$. Using the prior independence of the eigenvalue distribution and positive lower bound for the prior density at all concerned true value $\boldsymbol{\Sigma}_0$, it is easy to see that $\log \Pi \left\{\boldsymbol{\Sigma} : \|\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1}\|_F \leq \epsilon_n/c_1\right\} \gtrsim -d \log(1/\epsilon_n) \gtrsim -d \log n$. To lower bound $\Pi(\boldsymbol{P} : \|\boldsymbol{P} - \boldsymbol{P}_0\|_F \leq \epsilon_n/c_1)$, note that Π is the Haar measure on a compact Lie group of dimension $d(d-1)/2$. This means that all translates of $\{\boldsymbol{P} : \|\boldsymbol{P} - \boldsymbol{P}_0\|_F \leq \epsilon_n/c_1\}$ have the same probability, and N many such translates can cover the entire set of $d \times d$ orthogonal matrices, where N stands for the ϵ_n/c_1-covering number of the set of $d \times d$ orthogonal matrices in terms of the Frobenius distance. A crude upper bound for N is easily obtained by embedding the set of $d \times d$ orthogonal matrices in $[-1, 1]^{d^2}$, giving the estimate $N \leq (2c_1/\epsilon_n)^{d^2}$. This leads to the estimate $\log \Pi \{\boldsymbol{\Sigma} : \|\boldsymbol{P} - \boldsymbol{P}_0\|_F \leq \epsilon_n/c_1\} \gtrsim -d^2 \log(2c_1/\epsilon_n) \gtrsim -d^2 \log n$. Thus $\log \Pi(\mathcal{A}_1) \gtrsim -d^2 \log n$ using the prior independence of $\boldsymbol{D}$ and $\boldsymbol{P}$.

To derive a lower bound for $\Pi(\mathcal{A}_2|\mathcal{A}_1)$, we first note that $\|\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}\|_F \lesssim \epsilon_n$ implies that $\|\boldsymbol{\Sigma}^{-1}\|$ and $\|\boldsymbol{\Sigma}^{*-1}\|$ are bounded by a fixed constant, and hence $n^{-1}\sum_{i=1}^n X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i'$ and $n^{-1}\|\boldsymbol{\Sigma}^{*-1}\|\|\boldsymbol{\Sigma}_0^{-1}\| \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2$ are both bounded by a constant multiple of $n^{-1}\sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 = n^{-1}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_F^2$. Now by the inequality

$$\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_F \leq \|\boldsymbol{X}\|_\circ \sum_{j=1}^G \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{0j}\|_F \leq \|\boldsymbol{X}\|_\circ \sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1}, \tag{5.2}$$

to bound $\Pi(\mathcal{A}_2|\mathcal{A}_1)$ from below, it suffices to bound $\Pi\left(\sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1} \leq c r_n\right)$, where $r_n = \sqrt{n\epsilon_n^2}/\|\boldsymbol{X}\|_\circ$ and c is a positive constant. By (2.1), this can be further bounded below by

$$\pi(s_0)\frac{1}{\binom{Gd}{s_0}} \int_{\sum_{k=1}^d \|\beta_{S_{0,k}} - \beta_{0,S_{0,k}}\|_{2,1} \leq c r_n} \prod_{k=1}^d g_{s_{0,k}}(\beta_{S_{0,k}}) d\beta_{S_{0,1}} \dots d\beta_{S_{0,d}}. \tag{5.3}$$

By changing the variable $\beta_{S_{0,k}} - \beta_{0,S_{0,k}}$ to $\check{\beta}_{S_{0,k}}$ and using the fact that $\|x\| \leq \|x\|_1$ for any vector x, the integral in (5.3) is bounded below by

$$\begin{aligned}
&e^{-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1}} \int_{\sum_{k=1}^d \|\check{\beta}_{S_{0,k}}\|_1 \leq c r_n} \prod_{k=1}^d g_{s_{0,k}}(\check{\beta}_{S_{0,k}}) d\check{\beta}_{S_{0,1}} \dots d\check{\beta}_{S_{0,d}} \\
&\quad \geq e^{-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1}} \prod_{k=1}^d \prod_{j \in S_{0,k}} \left(\frac{2\lambda_k}{a_j \bar{\bar{\lambda}}}\right)^{p_j}
\end{aligned}$$

$$\times \int_{\sum_{k=1}^d \|\check{\beta}_{S_{0,k}}\|_1 \leq cr_n} \left(\frac{\overline{\lambda}}{2}\right)^{p_{S_0}} e^{-\overline{\lambda}\sum_{k=1}^d \|\check{\beta}_{S_{0,k}}\|_1} d\check{\beta}_{S_{0,1}} \dots d\check{\beta}_{S_{0,d}}.$$

Using the result that the integrand equals to the probability of the first p_{S_0} events of a Poisson process happen before time cr_n (similar to the argument used to derive (6.2) in Castillo et al., 2015), the last display is further bounded below by

$$e^{-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1}} \left\{ \prod_{k=1}^d \prod_{j\in S_{0,k}} \left(\frac{2\lambda_k}{a_j\overline{\lambda}}\right)^{p_j} \right\} e^{-\overline{\lambda}cr_n} \frac{1}{p_{S_0}!} \left(\overline{\lambda}cr_n\right)^{p_{S_0}}$$
$$\geq e^{-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1} - \overline{\lambda}cr_n} \left\{ \prod_{k=1}^d \prod_{j\in S_{0,k}} \left(\frac{2}{a_j}\right)^{p_j} \right\} \frac{1}{p_{S_0}!} (\underline{\lambda}cr_n)^{p_{S_0}}.$$

Hence, by Assumption 1, (5.3) is bounded below by

$$\frac{\pi(0)A_1^{s_0}}{(G\vee n^{p_{\max}})^{A_3 s_0}(Gd)^{s_0}} e^{-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1} - \overline{\lambda}cr_n} \frac{(\underline{\lambda}cr_n)^{p_{S_0}}}{p_{S_0}!} \prod_{k=1}^d \prod_{j\in S_{0,k}} \left(\frac{2}{a_j}\right)^{p_j},$$

implying that $\log \Pi(K(f_0, f) \leq n\epsilon_n^2, V(f_0, f) \leq n\epsilon_n^2)$ is bounded below by

$$\begin{aligned} &- d^2 \log n + \log \pi(0) + s_0 \log A_1 - c_{14}s_0(\log G + p_{\max} \log n + \log d) - \sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1} \\ &\quad - \overline{\lambda}cr_n + p_{S_0}\log(\underline{\lambda}cr_n) - \log(p_{S_0}!) - \sum_{k=1}^d \sum_{j\in S_{0,k}} p_j \log(a_j/2), \end{aligned} \tag{5.4}$$

for some constant $c_{14} > 0$. As $\pi(0)$ is bounded away from zero, and Assumption 2 gives $\overline{\lambda}cr_n - p_{S_0}\log(\underline{\lambda}cr_n) \lesssim \sqrt{n}\epsilon_n\sqrt{\log G} + (p_{S_0}/p_{\max})\log G \lesssim n\epsilon_n^2$, the second, sixth, and seventh terms are controlled.

Also, since $\sum_{k=1}^d \|\beta_{0,k}\|_{2,1} \leq \overline{\beta}$ with the expression of $\overline{\beta}$ is displayed in (3.1), we have $\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1} \leq \max_{1\leq k\leq d}\lambda_k \sum_{k=1}^d \|\beta_{0,k}\|_{2,1} \leq n\epsilon_n^2$. Furthermore, since $\log(p_{S_0}!) \leq p_{S_0}\log p_{S_0}$ and $a_j = O(p_j^{1/2})$, we obtain that $\log(p_{S_0}!) + \sum_{k=1}^d \sum_{j\in S_{0,k}} p_j \log(a_j/2) \lesssim s_0 p_{\max}\log n \leq n\epsilon_n^2$. Thus (5.4) is bounded below by a constant multiple of $-n\epsilon_n^2$. □

Proof of Lemma 3.2. Let $\mathcal{B}_n = \{\boldsymbol{\beta} : s_{\boldsymbol{\beta}} < r\}$. We show that $\mathbb{E}_0\Pi(\boldsymbol{\beta} \in \mathcal{B}_n^c | Y_1, \dots, Y_n) \to 0$ as $n \to \infty$ for $r \geq s_0$. By Lemma 5.1, the denominator of (3.9) in the expression for $\Pi(\boldsymbol{\beta} \in \mathcal{B}_n^c | Y_1, \dots, Y_n)$ with $\mathcal{B}_n$ as above, is bounded below by $e^{-C_1 n\epsilon_n^2}$ with a large probability. To derive an upper bound for the corresponding numerator, note that its expected value is

$$\mathbb{E}_0\Big(\int\int_{\mathcal{B}_n^c} (f/f_0) d\Pi(\boldsymbol{\beta}) d\Pi(\boldsymbol{\Sigma})\Big) \leq \int_{\mathcal{B}_n^c} d\Pi(\boldsymbol{\beta}) = \Pi(s_{\boldsymbol{\beta}} \geq r) = \sum_{s=r}^{\infty} \pi(s),$$

and by Assumption 1 and $A_2/(G \vee n^{p_{\max}})^{A_4} \leq 1/2$ as $n \to \infty$, the bound simplifies to

$$\pi(s_0)\Big(\frac{A_2}{(G \vee n^{p_{\max}})^{A_4}}\Big)^{r-s_0} \sum_{j=0}^{\infty} \Big(\frac{A_2}{(G \vee n^{p_{\max}})^{A_4}}\Big)^{j} \leq 2\Big(\frac{A_2}{(G \vee n^{p_{\max}})^{A_4}}\Big)^{r-s_0}.$$

Therefore, because $\mathbb{E}_0\Pi(\mathcal{B}_n^c|Y_1,\dots,Y_n) \leq \mathbb{E}_0\Pi(\mathcal{B}_n^c|Y_1,\dots,Y_n)\mathbb{1}_{E_n} + \mathbb{P}_0(E_n^c)$ and $\mathbb{P}_0(E_n^c) \to 0$, choosing $r = M_2\{s_0 \vee [d^2 \log n/(\log G \vee p_{\max} \log n)]\}$ for some M_2 large enough, we obtain that $\mathbb{E}_0\Pi(\mathcal{B}_n^c|Y_1,\dots,Y_n)$ is bounded above by

$$\exp\Big(C_1 n\epsilon_n^2 + \log 2 + (r - s_0)(\log A_2 - A_4(\log G \vee p_{\max} \log n))\Big) + o(1) \to 0.$$

□

Proof of Theorem 3.1. The proof contains two parts. In the first part, we obtain the posterior contraction rate with respect to the average negative log-affinity. In the second part, we use the results obtained from the first part to derive (3.2) and (3.3).

