Electronic Journal of Statistics
Vol. 15 (2021) 3040–3111
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1855

# Unified Bayesian theory of sparse linear regression with nuisance parameters[*]

**Seonghyun Jeong**

*Department of Statistics and Data Science, Department of Applied Statistics*
*Yonsei University, Seoul 03722, South Korea*
*e-mail:* sjeong@yonsei.ac.kr

**Subhashis Ghosal**

*Department of Statistics*
*North Carolina State University, Raleigh, NC 27607, USA*
*e-mail:* sghosal@ncsu.edu

**Abstract:** We study frequentist asymptotic properties of Bayesian procedures for high-dimensional Gaussian sparse regression when unknown nuisance parameters are involved. Nuisance parameters can be finite-, high-, or infinite-dimensional. A mixture of point masses at zero and continuous distributions is used for the prior distribution on sparse regression coefficients, and appropriate prior distributions are used for nuisance parameters. The optimal posterior contraction of sparse regression coefficients, hampered by the presence of nuisance parameters, is also examined and discussed. It is shown that the procedure yields strong model selection consistency. A Bernstein-von Mises-type theorem for sparse regression coefficients is also obtained for uncertainty quantification through credible sets with guaranteed frequentist coverage. Asymptotic properties of numerous examples are investigated using the theory developed in this study.

**MSC2020 subject classifications:** 62F15.
**Keywords and phrases:** Bernstein-von Mises theorems, High-dimensional regression, Model selection consistency, Posterior contraction rates, Sparse priors.

Received June 2020.

## Contents

[*] Research is partially supported by a Faculty Research and Professional Development Grant from College of Sciences of North Carolina State University and by ARO grant number 76643-MA.

## 1. Introduction

While Bayesian model selection for classical low-dimensional problems has a long history, sparse estimation in high-dimensional regression was studied much later; see Bondell and Reich [5], Johnson and Rossell [20], and Narisetty and He [24] for consistent Bayesian model selection methods in high-dimensional linear models. Extensive theoretical investigations, however, have been carried out only very recently. Since the pioneering work of Castillo et al. [8], frequentist asymptotic properties of Bayesian sparse regression have been discovered under various settings, and there is now a substantial body of literature [e.g., 23, 1, 28, 3, 26, 2, 10, 25, 14, 19, 18].

Most of the existing studies deal with sparse regression setups without nuisance parameters and there are only a few exceptions. An unknown variance parameter, the simplest type of nuisance parameters, was incorporated for high-dimensional linear regression in Song and Liang [28] and Bai et al. [2]. In these studies, the optimal properties of Bayesian procedures are characterized with continuous shrinkage priors. For more involved models, Chae et al. [10] adopted a nonparametric approach to estimate unknown symmetric densities in sparse linear regression. Ning et al. [25] considered a sparse linear model for vector-valued response variables with unknown covariance matrices.

Although nuisance parameters may not be of primary interest, modeling frameworks require the complete description of their roles as they explicitly parameterize models. Therefore, one may want to achieve optimal estimation properties for sparse regression coefficients, no matter what a nuisance parameter is. It may also be of interest to examine posterior contraction of nuisance parameters as a secondary objective. Despite these facts, however, there have not been attempts to consider a general class of high-dimensional regression models with nuisance parameters. In this study, we consider a general form of Gaussian sparse regression in the presence of nuisance parameters, and establish a theoretical framework for Bayesian procedures.