Part I. Note that for every $\epsilon > 0$,

$$\begin{aligned}&\mathbb{E}_0\Pi\left((\boldsymbol{\beta},\boldsymbol{\Sigma}) \in \mathbb{R}^{p\times d} \times \mathcal{H} : \frac{1}{n}\sum_{i=1}^{n} \rho(f_i, f_{0,i}) > \epsilon|Y_1,\dots,Y_n\right)\\ &\quad\leq \mathbb{E}_0\Pi\left((\boldsymbol{\beta},\boldsymbol{\Sigma}) \in \mathcal{B}_n \times \mathcal{H} : \frac{1}{n}\sum_{i=1}^{n} \rho(f_i, f_{0,i}) > \epsilon|Y_1,\dots,Y_n\right) + \mathbb{E}_0\Pi(\mathcal{B}_n^c|Y_1,\dots,Y_n),\end{aligned}$$

where $\mathcal{H}$ is the space of $d \times d$ positive definite matrices and $\mathcal{B}_n = \{\boldsymbol{\beta} : s_{\boldsymbol{\beta}} < M_2 s^\star\}$. The second term on the right hand side goes to zero by Lemma 3.2, and hence it suffices to show that the first term goes to zero for $\epsilon^2 = M_1\epsilon_n^2$.

Define the sieve

$$\mathcal{F}_n = \left\{(\boldsymbol{\beta},\boldsymbol{\Sigma}) \in \mathcal{B}_n \times \mathcal{H} : \max_{\substack{1\leq j\leq G\\ 1\leq k\leq d}} \|\beta_{jk}\| \leq H_n,\ n^{-1} < \operatorname{eig}_1(\boldsymbol{\Sigma}^{-1}),\ \operatorname{eig}_d(\boldsymbol{\Sigma}^{-1}) \leq n\right\},$$

where $H_n = p_{\max} n/\underline{\lambda}$ for $\underline{\lambda}$ given in (2.3). Then

$$\begin{aligned}\Pi((\mathcal{B}_n \times \mathcal{H}) \setminus \mathcal{F}_n) \leq &\sum_{S:s\leq M_2 s^\star} \frac{\pi(s)}{\binom{Gd}{s}} \sum_{k=1}^{d}\sum_{j\in S_k} \Pi(\|\beta_{jk}\| \geq H_n)\\ &+ \Pi\left(\operatorname{eig}_1(\boldsymbol{\Sigma}^{-1}) \leq n^{-1}\right) + \Pi\left(\operatorname{eig}_d(\boldsymbol{\Sigma}^{-1}) \geq n\right).\end{aligned} \tag{5.5}$$

It is easy to see that $\|\beta_{jk}\|$ is gamma distributed with shape parameter p_j and scale parameter λ_k. Applying the estimate of the tail of a gamma density on page 29 of

Boucheron et al. (2013) and the inequality $1 + x - \sqrt{1+2x} \geq (x-1)/2$, for any $x > 0$, we have that

$$\Pi(\|\beta_{jk}\| > H_n) \leq \exp\left(-p_j\left(1 + \frac{\lambda_k H_n}{p_j} - \sqrt{1 + 2\frac{\lambda_k H_n}{p_j}}\right)\right) \leq \exp\left(-\underline{\lambda} H_n + p_{\max}\right),$$

for $j = 1, \dots, G$, $k = 1, \dots, d$, leading to the estimate

$$\sum_{s=1}^{M_2 s^\star} \pi(s) s \exp\left(-\underline{\lambda} H_n + p_{\max}\right) \leq \exp\Big(\log(M_2 s^\star) - p_{\max}(n-1)\Big).$$

The second and third terms in (5.5) are both bounded by $e^{-c_2 n}$ for some $c_2 > 0$ by the tail property of inverse Gaussian distribution. Combining all these estimates, we obtain that for all sufficiently large n,

$$\Pi((\mathcal{B}_n \times \mathcal{H}) \setminus \mathcal{F}_n) \leq \exp\left(-(1 + C_1) n \epsilon_n^2\right).$$

Next, we construct a test φ_n such that

$$\mathbb{E}_{f_0} \varphi_n \lesssim e^{-M_1 n \epsilon_n^2/2}, \qquad \sup_{f \in \mathcal{F}_n : \rho(f_0, f) > M_1 n \epsilon_n^2} \mathbb{E}_f (1 - \varphi_n) \lesssim e^{-M_1 n \epsilon_n^2}, \tag{5.6}$$

for some $M_1 > C_1 + 1$, where $f_0 = \prod_{i=1}^n f_{0,i}$, $f_{0,i} = \mathcal{N}(X_i \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ and $f = \prod_{i=1}^n f_i$, $f_i = \mathcal{N}(X_i \boldsymbol{\beta}, \boldsymbol{\Sigma})$, $i = 1, \dots, n$, as required for an application of the general theory of posterior contraction. To this end, we first consider testing $H_0 : f = f_0$ against a single point $f = f_1$ in the alternative. Consider the most powerful Neyman-Pearson test $\phi_n = \mathbb{1}\{f_1/f_0 \geq 1\}$. If the average Réyni divergence $-n^{-1} \log \int f_0^{1/2} f_1^{1/2}$ between f_0 and f_1 is bigger than $\epsilon^2 > 0$, then

$$\mathbb{E}_{f_0} \phi_n = \mathbb{E}_{f_0}\left(\sqrt{f_1/f_0} \geq 1\right) \leq \int \sqrt{f_0 f_1} \leq e^{-n\epsilon^2},$$
$$\mathbb{E}_{f_1}(1 - \phi_n) = \mathbb{E}_{f_1}\left(\sqrt{f_0/f_1} \geq 1\right) \leq \int \sqrt{f_0 f_1} \leq e^{-n\epsilon^2}.$$

The test ϕ_n can also have exponentially small probability of type II error at other alternatives, because by the Cauchy-Schwarz inequality,

$$\mathbb{E}_f (1 - \phi_n) \leq \{\mathbb{E}_{f_1}(1 - \phi_n)\}^{1/2} \left\{\mathbb{E}_{f_1} (f/f_1)^2\right\}^{1/2}. \tag{5.7}$$

so that the expression can be controlled properly if the second factor grows at most like $e^{cn\epsilon^2}$ where $c > 0$ can be chosen suitably small. Now we show that $\mathbb{E}_{f_1}(f/f_1)^2$ is bounded for every density with parameters such that

$$\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}\|_\infty \leq \frac{1}{s^\star \sqrt{p_{\max} n} \|\boldsymbol{X}\|_\circ}, \quad \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\| \leq \frac{1}{n^2 d}, \quad \|\boldsymbol{\Sigma}^{-1}\| \leq n. \tag{5.8}$$

To see this, we observe that for $\boldsymbol{\Sigma}_1^\star = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1/2}$,

$$\begin{aligned}\mathbb{E}_{f_1}(f/f_1)^2 &= (\det(\boldsymbol{\Sigma}_1^\star))^{n/2}\Big(\det(2\boldsymbol{I} - \boldsymbol{\Sigma}_1^{\star\,-1})\Big)^{-n/2} \\ &\qquad \times \exp\Big(\sum_{i=1}^n X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_1)\boldsymbol{\Sigma}^{-1/2}(2\boldsymbol{\Sigma}_1^\star - \boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_1)'X_i'\Big).\end{aligned} \tag{5.9}$$

Because $\boldsymbol{\Sigma} \in \mathcal{F}_n$, the condition $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\| \le \delta_n' = 1/(n^2 d)$ implies that

$$\|\boldsymbol{\Sigma}_1^\star - \boldsymbol{I}\| \le \|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\| \le n\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\| \le n\delta_n',$$

and hence $1 - n\delta_n' \le \operatorname{eig}_1(\boldsymbol{\Sigma}_1^\star) \le \operatorname{eig}_d(\boldsymbol{\Sigma}_1^\star) \le 1 + n\delta_n'$. Therefore, we obtain that

$$\begin{aligned}\left(\frac{\det(\boldsymbol{\Sigma}_1^\star)}{\det(2\boldsymbol{I} - \boldsymbol{\Sigma}_1^{\star\,-1})}\right)^{n/2} &= \exp\left(\frac{n}{2}\sum_{k=1}^d \log\left(\operatorname{eig}_k(\boldsymbol{\Sigma}_1^\star)\right) - \frac{n}{2}\sum_{k=1}^d \log\left(2 - \frac{1}{\operatorname{eig}_k(\boldsymbol{\Sigma}_1^\star)}\right)\right) \\ &\le \exp\left(\frac{dn}{2}\log(1 + n\delta_n') - \frac{dn}{2}\log\left(1 - \frac{n\delta_n'}{1 - n\delta_n'}\right)\right).\end{aligned}$$

By the inequalities $1 - x^{-1} \le \log x \le x - 1$ for $x > 0$, the display is further bounded by

$$\exp\left(\frac{n^2 d\delta_n'}{2} + \frac{dn}{2}\left(\frac{n\delta_n'}{1 - 2n\delta_n'}\right)\right) \le \exp\left(n^2 d\delta_n'\right) = e.$$

By the inequality (5.2), we bound the exponential term in (5.9) by

$$\begin{aligned}&\|\boldsymbol{\Sigma}^{-1}\|\|(2\boldsymbol{\Sigma}_1^\star - \boldsymbol{I})^{-1}\|\sum_{i=1}^n \|X_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta})\|_2^2 \\ &\quad \le \|\boldsymbol{\Sigma}^{-1}\|\ \|(2\boldsymbol{\Sigma}_1^\star - \boldsymbol{I})^{-1}\|\|\boldsymbol{X}\|_\circ^2\Big(\sum_{k=1}^d \|\beta_{1,k} - \beta_k\|_{2,1}\Big)^2.\end{aligned}$$

Since $\|(2\boldsymbol{\Sigma}_1^\star - \boldsymbol{I})^{-1}\| \le 2$, $\|\boldsymbol{\Sigma}^{-1}\| \le n$, and $\sum_{k=1}^d \|\beta_{1,k} - \beta_k\|_{2,1} \le s_{\boldsymbol{\beta}_1 - \boldsymbol{\beta}}\sqrt{p_{\max}}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}\|_\infty \le 2M_2 s^\star\sqrt{p_{\max}}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}\|_\infty$ on $\mathcal{F}_n$, the display is further bounded by

$$8M_2^2 n s^{\star 2} p_{\max}\|\boldsymbol{X}\|_\circ^2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}\|_\infty^2 \le 8M_2^2.$$

Hence we conclude that (5.7) is bounded by a multiple of $e^{-n\epsilon^2}$ for every density with a parameter in the piece.

The desired test φ_n satisfying (5.6) is obtained as the maximum of all tests ϕ_n described above, for each piece required to cover the sieve. To complete the proof of (5.6), we need to show that $\log N_* \lesssim n\epsilon_n^2$, where N_* is the number of pieces satisfying (5.8) needed to cover the sieve $\mathcal{F}_n$ (see Lemma D.3 of Ghosal and van der Vaart (2017)). It is easy to see that $\log N_*$ is bounded by

$$\log N\Big(\frac{1}{s^\star\sqrt{p_{\max}n}\|\boldsymbol{X}\|_\circ}, \big\{\boldsymbol{\beta} : s_{\boldsymbol{\beta}} \le M_2 s^\star, \max_{\substack{1\le j\le G \\ 1\le k\le d}} \|\beta_{jk}\| < H_n\big\}, \|\cdot\|_\infty\Big)$$

$$+\log N\Big(\frac{1}{n^2 d}, \left\{\boldsymbol{\Sigma} : n^{-1} < \mathrm{eig}_1(\boldsymbol{\Sigma}^{-1}),\ \mathrm{eig}_d(\boldsymbol{\Sigma}^{-1}) < n\right\}, \|\cdot\|\Big).$$

The first term of the display is bounded by

$$\begin{aligned}
&\log N\Big(\frac{1}{s^\star\sqrt{p_{\max}n}\|\boldsymbol{X}\|_\circ}, \{\boldsymbol{\beta} : s_{\boldsymbol{\beta}} \le M_2 s^\star, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty < H_n\}, \|\cdot\|_\infty\Big)\\
&\le \log\left\{\binom{Gd}{M_2 s^\star}\Big(3\sqrt{p_{\max}n}s^\star H_n\|\boldsymbol{X}\|_\circ\Big)^{M_2 s_n^\star p_{\max}}\right\}\\
&\lesssim s^\star \log G + s^\star p_{\max}(\log n + \log(H_n\|\boldsymbol{X}\|_\circ) \qquad (5.10)
\end{aligned}$$

while the second term is bounded by

$$\begin{aligned}
\log N\Big(\frac{1}{n^2 d}, \left\{\boldsymbol{\Sigma} : n^{-1} < \mathrm{eig}_1(\boldsymbol{\Sigma}^{-1})\right\}, \|\cdot\|\Big) &\le \log N\Big(\frac{1}{n^2 d}, \Big\{\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma}\|_F < n\sqrt{d}\Big\}, \|\cdot\|_F\Big)\\
&\le d^2 \log\left(n^3 d^{3/2}\right),
\end{aligned}$$

both of which are bounded by a constant multiple of $n\epsilon_n^2$.

Choosing $\epsilon = M_1\epsilon_n^2$ for a sufficiently large $M_1 > 1 + C_1$, we thus have (5.6). We finally obtain that the posterior $\Pi\left(\sum_{i=1}^n \rho(f_i, f_{0,i}) > M_1 n\epsilon_n^2 | Y_1, \dots, Y_n\right)$ goes to zero in $\mathbb{P}_0$-probability.

Part II. Observe that $n^{-1}\sum_{i=1}^n \rho(f_i, f_{0,i})$ is equal to

$$-\log\left(\frac{(\det(\boldsymbol{\Sigma}))^{1/4}(\det(\boldsymbol{\Sigma}_0))^{1/4}}{(\det((\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0)/2))^{1/2}}\right) + \frac{1}{8n}\sum_{i=1}^n X_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\left(\frac{\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0}{2}\right)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)' X_i'.$$

Then $\sum_{i=1}^n \rho(f_i, f_{0,i}) \lesssim n\epsilon_n^2$ implies the relations

$$-\log\left(\frac{(\det(\boldsymbol{\Sigma}))^{1/4}(\det(\boldsymbol{\Sigma}_0))^{1/4}}{(\det((\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0)/2))^{1/2}}\right) \lesssim \epsilon_n^2, \qquad (5.11)$$

$$\frac{1}{8n}\sum_{i=1}^n X_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\left((\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0)/2\right)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)' X_i' \lesssim \epsilon_n^2. \qquad (5.12)$$

First, we show that the probability of (5.11) goes to 1 implies (3.3). Let

$$d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0) = h^2\left(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)\right) = 1 - \frac{(\det(\boldsymbol{\Sigma}))^{1/4}(\det(\boldsymbol{\Sigma}_0))^{1/4}}{(\det((\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0)/2))^{1/2}}.$$

Since the eigenvalues of $\boldsymbol{\Sigma}_0$ lie in $[b_1, b_2]$, by Lemma 2 of Suarez and Ghosal (2017), we obtain that $d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0) \gtrsim \|\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}_0^{-1/2}\|_F^2$, if the left hand side is sufficiently small. Since

$$-\log\left(\frac{(\det(\boldsymbol{\Sigma}))^{1/4}(\det(\boldsymbol{\Sigma}_0))^{1/4}}{(\det((\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0)/2))^{1/2}}\right) = -\log(1 - d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)) \ge d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0),$$

we obtain that $\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0\|_F^2 \lesssim \epsilon_n^2$. This proves (3.3).

Next, we show that the probability (5.12) goes to 1 implies (3.2). Given (3.3) and by Assumption 3, we obtain that

$$\|\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0\|^2 = \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0+2\boldsymbol{\Sigma}_0\|^2 \leq 2\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0\|_F^2 + 8\|\boldsymbol{\Sigma}_0\|^2 \lesssim \epsilon_n^2+1.$$

Hence using $\mathrm{eig}_1\left((\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0/2)^{-1}\right) = \left(\mathrm{eig}_d(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0/2)\right)^{-1} = \|(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0/2)\|^{-1} \geq (1+\epsilon_n^2)^{-1/2}$, (5.12) implies that

$$\epsilon_n^2 \geq \frac{1}{8n}\sum_{i=1}^n \|X_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|^2 \left\|\frac{\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0}{2}\right\|^{-1} \gtrsim \frac{1}{n}\sum_{i=1}^n \|X_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|^2/\sqrt{\epsilon_n^2+1}.$$

Combining with (3.3), we obtain (3.2). □

Proof of Theorem 3.6. Let $\mathcal{H}_n = \{\boldsymbol{\Sigma}\in\mathcal{H} : \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0\|_F^2 \leq M_1\epsilon_n^2\}$ and

$$\Theta_n = \left\{\boldsymbol{\beta}\in\mathbb{R}^{p\times d} : s_{\boldsymbol{\beta}} \leq M_2 s^\star, \left(\sum_{k=1}^d \|\beta_k-\beta_{0,k}\|_{2,1}\right)^2 \leq \frac{M_4 n\epsilon_n^2 s^\star}{\|\boldsymbol{X}\|_\circ^2 \phi_{\ell_{2,1}}^2(s_0+M_2 s^\star)}\right\},$$

where $\mathcal{H}$ is a space of $d\times d$ positive definite matrices. The proof contains two parts. In the first part, we show that the total variation metric between $\Pi(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)$ and $\check{\Pi}_n(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n) := \check{\Pi}_n((\boldsymbol{\beta},\boldsymbol{\Sigma})\in\cdot\times\mathcal{H}_n|Y_1,\dots,Y_n)$ is small, where $\check{\Pi}_n((\boldsymbol{\beta},\boldsymbol{\Sigma})\in\cdot\times\cdot|Y_1,\dots,Y_n)$ is the renormalized measure of $\Pi((\boldsymbol{\beta},\boldsymbol{\Sigma})\in\cdot\times\cdot|Y_1,\dots,Y_n)$ restricted to the set $\Theta_n\times\mathcal{H}_n$. We also show that the total variation distance between $\Pi^\infty(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)$ and $\check{\Pi}_n^\infty(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)$ is small, where $\check{\Pi}_n^\infty(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)$ is the measure $\Pi^\infty(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)$ restricted and renormalized to Θ_n. In the second part, we show that the total variation distance between $\check{\Pi}_n(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)$ and $\check{\Pi}_n^\infty(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)$ is small.

For any set A, let $\Pi_A(\cdot)$ be the renormalized measure of $\Pi(\cdot)$ which is restricted to the set A. Then $\|\Pi(\cdot)-\Pi_A(\cdot)\| \leq 2\Pi(A^c)$. Clearly,

$$\mathbb{E}_0\|\Pi(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)-\check{\Pi}_n(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)\|_{TV}\to 0,$$

by (3.3) and (3.7). To show that

$$\mathbb{E}_0\|\Pi^\infty(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)-\check{\Pi}_n^\infty(\boldsymbol{\beta}\in\cdot|Y_1,\dots,Y_n)\|_{TV}\to 0,$$

we write

$$\Pi^\infty(\boldsymbol{\beta}\in\Theta_n^c|Y_1,\dots,Y_n) = \frac{\int_{\Theta_n^c}\exp\{\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}_0)-\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0)\}dU(\boldsymbol{\beta})}{\int\exp\{\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}_0)-\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0)\}dU(\boldsymbol{\beta})}, \tag{5.13}$$

with $dU(\boldsymbol{\beta})$ defined in (3.10). By (3.8), $\ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}_0)-\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0)$ equals to

$$-\frac{1}{2}\sum_{i=1}^n \|\boldsymbol{\Sigma}_0^{-1/2}\tilde{\boldsymbol{X}}_i'\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\|^2 + \sum_{i=1}^n \left(Y_i-\mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\right)\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}_i'\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'.$$

By plugging-in the last display into (5.13), the denominator is bounded below by

$$
\begin{aligned}
&\frac{\pi(s_0)}{\binom{Gd}{s_0}}\left(\prod_{k=1}^{d}\prod_{j\in S_{0,k}}\left(\frac{\lambda_k}{a_j}\right)^{p_j}\right)\\
&\quad\times\int \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{\Sigma}_0^{-1/2}\tilde{\boldsymbol{X}}'_{i,S_0}\tilde{\beta}_{S_0}\|^2+\sum_{i=1}^{n}\left(Y_i-\mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\right)\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}'_{i,S_0}\tilde{\beta}_{S_0}\right)d\tilde{\beta}_{S_0},
\end{aligned}
$$

where $\tilde{\beta}_{S_0}=\left((\beta_{1,S_{0,1}}-\beta_{0,1,S_{0,1}})',\dots,(\beta_{d,S_{0,d}}-\beta_{0,d,S_{0,d}})'\right)'$. By Jensen's inequality, the display is bounded below by

$$
\begin{aligned}
&\frac{\pi(s_0)}{\binom{Gd}{s_0}}\left(\prod_{k=1}^{d}\prod_{j\in S_{0,k}}\left(\frac{\lambda_k}{a_j}\right)^{p_j}\right)\int \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{\Sigma}_0^{-1/2}\tilde{\boldsymbol{X}}'_{i,S_0}\tilde{\beta}_{S_0}\|^2\right)d\tilde{\beta}_{S_0},\\
&\quad=\frac{\pi(s_0)}{\binom{Gd}{s_0}}\left(\prod_{k=1}^{d}\prod_{j\in S_{0,k}}\left(\frac{\lambda_k}{a_j}\right)^{p_j}\right)\sqrt{\frac{(2\pi)^{p_{S_0}}}{\det\left(\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_{i,S_0}\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}'_{i,S_0}\right)}}. \qquad (5.14)
\end{aligned}
$$

Letting $\boldsymbol{\Gamma}_{S_0}=\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_{i,S_0}\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}'_{i,S_0}$, we apply Jensen's inequality to obtain that

$$
\det(\boldsymbol{\Gamma}_{S_0})\le(\mathrm{Tr}(\boldsymbol{\Gamma}_{S_0})/p_{S_0})^{p_{S_0}}\le\left(\max_{l}(\boldsymbol{\Gamma}_{S_0})_{l,l}\right)^{p_{S_0}},
$$

where $(\boldsymbol{\Gamma}_{S_0})_{l,l}$ is the lth diagonal element of $\boldsymbol{\Gamma}_{S_0}$. Note that

$$
\max_{l}(\boldsymbol{\Gamma}_{S_0})_{l,l}\le\frac{1}{b_1}\max_{1\le j\le G}\left\|\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_{i,j}\tilde{\boldsymbol{X}}'_{i,j}\right\|=\frac{1}{b_1}\max_{1\le j\le G}\|\boldsymbol{X}_j\|^2=\frac{\|\boldsymbol{X}\|_\circ^2}{b_1},
$$

where $\tilde{\boldsymbol{X}}_{i,j}=\boldsymbol{I}_d\otimes X'_{ij}$, and hence (5.14) is further bounded below by

$$
\begin{aligned}
&\frac{\pi(s_0)}{\binom{Gd}{s_0}}\left(\prod_{k=1}^{d}\prod_{j\in S_{0,k}}\left(\frac{\lambda_k}{a_j}\right)^{p_j}\right)\left(\frac{2b_1\pi}{\|\boldsymbol{X}\|_\circ^2}\right)^{p_{S_0}/2}\\
&\quad\ge\frac{\pi(s_0)}{(Gd)^{s_0}\prod_{k=1}^{d}\prod_{j\in S_{0,k}}a_j^{p_j}}\left(\frac{\sqrt{2b_1\pi}}{B_1(G^{1/p_{\max}}\vee n)^{B_2}}\right)^{p_{S_0}}\\
&\quad\ge\frac{\pi(s_0)}{(Gd)^{s_0}a_j^{s_0p_{\max}}}\left(\frac{\sqrt{2b_1\pi}}{B_1(G^{1/p_{\max}}\vee n)^{B_2}}\right)^{s_0p_{\max}}. \qquad (5.15)
\end{aligned}
$$

We thus obtain a lower bound for the denominator.

The numerator of (5.13) can be written as

$$
\begin{aligned}
&\int_{\Theta_n^c}\left\{\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\|\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\boldsymbol{\Sigma}_0^{-1/2}\|^2\right)\right.\\
&\qquad\left.\times\exp\left(\sum_{i=1}^{n}\left(Y_i-\mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\right)\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}'_i\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\right)\right\}dU(\boldsymbol{\beta}). \qquad (5.16)
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\sum_{i=1}^{n}\Big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\Big)\,\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}_i'\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)' \\
&\quad = \sum_{j=1}^{G}\sum_{i=1}^{n}\Big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\Big)\,\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}_{i,j}'\mathrm{Vec}(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{0,j})' \\
&\quad \le \sum_{j=1}^{G}\left\|\sum_{i=1}^{n}\Big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\Big)\,\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}_{i,j}'\right\|\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_{0,j}\|_F. \qquad (5.17)
\end{aligned}
$$

Using the tail inequality for quadratic forms of Gaussian random variables (Proposition 1 of Hsu et al. (2012)), we obtain for every $t > 0$,

$$
\begin{aligned}
\mathbb{P}\Bigg(&\max_{1\le j\le G}\left\|\sum_{i=1}^{n}\Big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\Big)\,\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}_{i,j}'\right\|^2 \\
&\qquad \ge \mathrm{Tr}(\boldsymbol{\Delta}'\boldsymbol{\Delta}) + 2\sqrt{\mathrm{Tr}((\boldsymbol{\Delta}'\boldsymbol{\Delta})^2)t} + 2\|\boldsymbol{\Delta}\|^2 t\Bigg) \le Ge^{-t},
\end{aligned}
$$

where $\boldsymbol{\Delta} = (\tilde{\boldsymbol{X}}_{1,j}\boldsymbol{\Sigma}_0^{-1}, \dots, \tilde{\boldsymbol{X}}_{n,j}\boldsymbol{\Sigma}_0^{-1}) \in \mathbb{R}^{p_j\times dn}$. Since $\mathrm{Tr}(\boldsymbol{\Delta}'\boldsymbol{\Delta}) \le p_j\|\boldsymbol{\Delta}\|^2$ and $\|\boldsymbol{\Delta}\| \lesssim \|(\tilde{\boldsymbol{X}}_{1,j}, \dots, \tilde{\boldsymbol{X}}_{n,j})\| = \|\boldsymbol{X}_j\| \le \|\boldsymbol{X}\|_\circ$, choosing $t = 2(\log G \vee p_{\max}\log n)$, we obtain

$$
\mathbb{P}\left(\max_{1\le j\le G}\left\|\sum_{i=1}^{n}\Big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\Big)\,\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}_{i,j}'\right\| \ge c_4\|\boldsymbol{X}\|_\circ\sqrt{\log G \vee p_{\max}\log n}\right) \le \frac{1}{G},
$$

for some $c_4 > 0$. Let $D_n = c_4\|\boldsymbol{X}\|_\circ\sqrt{\log G \vee p_{\max}\log n}$. Then, with probability tending to one, (5.17) is further bounded by

$$
\begin{aligned}
D_n\sum_{k=1}^{d}\|\beta_k - \beta_{0,k}\|_{2,1} &\le \frac{2D_n\|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_F|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|^{1/2}}{\|\boldsymbol{X}\|_\circ\phi_{\ell_{2,1}}(|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|)} - D_n\sum_{k=1}^{d}\|\beta_k - \beta_{0,k}\|_{2,1} \\
&= \frac{2D_n\sqrt{|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|\sum_{i=1}^{n}\|\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\|^2}}{\|\boldsymbol{X}\|_\circ\phi_{\ell_{2,1}}(|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|)} - D_n\sum_{k=1}^{d}\|\beta_k - \beta_{0,k}\|_{2,1}.
\end{aligned}
$$

The display is further bounded by

$$
\begin{aligned}
&\frac{2b_2D_n\sqrt{|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|\sum_{i=1}^{n}\|\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\boldsymbol{\Sigma}_0^{-1}\|^2}}{\|\boldsymbol{X}\|_\circ\phi_{\ell_{2,1}}(|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|)} - D_n\sum_{k=1}^{d}\|\beta_k - \beta_{0,k}\|_{2,1} \\
&\quad \le \frac{1}{2}\sum_{i=1}^{n}\|\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\boldsymbol{\Sigma}_0^{-1}\|^2 + \frac{2b_2^2D_n^2|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|}{\|\boldsymbol{X}\|_\circ^2\phi_{\ell_{2,1}}^2(|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|)} - D_n\sum_{k=1}^{d}\|\beta_k - \beta_{0,k}\|_{2,1},
\end{aligned}
$$

by the Cauchy-Schwarz inequality. Therefore, with probability tending to one, (5.16) is bounded by

$$\begin{aligned}
&\int_{\Theta_n^c} \exp\left(\frac{2b_2^2 D_n^2 |S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|}{\|\boldsymbol{X}\|_\circ^2 \phi_{\ell_{2,1}}^2(|S_{\boldsymbol{\beta}-\boldsymbol{\beta}_0}|)} - D_n \sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1}\right) dU(\boldsymbol{\beta}) \\
&\quad \le \exp\left(\frac{2b_2^2 D_n^2 (s_0 + M_2 s^\star)}{\|\boldsymbol{X}\|_\circ^2 \phi_{\ell_{2,1}}^2 (s_0 + M_2 s^\star)} - \frac{D_n \sqrt{M_4 n \epsilon_n^2 s^\star}}{2\|\boldsymbol{X}\|_\circ \phi_{\ell_{2,1}}(s_0 + M_2 s^\star)}\right) \\
&\qquad \times \sum_{S: s \le M_2 s^\star} \frac{\pi(s)}{\binom{Gd}{s}} \int_{\Theta_n^c} \prod_{k=1}^d \left(\prod_{j \in S_k} \left(\frac{\lambda_k}{a_j}\right)^{p_j} \right) \exp\left(-\frac{D_n}{2}\|\beta_k - \beta_{0,k}\|_{2,1}\right) d\beta_{S_k} \otimes \delta_{S_k^c}.
\end{aligned}$$

Since $c_4 \lambda_k / B_3 \le D_n$ for every $k \le d$, the last summation is bounded by

$$\sum_{S: s \le M_2 s^\star} \frac{\pi(s)}{\binom{Gd}{s}} \left(\frac{2B_3}{c_4}\right)^{p_S} \le 1,$$

where the inequality holds by making c_4 large enough. Now, plug in D_n and combine the display with (5.15) to obtain an upper bound of the expectation of (5.13). Since $a_j = O(p_j^{1/2})$ and $\pi(s_0) \gtrsim A_1^{s_0} / (G^{A_3} \vee n^{A_5 p_{\max}})^{s_0}$, the upper bound goes to zero as long as M_4 is chosen sufficiently large.

For a measurable subset $\mathcal{B}$ of $\mathbb{R}^{p \times d}$, we can write

$$\begin{aligned}
&\check{\Pi}_n(\mathcal{B} | Y_1, \dots, Y_n) \\
&\propto \int_{(\mathcal{B} \cap \Theta_n)} \int_{\mathcal{H}_n} \frac{\exp(\ell_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}))}{\exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}))} \exp\left(-\sum_{k=1}^d \lambda_k \|\beta_k\|_{2,1}\right) \exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})) d\Pi(\boldsymbol{\Sigma}) dU(\boldsymbol{\beta}),
\end{aligned}$$

and

$$\begin{aligned}
&\check{\Pi}_n^\infty(\mathcal{B} | Y_1, \dots, Y_n) \\
&\propto \int_{(\mathcal{B} \cap \Theta_n)} \int_{\mathcal{H}_n} \frac{\exp(\ell_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}_0))}{\exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0))} \exp\left(-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1}\right) \exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})) d\Pi(\boldsymbol{\Sigma}) dU(\boldsymbol{\beta}).
\end{aligned}$$

Note that for sequences of measures (μ_S) and (ν_S),

$$\left\| \frac{\sum_S \mu_S}{\|\sum_S \mu_S\|_{TV}} - \frac{\sum_S \nu_S}{\|\sum_S \nu_S\|_{TV}} \right\|_{TV} \le 2 \sup_S \left\| 1 - \frac{d\nu_S}{d\mu_S} \right\|_\infty,$$

(see e.g. page 2011 of Castillo et al. (2015)). Hence it suffices to show that

$$\mathbb{E}_0 \sup_{\boldsymbol{\beta} \in \Theta_n} \left| 1 - \frac{\int_{\mathcal{H}_n} \frac{\exp(\ell_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}))}{\exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}))} \exp\left(-\sum_{k=1}^d \lambda_k \|\beta_k\|_{2,1}\right) \exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})) d\Pi(\boldsymbol{\Sigma})}{\int_{\mathcal{H}_n} \frac{\exp(\ell_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}_0))}{\exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0))} \exp\left(-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1}\right) \exp(\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})) d\Pi(\boldsymbol{\Sigma})} \right| \to 0.$$

Using the property that $|1-\int f/\int g| \le (1-\inf(f/g)) \vee (\sup(f/g)-1) \le \sup|1-f/g|$, the expression in the last display is bounded by

$$\begin{aligned}
&\mathbb{E}_0 \sup_{\boldsymbol{\beta}\in\Theta_n} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \left|1-\exp\left(\tilde{\ell}_n(\boldsymbol{\beta},\boldsymbol{\Sigma}) - \sum_{k=1}^d \lambda_k(\|\beta_k\|_{2,1} - \|\beta_{0,k}\|_{2,1})\right)\right| \\
&\quad \le \mathbb{E}_0 \sup_{\boldsymbol{\beta}\in\Theta_n} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \Bigg\{ \left(|\tilde{\ell}_n(\boldsymbol{\beta},\boldsymbol{\Sigma})| + \max_{1\le k\le d}\lambda_k \sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1}\right) \\
&\qquad\qquad \times \exp\left(|\tilde{\ell}_n(\boldsymbol{\beta},\boldsymbol{\Sigma})| + \max_{1\le k\le d}\lambda_k \sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1}\right)\Bigg\},
\end{aligned}$$

where $\tilde{\ell}_n(\boldsymbol{\beta},\boldsymbol{\Sigma}) = \ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}) + \ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0) - \ell_n(\boldsymbol{\beta},\boldsymbol{\Sigma}_0) - \ell_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma})$. First, it is easy to see that $\sup\{\lambda_k \sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1} : \boldsymbol{\beta}\in\Theta_n,\ 1\le k\le d\} \to 0$ due to the small λ regime. To complete the proof, we shall show that

$$\mathbb{E}_0 \sup_{\boldsymbol{\beta}\in\Theta_n} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} |\tilde{\ell}_n(\boldsymbol{\beta},\boldsymbol{\Sigma})| \to 0. \tag{5.18}$$

It can be easily verified that

$$\begin{aligned}
|\tilde{\ell}_n(\boldsymbol{\beta},\boldsymbol{\Sigma})| \le \frac{1}{2}&\left|\sum_{i=1}^n \mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}_0^{-1})\tilde{\boldsymbol{X}}_i'\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\right| \\
&+\left|\sum_{i=1}^n \mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}_0^{-1})\big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\big)'\right|.
\end{aligned}$$

First note that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\Theta_n} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \left|\sum_{i=1}^n \mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}_0^{-1})\tilde{\boldsymbol{X}}_i'\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\right| \\
&\quad \le \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \|\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}_0^{-1}\| \sup_{\boldsymbol{\beta}\in\Theta_n} \|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_F^2 \\
&\quad \lesssim \|\boldsymbol{X}\|_\circ^2 \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0\| \sup_{\boldsymbol{\beta}\in\Theta_n} \left(\sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1}\right)^2,
\end{aligned}$$

where the last inequality holds by (5.2) and Assumption 3 since $\sup\{\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0\| : \boldsymbol{\Sigma}\in\mathcal{H}_n\}$ is small. The rightmost side of the display is bounded by $s^\star n\epsilon_n^3$ which goes to zero by the assumption. Similar to (5.17), we also obtain that

$$\begin{aligned}
&\mathbb{E}_0 \sup_{\boldsymbol{\beta}\in\Theta_n} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \left|\mathrm{Vec}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\sum_{i=1}^n \tilde{\boldsymbol{X}}_i(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}_0^{-1})\big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\big)'\right| \\
&\quad \le \mathbb{E}_0 \sup_{\boldsymbol{\beta}\in\Theta_n} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \sum_{j=1}^G \|\mathrm{Vec}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{0,j})\|_F \|W_{\boldsymbol{\Sigma},j}\|
\end{aligned}$$

$$\leq \mathbb{E}_0 \max_{1\leq j\leq G} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \|W_{\boldsymbol{\Sigma},j}\| \sup_{\boldsymbol{\beta}\in\Theta_n} \sum_{k=1}^{d} \|\beta_k - \beta_{0,k}\|_{2,1}, \tag{5.19}$$

where $W_{\boldsymbol{\Sigma},j} = \sum_{i=1}^n \tilde{\boldsymbol{X}}_{i,j}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1})\big(Y_i - \mathrm{Vec}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{X}}_i\big)'$. By Lemma 2.2.2 of van der Vaart and Wellner (1996) applied with $\psi_2(x) = \exp(x^2) - 1$, we have

$$\begin{aligned}
\mathbb{E}_0 \max_{1\leq j\leq G} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} \|W_{\boldsymbol{\Sigma},j}\| &\leq \sqrt{p_{\max} d}\, \mathbb{E}_0 \max_{1\leq j\leq G} \max_{1\leq \ell\leq p_j d} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} |W_{\boldsymbol{\Sigma},j,\ell}| \\
&\leq \sqrt{p_{\max} d} \left\| \max_{1\leq j\leq G} \max_{1\leq \ell\leq p_j d} \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} |W_{\boldsymbol{\Sigma},j,\ell}| \right\|_{\psi_2} \\
&\lesssim \sqrt{p_{\max} d \log G} \max_{1\leq j\leq G} \max_{1\leq \ell\leq p_j d} \left\| \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} |W_{\boldsymbol{\Sigma},j,\ell}| \right\|_{\psi_2},
\end{aligned}$$

where $\|\cdot\|_{\psi_2}$ denotes the Orlicz norm and $W_{\boldsymbol{\Sigma},j,\ell}$ is the ℓth element of $W_{\boldsymbol{\Sigma},j}$. By Lemma 2.2.1 of van der Vaart and Wellner (1996), we have that for every $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{H}_n$,

$$\|W_{\boldsymbol{\Sigma}_1,j,\ell} - W_{\boldsymbol{\Sigma}_2,j,\ell}\|_{\psi_2} \lesssim \sqrt{\mathrm{Var}(W_{\boldsymbol{\Sigma}_1,j,\ell} - W_{\boldsymbol{\Sigma}_2,j,\ell})} \leq \|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\| \|\boldsymbol{X}_j\|,$$

which is bounded by $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_F \|\boldsymbol{X}\|_\circ$, by the relations $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_F \leq \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0\|_F + \|\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_0\|_F \lesssim \epsilon_n$ and $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1} = -\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\boldsymbol{\Sigma}_2^{-1}$, and the fact that the eigenvalues of $\boldsymbol{\Sigma}$, and hence also those of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_2$, lie between two fixed positive numbers. We see that $W_{\boldsymbol{\Sigma},j,\ell}$ is a separable Gaussian process as $\mathcal{H}_n$ is a separable metric space under the Frobenius norm. Hence, by Corollary 2.2.5 of van der Vaart and Wellner (1996), for any fixed $\boldsymbol{\Sigma}' \in \mathcal{H}_n$ and some $c_5 > 0$,

$$\left\| \sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n} |W_{\boldsymbol{\Sigma},j,\ell}| \right\|_{\psi_2} \lesssim \|W_{\boldsymbol{\Sigma}',j,\ell}\|_{\psi_2} + \int_0^{c_5\|\boldsymbol{X}\|_\circ \mathrm{diam}_j(\mathcal{H}_n)} \sqrt{\log N\left(\frac{\epsilon}{2c_5\|\boldsymbol{X}\|_\circ}, \mathcal{H}_n, \|\cdot\|_F\right)} d\epsilon,$$

where $\mathrm{diam}_j(\mathcal{H}_n) = \sup\{\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_F : \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{H}_n\}$. By Lemma 2.2.1 of van der Vaart and Wellner (1996) again, we have that

$$\|W_{\boldsymbol{\Sigma}',j,\ell}\|_{\psi_2} \lesssim \sqrt{\mathrm{Var}(W_{\boldsymbol{\Sigma}',j,\ell})} \leq \|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Sigma}'^{-1} - \boldsymbol{\Sigma}_0^{-1})\| \|\boldsymbol{X}_j\| \lesssim \|\boldsymbol{\Sigma}' - \boldsymbol{\Sigma}_0\|_F \|\boldsymbol{X}\|_\circ.$$

We also obtain that

$$\begin{aligned}
&\int_0^{c_5\|\boldsymbol{X}\|_\circ \mathrm{diam}_j(\mathcal{H}_n)} \sqrt{\log N\left(\frac{\epsilon}{2c_5\|\boldsymbol{X}\|_\circ}, \mathcal{H}_n, \|\cdot\|_F\right)} d\epsilon \\
&\quad\leq \int_0^{2c_5\sqrt{M_1}\|\boldsymbol{X}\|_\circ \epsilon_n} \sqrt{d^2 \log\left(\frac{6c_5\sqrt{M_1}\|\boldsymbol{X}\|_\circ \epsilon_n}{\epsilon}\right)} d\epsilon \\
&\quad= 12c_5\sqrt{M_1}\|\boldsymbol{X}\|_\circ d\epsilon_n \int_{\sqrt{\log 3}}^{\infty} t^2 e^{-t^2} dt.
\end{aligned}$$

Since the integral term on the rightmost side of the last display is bounded, we finally verify that $\big\|\sup_{\boldsymbol{\Sigma}\in\mathcal{H}_n}|W_{\boldsymbol{\Sigma},j,\ell}|\big\|_{\psi_2} \lesssim \|\boldsymbol{X}\|_\circ d\epsilon_n$ for every j and ℓ. Putting everything together, (5.19) is bounded by a multiple of $\epsilon_n^2\sqrt{p_{\max} n d^3 s^\star \log G}$ which goes to zero by the assumption. We finally verify (5.18), and hence the proof is complete. □

Proof of Theorem 3.7. We only need to show that

$$\sup_{\substack{\boldsymbol{\beta}_0\in\{\mathcal{B}_0:s_0\le s_n\delta_n(s_0)\le\eta_n,\\ \phi_{\ell_{2,1}}(s_0+M_2s^\star)>c\},\,\boldsymbol{\Sigma}_0\in\mathcal{H}_0}} \mathbb{E}_0\Pi(\boldsymbol{\beta}:S_{\boldsymbol{\beta},1}\supset S_{0,1},\dots,S_{\boldsymbol{\beta},d}\supset S_{0,d},S_{\boldsymbol{\beta}}\neq S_0|Y_1,\dots,Y_n)\to 0.$$

Then the theorem follows by the Beta-min condition. Our proof is similar to the proof of Theorem 4 of Castillo et al. (2015).

Let $\mathcal{S}_n=\{S:s\le M_2s^\star,S_1\supset S_{0,1},\dots,S_d\supset S_{0,d},S\neq S_0\}$ and $\boldsymbol{\Gamma}_S=\sum_{i=1}^n\tilde{\boldsymbol{X}}_{i,S}\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{X}}_{i,S}'$. By Theorem 3.6, it suffices to show that $\Pi^\infty(\boldsymbol{\beta}:S_{\boldsymbol{\beta}}\in\mathcal{S}_n|Y_1,\dots,Y_n)\to 0$ in probability. By (3.11), we obtain that $\Pi^\infty(\boldsymbol{\beta}:S_{\boldsymbol{\beta}}\in\mathcal{S}_n|Y_1,\dots,Y_n)\le\sum_{S\in\mathcal{S}_n}w_S^\infty/w_{S_0}^\infty$ which can be bounded by

$$\sum_{\bar{s}=s_0+1}^{M_2s^\star}\left\{\frac{\pi(\bar{s})\binom{Gd}{s_0}\binom{Gd-s_0}{\bar{s}-s_0}}{\pi(s_0)\binom{Gd}{\bar{s}}}\max_{S\in\mathcal{S}_n:s=\bar{s}}\left[\left(\prod_{k=1}^d\prod_{j\in S_k}\left(\frac{\lambda_j\sqrt{2\pi}}{a_j}\right)^{p_j}\right)\left(\frac{\det\boldsymbol{\Gamma}_{S_0}}{\det\boldsymbol{\Gamma}_S}\right)^{1/2}\right.\right.$$
$$\left.\left.\times\exp\left(\frac{1}{2}\sum_{i=1}^n\|\boldsymbol{\Sigma}_0^{-1/2}\tilde{\boldsymbol{X}}_{i,S}'\hat{\beta}_S^\star\|^2-\frac{1}{2}\sum_{i=1}^n\|\boldsymbol{\Sigma}_0^{-1/2}\tilde{\boldsymbol{X}}_{i,S_0}'\hat{\beta}_{S_0}^\star\|^2\right)\right]\right\}. \tag{5.20}$$

To bound further, we bound each factor in the above expression.

The interlacing theorem applied to $\boldsymbol{\Gamma}_S$ and its principal submatrix $\boldsymbol{\Gamma}_{S_0}$ gives $\operatorname{eig}_m(\boldsymbol{\Gamma}_{S_0})\le\operatorname{eig}_m(\boldsymbol{\Gamma}_S)$, $m=1,\dots,\sum_{k=1}^d\sum_{j\in S_{0,k}}p_j$, we have

$$\det(\boldsymbol{\Gamma}_{S_0})\le\prod_m\operatorname{eig}_m(\boldsymbol{\Gamma}_S)\le(\operatorname{eig}_1(\boldsymbol{\Gamma}_S))^{p_{S_0}-p_S}\det(\boldsymbol{\Gamma}_S),$$

so by (3.6), $\det(\boldsymbol{\Gamma}_{S_0})/\det(\boldsymbol{\Gamma}_S)$ is bounded by $(b_2^{-1}\phi_{\ell_2}(s)\|\boldsymbol{X}\|_\circ)^{2(p_{S_0}-p_S)}$.

The exponential term $Q_S:=\sum_{i=1}^n\|\boldsymbol{\Sigma}_0^{-1/2}\tilde{\boldsymbol{X}}_{i,S}'\hat{\beta}_S^\star\|^2-\sum_{i=1}^n\|\boldsymbol{\Sigma}_0^{-1/2}\tilde{\boldsymbol{X}}_{i,S_0}'\hat{\beta}_{S_0}^\star\|^2$ in (5.20) has a χ^2-distribution with degree of freedom $p_{S_0}-p_S$. By Markov's inequality on the exponential moment, we have that for every $0<u<1/2$ and $r>0$,

$$\begin{aligned}&\mathbb{P}_0\left(\max_{S\in\mathcal{S}_n:s=\bar{s}}Q_S\ge r(\bar{s}-s_0)(\log G\vee p_{\max}\log n)\right)\\&\quad\le\exp\Big(-ur(\bar{s}-s_0)(\log G\vee p_{\max}\log n)\Big)\mathbb{E}_0\left(\max_{S\in\mathcal{S}_n:s=\bar{s}}e^{uQ_S}\right)\\&\quad\le N_{\bar{s}}\exp\Big(-ur(\bar{s}-s_0)(\log G\vee p_{\max}\log n)\Big)(1-2u)^{-(p_{S_0}-p_S)/2},\end{aligned}$$

where $N_{\bar{s}}=\binom{Gd-s_0}{\bar{s}-s_0}$ is the cardinality of the set $\{S\in\mathcal{S}_n:s=\bar{s}\}$. Since $N_{\bar{s}}\le(Gd)^{\bar{s}-s_0}$ and $d^2\log n\ll n$, we have that for some $c>0$,

$$\mathbb{P}\Big(Q_S\ge r(\bar{s}-s_0)(\log G\vee p_{\max}\log n),\ \text{for any } S\in\mathcal{S}_n\Big)$$

$$\leq \sum_{\bar{s}>s_0} \exp\left(-ur(\bar{s}-s_0)(\log G \vee p_{\max}\log n) + \frac{3}{2}(\bar{s}-s_0)\log G + c(\bar{s}-s_0)p_{\max}\right)$$

which goes to 0 whenever $ur > 3/2$. If $r > 3$, this is ensured by choosing u arbitrarily close to $1/2$. Thus with probability tending to 1, (5.20) is bounded by

$$\sum_{s=s_0+1}^{M_2 s^\star} \frac{A_1^{s-s_0} s^{s-s_0}}{(G \vee n^{p_{\max}})^{A_4(s-s_0)}} \left(\frac{\max_{1\leq k\leq d}\lambda_k\sqrt{2\pi}}{b_2^{-1}\|\boldsymbol{X}\|_\circ \phi_{\ell_2}(s)}\right)^{p_{\max}(\bar{s}-s_0)} \frac{1}{(G\vee n^{p_{\max}})^{r(\bar{s}-s_0)/2}}. \quad (5.21)$$

Under the small λ regime, for every S such that $s \leq M_2 s^\star \lesssim G^a$,

$$\left(\frac{\max_{1\leq k\leq d}\lambda_k\sqrt{2\pi}}{b_2^{-1}\|\boldsymbol{X}\|_\circ \phi_{\ell_2}(s)}\right)^{p_{\max}(\bar{s}-s_0)} \leq \left(\frac{\max_{1\leq k\leq d}\lambda_k\sqrt{2\pi M_2 s^\star}}{b_2^{-1}\|\boldsymbol{X}\|_\circ \phi_{\ell_{2,1}}(M_2 s^\star)}\right)^{p_{\max}(\bar{s}-s_0)} \lesssim 1.$$

and hence (5.21) goes to 0 if $a - A_4 + r/2 < 0$. If $A_4 > a + 3/2$, this is ensured by choosing r arbitrarily close to 3.

□

6. Supplement to "Bayesian Linear Regression for Multivariate Responses Under Group Sparsity"

The following lemma obtains the normalizing constant in the density proportional to $e^{-\lambda\|x\|}$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$.

Lemma 6.1. *For $a_m = \sqrt{\pi}\,(\Gamma(m+1)/\Gamma(m/2+1))^{1/m}$,*

$$\int_{\mathbb{R}^m} \left(\frac{\lambda}{a_m}\right)^m \exp(-\lambda\|(x_1,\ldots,x_m)\|)dx_1\cdots dx_m = 1. \quad (6.1)$$

Also as $m \to \infty$, $a_m \asymp m^{1/2}$.

If x is expressed in terms of the spherical polar coordinates by a radius r, a base angle $\theta_{m-1} \in (0, 2\pi)$, and $m-2$ angles $\theta_1, \ldots, \theta_{m-2}$ ranging over $(-\pi/2, \pi/2)$, then the density of r is given by

$$f(r|\lambda) = \frac{\lambda^m}{\Gamma(m)} r^{m-1}\exp(-\lambda r), \quad (6.2)$$

the gamma density with the shape parameter m and rate parameter λ.

Proof. Applying the polar transformation $(x_1, \ldots, x_m) \mapsto (r, \theta_1, \ldots, \theta_{m-1})$, evaluating the Jacobian, and applying the results shown in Chapter 1.5.1 of Scott (2015), the integral in (6.1) equals to

$$\int_0^{2\pi}\int_{-\pi/2}^{\pi/2}\cdots\int_{-\pi/2}^{\pi/2}\int \left(\frac{\lambda}{a}\right)^m r^{m-1}e^{-\lambda r}\prod_{i=1}^{m-2}(\cos\theta_{m-i-1})^i dr\, d\theta_1\cdots d\theta_{m-2}\, d\theta_{m-1}$$

$$= \int_0^\infty \frac{2\pi^{m/2}\lambda^m}{\Gamma(m/2)a^m} r^{m-1} e^{-\lambda r} dr. \tag{6.3}$$

The second line of the last display is obtained by using the results in Chapter 1.5.2 of Scott (2015). Since $\int_0^\infty r^{m-1}e^{-\lambda r}dr = \Gamma(m)/\lambda^m$, the value

$$a_m = \sqrt{\pi}\left(\frac{2\Gamma(m)}{\Gamma(m/2)}\right)^{1/m} = \sqrt{\pi}\left(\frac{\Gamma(m+1)}{\Gamma(m/2+1)}\right)^{1/m}$$

makes $(\lambda/a_m)^m \exp\left(-\lambda\|(x_1,\ldots,x_m)\|\right)$ a probability density function.

Now by Stirling's approximation to the gamma functions, we obtain that

$$\frac{\sqrt{2}\pi}{e}\left(\frac{2m}{e}\right)^{1/2} \le a_m \le \frac{e}{\sqrt{2}}\left(\frac{2m}{e}\right)^{1/2},$$

which implies that $a_m \asymp m^{1/2}$. The relation (6.2) is evident from (6.3). □

Theorem 6.2. *Consider the setup of Theorem 3.1 except that the prior on $\boldsymbol{\Sigma}$ is given by the inverse-Wishart distribution $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}_d(\kappa d^2, \boldsymbol{\Phi})$ such that $c^{-1} \le \mathrm{eig}_1(\boldsymbol{\Phi}) \le \mathrm{eig}_d(\boldsymbol{\Phi}) \le c$ for some constant $c > 1$, where $\mathcal{W}_d(\nu, \boldsymbol{\Psi})$ stands for the Wishart distribution in dimension d with ν degrees of freedom and scale matrix $\boldsymbol{\Psi}$. Then for a sufficiently large $M' > 0$,*

$$\sup_{\boldsymbol{\beta}_0\in\mathcal{B}_0,\boldsymbol{\Sigma}_0\in\mathcal{H}_0} \mathbb{E}_0\Pi\Big(\boldsymbol{\beta} : s_{\boldsymbol{\beta}} \ge M'\tilde{s}^\star \Big| Y_1,\ldots,Y_n\Big) \to 0, \tag{6.4}$$

$$\sup_{\boldsymbol{\beta}_0\in\mathcal{B}_0,\boldsymbol{\Sigma}_0\in\mathcal{H}_0} \mathbb{E}_0\Pi\Big(\boldsymbol{\beta} : \|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_F^2 \ge M'n\tilde{\epsilon}_n^2 \Big| Y_1,\ldots,Y_n\Big) \to 0, \tag{6.5}$$

$$\sup_{\boldsymbol{\beta}_0\in\mathcal{B}_0,\boldsymbol{\Sigma}_0\in\mathcal{H}_0} \mathbb{E}_0\Pi\Big(\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0\|_F^2 \ge M'\tilde{\epsilon}_n^2 \Big| Y_1,\ldots,Y_n\Big) \to 0, \tag{6.6}$$

where

$$\tilde{s}^\star = \max\left\{s_0, \frac{d^3\log n}{\log G \vee p_{\max}\log n}\right\}, \tag{6.7}$$

$$\tilde{\epsilon}_n = \max\left\{\sqrt{\frac{s_0\log G}{n}}, \sqrt{\frac{s_0 p_{\max}\log n}{n}}, \sqrt{\frac{d^3\log n}{n}}\right\}. \tag{6.8}$$

Remark 3. Once $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are confined in small neighborhoods around the true values, the distributional approximation in Theorem 3.6 and the selection consistency in Theorem 3.7 remain valid with the revised rate $\tilde{\epsilon}_n$ given in (6.8). This can be shown by imitating the proofs of these theorems with the inverse Wishart prior on $\boldsymbol{\Sigma}$, as the proofs of these results do not require a specific prior.

To prove the theorem, we need the following lemma giving estimates on the distribution of eigenvalues of a Wishart matrix.

Lemma 6.3. *If $\mathbf{\Sigma}^{-1} \sim \mathcal{W}_d(\nu, \mathbf{\Psi})$, where $\nu \geq d$ is an integer, $0 < \rho_1 < \cdots < \rho_d$ are its eigenvalues, then for $t_1 > \nu d$, $t_2 > 0$, $0 \leq t_3 \leq 1$, and $0 \leq a_1 \leq \cdots \leq a_d$,*

$$\begin{aligned}
\mathbb{P}\left(\rho_d \geq t_1 \|\mathbf{\Psi}\|\right) &\leq \left(\frac{t_1}{\nu d}\right)^{\nu d/2} \exp(\nu d/2 - t_1/2), \qquad (6.9)\\
\mathbb{P}\left(\rho_1 \leq t_2\right) &\leq \left(\frac{\nu + d}{2e}\right)^{d(\nu+d)/2} \frac{(e(\nu+d)/\sqrt{\pi})^d}{2^{(\nu+d+1)/2}} t_2^{(\nu-d-1)/2}\\
&\quad \times (\det(\mathbf{\Psi}))^{-\nu/2} \|\mathbf{\Psi}\|^{(d-1)(\nu+1)/2}, \qquad (6.10)\\
\mathbb{P}\left(\bigcap_{k=1}^{d} \{a_k \leq \rho_k \leq a_k(1+t_3)\}\right) &\geq \left(\frac{a_1 t_3 e^2 \nu}{8\sqrt{\pi}}\right)^{-d} \left(\frac{2\nu d}{e a_1 t_3}\right)^{-\nu d/2} \left(\frac{d}{2e}\right)^{-d^2/2}\\
&\quad \times (\det(\mathbf{\Psi}))^{-\nu/2} \exp\left(-\frac{a_1(1+t_3)\operatorname{Tr}(\mathbf{\Psi}^{-1})}{2}\right). \qquad (6.11)
\end{aligned}$$

Remark 4. To control the sequences appearing in the estimates of prior probabilities in Lemma 6.3 such that explicit growth estimates can be obtained for use in the rate theorem, the degrees of freedom of the Wishart prior on $\mathbf{\Sigma}^{-1}$ needs to be taken approximately proportional to the dimension d^2. By choosing ν to be the integer part of κd^2 for some constant $\kappa \geq 1$, the estimates in Lemma 6.3 simplify to

$$\begin{aligned}
\mathbb{P}\left(\rho_d \geq t_1 \|\mathbf{\Psi}\|\right) &\leq \left(b_1 t_1/d^3\right)^{b_2 d^3} \exp(b_3 d^3 - t_1/2),\\
\mathbb{P}\left(\rho_1 \leq t_2\right) &\leq \left(b_4 d^2\right)^{b_5 d^3} t_2^{b_6 d^2} (\det(\mathbf{\Psi}))^{-\kappa d^2/2} \|\mathbf{\Psi}\|^{b_7 d^3}\\
\mathbb{P}\left(\bigcap_{k=1}^{d} \{a_k \leq \rho_k \leq a_k(1+t_3)\}\right) &\geq (b_8 a_1 t_3 d^2)^{-d} (b_9 d^3/(a_1 t_3))^{-b_{10} d^3} (b_{11} d)^{-d^2/2}\\
&\quad \times (\det(\mathbf{\Psi}))^{-\kappa d^2/2} \exp\left(-a_1(1+t_3)\operatorname{Tr}(\mathbf{\Psi}^{-1})/2\right),
\end{aligned}$$

for some constants $b_1, \ldots, b_{11} > 0$.

Remark 5. In Theorem 6.2, the degree of freedom ν is chosen to grow in proportion to d^2. This choice makes the first two terms of $\tilde{\epsilon}_n$ the same as those in ϵ_n, but induces slightly increased $\tilde{s}^\star$ in (6.7) compared to $s^\star$ as the prior concentration decreases. Instead, one may choose ν that is proportional to d. Then the prior concentration stays and the assertion (6.4) holds with $s^\star$ instead of $\tilde{s}^\star$, so the same dimension recovery result is obtained as Lemma 3.2. However, this significantly weakens the rate to $\max\{\sqrt{(ds_0 \log G)/n}, \sqrt{(ds_0 p_{\max} \log n)/n}, \sqrt{(d^3 \log n)/n}\}$, because a weaker right-tail decay of the largest eigenvalue of $\mathbf{\Sigma}$ necessitates the use of a larger sieve, which increases the entropy.

Proof of Lemma 6.3. If the dimension d remains bounded, the result is already given in Lemma 9.16 of Ghosal and van der Vaart (2017). For $d \to \infty$, the dependence of the constants on d must be explicitly identified. Below we carefully estimate the normalizing constant, assuming that d is sufficiently large.

To prove (6.9), consider the random matrix $\boldsymbol{\Omega} = \boldsymbol{\Psi}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{1/2}$. Observe that then $\boldsymbol{\Omega}^{-1} \sim \mathcal{W}_d(\nu, \boldsymbol{I}_d)$. Since $\rho_d = \|\boldsymbol{\Sigma}^{-1}\| = \|\boldsymbol{\Psi}^{1/2}\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}^{1/2}\| \leq \|\boldsymbol{\Psi}\|\|\boldsymbol{\Omega}^{-1}\|$, we have that

$$\mathbb{P}(\rho_d \geq t_1\|\boldsymbol{\Psi}\|) \leq \mathbb{P}(\|\boldsymbol{\Omega}^{-1}\| \geq t_1) \leq \mathbb{P}(\mathrm{Tr}(\boldsymbol{\Omega}^{-1}) \geq t_1)$$

and $\mathrm{Tr}(\boldsymbol{\Omega}^{-1}) \sim \chi^2_{\nu d}$. Then apply the Chernoff bound for a χ^2-distribution, we obtain (6.9).

To prove (6.10) and (6.11), we need estimates for the multivariate gamma function from both sides. By Stirling's approximation to the gamma functions, $e(n/e)^n \leq \Gamma(n+1) \leq en(n/e)^n$. Thus we have

$$\begin{aligned}\Gamma_d(\nu/2) &= \pi^{d(d-1)/4}\prod_{k=1}^{d}\Gamma\left(\frac{\nu+1-k}{2}\right)\\ &\leq \pi^{d(d-1)/4}\left(\Gamma\left(\nu/2+1\right)\right)^d\\ &\leq \pi^{d(d-1)/4}e^{-\nu d/2+d}\left(\nu/2\right)^{(\nu/2+1)d}\end{aligned}$$

and since $\nu \geq d$,

$$\Gamma_d(\nu/2) \geq \pi^{d(d-1)/4}(\Gamma(1/2))^d = \pi^{d(d-1)/4+d/2}.$$

To prove (6.11), we need the following three inequalities:

1. $\displaystyle\prod_{1\leq k<k'\leq d}(\rho_{k'}-\rho_k) \leq \prod_{1\leq k<k'\leq d}\rho_{k'} = \prod_{k=2}^{d}\rho_k^{k-1}$;
2. $\displaystyle\exp\left(-\frac{\mathrm{Tr}(\boldsymbol{\Psi}^{-1}\boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}')}{2}\right) \leq \exp\left(-\frac{\mathrm{Tr}(\boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}')}{2\|\boldsymbol{\Psi}\|}\right) = \exp\left(-\sum_{k=1}^{d}\frac{\rho_k}{2\|\boldsymbol{\Psi}\|}\right)$;
3. $\displaystyle\frac{\pi^{d^2/2}2^{-d\nu/2}\left(\det(\boldsymbol{\Psi})\right)^{-\nu/2}}{\Gamma_d(d/2)\Gamma_d(\nu/2)} \leq \pi^{-d/2}2^{-d\nu/2}\left(\det(\boldsymbol{\Psi})\right)^{-\nu/2}$, which is a consequence of the lower bound for the multivariate gamma function.

Note that $\boldsymbol{D}^{-1} = \mathrm{diag}(\rho_1, \ldots, \rho_d)$. Then by plugging-in the above three upper bounds in the expression for the joint density of the eigenvalues of a Wishart matrix (see, e.g., equation (9.6) of Ghosal and van der Vaart (2017)) and integrating, the marginal density of ρ_1 is bounded by

$$\begin{aligned}&\pi^{-d/2}2^{-d\nu/2}\left(\det(\boldsymbol{\Psi})\right)^{-\nu/2}\rho_1^{(\nu-1-d)/2}e^{-\rho_1/(2\|\boldsymbol{\Psi}\|)}\\ &\times\prod_{k=2}^{d}\int_0^\infty\rho_k^{(\nu-1-d)/2+k-1}\exp(-\rho_k/(2\|\boldsymbol{\Psi}\|))d\rho_k.\end{aligned}$$

Each integral function in the last display equals to $\Gamma\left(\frac{\nu-d-1}{2}+k\right)(2\|\boldsymbol{\Psi}\|)^{(\nu-d-1)/2+k}$. Now applying the upper bound of the gamma function, we obtain

$$\Gamma\left(\frac{\nu-d-1}{2}+k\right)\leq\Gamma\left(\frac{\nu+d}{2}+1\right)\leq\left(\frac{\nu+d}{2}\right)^{(\nu+d)/2+1}e^{1-\nu+d/2}.$$

With $\rho_1\leq t_2$, the marginal density of ρ_1 can be further bounded above by

$$\pi^{-d/2}2^{-\nu d/2}(\det(\boldsymbol{\Psi}))^{-\nu/2}\left(\frac{\nu+d}{2}\right)^{d(\nu+d)/2+d}t_2^{(\nu-d-1)/2}e^{-d(\nu+d)/2+d}\,(2\|\boldsymbol{\Psi}\|)^{(d-1)(\nu+1)/2},$$

as $\sum_{k=2}^{d}((\nu-d-1)/2+k)=(d-1)(\nu+1)/2$ if $d\geq 2$. The last display equals to the upper bound of (6.10).

To prove (6.11), let $I_k=\{a_k(1+(k-1/2)t_3/d),a_k(1+kt_3/d)\}$ for each $k\in\{1,\dots,d\}$. Then $\rho_k\in I_k$ implies that $\rho_k\in[a_k,a_k(1+t_3)]$. Therefore integrating the expression for the joint density of the eigenvalues and using the estimates of the normalizing constants given above, we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{k=1}^{d}\{\boldsymbol{\Sigma}:a_k\leq\rho_k\leq a_k(1+t_3)\}\right)\\
&\geq\frac{\pi^{d^2/2}2^{-d\nu/2}\,(\det(\boldsymbol{\Psi}))^{-\nu/2}}{\Gamma_d(d/2)\Gamma_d(\nu/2)}\int_{I_d}\cdots\int_{I_1}\Bigg\{\prod_{k=1}^{d}\rho_k^{(\nu-d-1)/2}\prod_{k<k'}(\rho_{k'}-\rho_k)\\
&\qquad\qquad\times\int_{\mathscr{O}(d)}\exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{\Psi}^{-1}\boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}')\right)\Bigg\}d\boldsymbol{P}\,d\rho_1\cdots d\rho_d\\
&\geq\frac{\pi^{d^2/2}2^{-d\nu/2}\,(\det(\boldsymbol{\Psi}))^{-\nu/2}}{\Gamma_d(d/2)\Gamma_d(\nu/2)}\left(\frac{a_1t_3}{2d}\right)^{(\nu-2)d/2}\exp\left(-\frac{a_d(1+t_3)}{2}\mathrm{Tr}(\boldsymbol{\Psi}^{-1})\right).\qquad(6.12)
\end{aligned}$$

The lower bound in the third line of the last display is obtained by noticing that for $k'>k$, $\rho_{k'}-\rho_k\geq a_1t_3/(2d)$, $-\boldsymbol{D}^{-1}\geq-\rho_d\boldsymbol{I}_d>-a_d(1+t_3)\boldsymbol{I}_d$, and $\boldsymbol{P}\boldsymbol{P}'=\boldsymbol{I}_d$. Now we plug the upper bound for the multivariate gamma function in (6.12) to obtain the lower bound in (6.11). □

Proof of Theorem 6.2. It only suffices to obtain estimates of prior concentration and define an appropriate sieve for this prior such that the complement has exponentially small prior probability. The proof is very similar to that of Theorem 3.1 employing the same overall strategy, except when estimates regarding the prior concentration of the covariance matrix are involved. The estimates of the prior mass outside the sieve and that of the entropy of the sieve must be obtained afresh since a different sieve is used.

Since the negative logarithm of the average Kullback-Leibler neighborhood of size $\tilde{\epsilon}_n^2$ should be controlled, the probabilities of the sets $\{\boldsymbol{\beta}:\sum_{k=1}^{d}\|\beta_k-\beta_{0,k}\|_{2,1}\leq c\tilde{r}_n\}$ and $\{\boldsymbol{\Sigma}:\|\boldsymbol{\Sigma}^{*-1}-\boldsymbol{I}\|\leq\tilde{\epsilon}_n\}$ need to be obtained, where $\tilde{r}_n=\sqrt{n\tilde{\epsilon}_n^2}/\|\boldsymbol{X}\|_\circ$ and $\boldsymbol{\Sigma}^*=\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1/2}$ as before. For the former, it is easy to see that the prior concentration of $\boldsymbol{\beta}$ is bounded by a constant multiple of $n\tilde{\epsilon}_n^2$, in view of the proof of Lemma 5.1. The condition for the latter clearly holds if all eigenvalues of $\boldsymbol{\Sigma}^{*-1}$ lie between 1 and $1+d^{-1/2}\tilde{\epsilon}_n$. In view of the third assertion in Remark 4, the prior probability of this event is at least

$$-d\log(b_8 d^{3/2}\tilde{\epsilon}_n) - b_{10}d^3\log\left(\frac{b_9 d^{7/2}}{\tilde{\epsilon}_n}\right) - \frac{d^2}{2}\log(b_{11}d) - \frac{\kappa d^3}{2}\log\|\boldsymbol{\Psi}\| - \frac{d+d^{1/2}\tilde{\epsilon}_n}{2}\log\|\boldsymbol{\Psi}^{-1}\|,$$

which is bounded below by a constant multiple of $-d^3\log n$. Thus, the estimate for the prior concentration is controlled. Then using the same techniques in the proof of Lemma 3.2, dimension recovery is still valid with $s^\star$ replaced by $\tilde{s}^\star$, which verifies (6.4).

Next, to prove (6.5) and (6.6) define the sieve

$$\tilde{\mathcal{F}}_n = \Big\{(\boldsymbol{\beta},\boldsymbol{\Sigma}) \in \mathcal{B}_n \times \mathcal{H} : s \le M'\tilde{s}^\star,\ \max_{\substack{1\le j\le G\\ 1\le k\le d}} \|\beta_{jk}\| \le H_n,$$
$$\exp(-Mn\tilde{\epsilon}_n^2/d^2) < \operatorname{eig}_1(\boldsymbol{\Sigma}^{-1}),\ \operatorname{eig}_d(\boldsymbol{\Sigma}^{-1}) \le n\Big\},$$

for a sufficiently large $M>0$, where $H_n = p_{\max}n/\underline{\lambda}$. Recall that the expression of $\underline{\lambda}$ is shown in (2.3).

We shall verify that

$$\Pi((\mathcal{B}_n \times \mathcal{H}) \setminus \tilde{\mathcal{F}}_n) \le \exp\left(-(1+C_1)n\tilde{\epsilon}_n^2\right), \tag{6.13}$$

as long as M is chosen sufficiently large.

Following (5.5), it suffices to bound only the terms $\Pi\left(\operatorname{eig}_1(\boldsymbol{\Sigma}^{-1}) \le \exp(-Mn\tilde{\epsilon}_n^2/d^2)\right)$ and $\Pi\left(\operatorname{eig}_d(\boldsymbol{\Sigma}^{-1}) \ge n\right)$ as the rest is unchanged.

In view of Remark 4, these terms are bounded above by

$$\exp\left(c_1 d^3\log n - c_2 Mn\tilde{\epsilon}_n^2\right) + \exp\left(c_3 d^3 \log n - c_4 n\right),$$

where $c_1,\dots,c_4$ are positive constants. Thus if M is chosen sufficiently large, we have (6.13).

To complete the proof of (5.6), we need to show that $\log N_* \lesssim n\epsilon_n^2$, where N_* is the number of pieces satisfying (5.8) needed to cover the sieve $\tilde{\mathcal{F}}_n$. It is easy to see that $\log N_*$ is bounded by

$$\log N\Big(\frac{1}{\tilde{s}^\star\sqrt{p_{\max}n}\|\boldsymbol{X}\|_\circ}, \Big\{\boldsymbol{\beta} : s_{\boldsymbol{\beta}} \le M'\tilde{s}^\star,\ \max_{\substack{1\le j\le G\\ 1\le k\le d}} \|\beta_{jk}\| \le H_n\Big\}, \|\cdot\|_\infty\Big)$$
$$+ \log N\Big(\frac{1}{n^2 d}, \left\{\boldsymbol{\Sigma} : \exp(-Mn\tilde{\epsilon}_n^2/d^2) < \operatorname{eig}_1(\boldsymbol{\Sigma}^{-1}),\ \operatorname{eig}_d(\boldsymbol{\Sigma}^{-1}) < n\right\}, \|\cdot\|\Big).$$

Similar to (5.10), it can be easily verified that the estimate of the first term is bounded by a constant multiple of $n\tilde{\epsilon}_n^2$, so we only need to estimate the second term, which is bounded by

$$\log N\left(\frac{1}{n^2 d}, \left\{\boldsymbol{\Sigma} : \exp(-n\tilde{\epsilon}_n^2/d^2) \le \operatorname{eig}_1(\boldsymbol{\Sigma}^{-1})\right\}, \|\cdot\|\right)$$

$$\begin{aligned}
&\leq \log N\left(\frac{1}{n^2 d}, \left\{\mathbf{\Sigma} : \|\mathbf{\Sigma}\|_F < \sqrt{d}\exp(Mn\tilde{\epsilon}_n^2/d^2)\right\}, \|\cdot\|_F\right) \\
&\leq d^2 \log\left(n^2 d^{3/2}\exp(Mn\tilde{\epsilon}_n^2/d^2)\right).
\end{aligned}$$

The last expression is easily seen to be bounded by a constant multiple of $n\tilde{\epsilon}_n^2$.

□

References

Banerjee, S. and S. Ghosal (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics 8*, 2111–2137.

Banerjee, S. and S. Ghosal (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis 136*, 147–162.

Belitser, E. and S. Ghosal (2019). Empirical Bayes oracle uncertainty quantification for regression. *The Annals of Statistics (to appear)*.

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

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bühlmann, P. and S. van der Geer (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer-Verlag.

Castillo, I. and R. Mismer (2018). Empirical Bayes analysis of spike and slab posterior distributions. *arXiv:1801.01696*.

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

Chae, M., L. Lin, and D. B. Dunson (2019). Bayesian sparse linear regression with unknown symmetric error. *Information and Inference: A Journal of the IMA 01*, 1–33.

Chen, R.-B., C.-H. Chu, S. Yuan, and Y. N. Wu (2016). Bayesian sparse group selection. *Journal of Computational and Graphical Statistics 25*, 665–683.

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

de Jonge, R. and H. van Zanten (2013). Semiparametric Bernstein-von Mises for the error standard deviation. *Electronic Journal of Statistics 7*, 217–243.

Gao, C., A. W. van der Vaart, and H. H. Zhou (2015). A general framework for Bayes structured linear models. *arXiv preprint arXiv:1506.02174*.

Gao, C. and H. H. Zhou (2016). Bernstein-von Mises theorems for functionals of the covariance matrix. *Electronic Journal of Statistics 10*, 1751–1806.

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

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis 74*(1), 49–68.

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Greenlaw, K., E. Szefer, J. Graham, M. Lesperance, and F. S. Nathoo (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics 33*, 2513–2522.

Hsu, D., S. Kakade, and T. Zhang (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability 17*.

Huang, J. and T. Zhang (2010). The benefit of group sparsity. *The Annals of Statistics 38*, 1978–2004.

Li, F. and N. R. Zhan (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association 105*(491), 1202–1214.

Liquet, B., K. Mengersen, A. N. Pettitt, and M. Sutton (2017). Bayesian variable selection regression of multivariate responses for group data. *Bayesian Analysis 12*, 1039–1067.

Lounici, K., M. Pontil, A. B. Tsybakov, and S. van de Geer (2009). Taking advantage of sparsity in multi-task learning. *In Proceedings of the 22nd Annual Conference on Learning Theory (COLT-2009)*, 73–82.

Lounici, K., M. Pontil, S. van de Geer, and A. B. Tsybakov (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics 39*, 2164–2204.

Martin, R., R. Mess, and S. G. Walker (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli 23*, 1822–1857.

Nardi, Y. and A. Rinaldo (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics 2*, 605–633.

Ning, B., S. Ghosal, and J. Thomas (2018). Bayesian method for causal inference in spatially-correlated multivariate time series. *Bayesian Analysis 14*, 1–28.

Pati, D., A. Bhattacharya, N. S. Pillai, and D. Dunson (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics 42*(3), 1102–1130.

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics 46*, 401–437.

Ročková, V. and E. Lesaffre (2014). Incorporating grouping information in Bayesian variable selection with applications in genomics. *Bayesian Analysis 9*, 221–258.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, Inc.

Song, Q. and F. Liang (2017). Nearly optimal Bayesian shrinkage for high-dimensional regression. *arXiv:1712.08964*.

Suarez, A. J. and S. Ghosal (2017). Bayesian estimation of principal components for functional data. *Bayesian Analysis 12*, 311–333.

Sun, D. and J. O. Berger (2007). Objective bayesian analysis for the multivariate normal model. *Bayesian Statistics 8 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press*, 525–562 (with discussion).

van der Vaart, A. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer.

Xu, X. and M. Ghosh (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis 10*, 909–936.

Yang, R. and J. O. Berger (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics 22*(3), 1195–1211.

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