We formulate a general framework to treat sparse regression models in a unified way as follows. Let $\eta$ be possibly an infinite-dimensional nuisance parameter taking values in a set $\mathbb{H}$. For each $\eta \in \mathbb{H}$ and an integer $m_i \in \{1,\dots,\overline{m}\}$ for some $\overline{m} \geq 1$, suppose that there are a vector $\xi_{\eta,i} \in \mathbb{R}^{m_i}$ and a positive definite matrix $\Delta_{\eta,i} \in \mathbb{R}^{m_i \times m_i}$ which define a regression model for a vector-valued response variable $Y_i \in \mathbb{R}^{m_i}$ against covariates $X_i \in \mathbb{R}^{m_i \times p}$ given by

$$Y_i = X_i\theta + \xi_{\eta,i} + \varepsilon_i, \quad \varepsilon_i \overset{\text{ind}}{\sim} \mathrm{N}_{m_i}(0, \Delta_{\eta,i}), \quad i = 1,\dots,n, \tag{1}$$

where $\theta \in \mathbb{R}^p$ is a vector of regression coefficients. Here $m_i$ (and $\overline{m}$) can increase with $n$. We consider the high-dimensional situation where $p > n$, but $\theta$ is assumed to be sparse, with many coordinates zero. The form in (1) clearly includes sparse linear regression with unknown error variances. Our main interest lies in more complicated setups. As will be shortly discussed in Section 1.1, many interesting examples belong to form (1).

In this paper, we develop a unified theory of posterior asymptotics in the high-dimensional sparse regression models described by form (1). To the best of our knowledge, there is no study thus far considering a general modeling framework of sparse regression as in (1), even from the frequentist perspective. The results on complicated high-dimensional regression models are only available at model-specific levels and cannot be universally used for different model classes. On the other hand, our approach is a unified theoretical treatment of the general model structure in (1) under the Bayesian framework. We establish general theorems on nearly optimal posterior contraction rates, a Bernstein-von Mises theorem via shape approximation to the posterior distribution of $\theta$, and model selection consistency.

The general theory of posterior contraction using the canonical root-average-squared Hellinger metric on the joint density [16] is not very useful in this context, since to recover rates in terms of the metric of interest on the regression coefficients, some boundedness conditions are needed [19]. To deal with this issue, we construct an exponentially powerful likelihood ratio test in small pieces that are sufficiently separated from the true parameters in terms of the average Rényi divergence of order 1/2 (which coincides with the average negative log-affinity). This test provides posterior contraction relative to the corresponding divergence. The posterior contraction rates of $\theta$ and $\eta$ can then be recovered in terms of the metrics of interest under mild conditions on the parameter space. Due to a nuisance parameter $\eta$, the resulting posterior contraction for $\theta$ may be suboptimal. Conditions for the optimal posterior contraction will also be examined. Our results show that the obtained posterior contraction rates are adaptive to the unknown sparsity level.

For a Bernstein-von Mises theorem and selection consistency, stronger conditions are required than those used for posterior contraction, in line with the existing literature [e.g., 8, 23]. As pointed out by Chae et al. [10], the Bernstein-von Mises theorems for finite dimensional parameters in classical semiparametric models [e.g., 7] may not be directly useful in the high-dimensional context. We thus directly characterize a version of the Bernstein von-Mises theorem for model (1). The key idea is to find a suitable orthogonal projection that satisfies some required conditions, which is typically straightforward if the support of a prior for $\xi_{\eta,i}$ is a linear space. The complexity of the space of covariance matrices, measured by its metric entropy, also has an important role in deriving the Bernstein-von Mises theorem and selection consistency. Combining these two leads to a single component of normal distributions for an approximation, which enables to correctly quantify remaining uncertainty on the parameter through the posterior distribution.

### *1.1. Sparse linear regression with nuisance parameters*

As briefly discussed above, the form in (1) is general and includes many interesting statistical models. Here we provide specific examples belonging to (1) in detail. In Section 5, these examples will be used to apply the main results developed in this study.

**Example 1** (Multiple response models with missing components)**.** We consider a general multiple response model with missing values, which is very common in practice. Suppose that for each $i$, a vector of $\overline{m}$ responses with covariance matrix $\Sigma$ are supposed to be observed, but for the $i$th group (or subject) only $m_i$ entries are actually observed with the rest missing. Letting $Y_i \in \mathbb{R}^{m_i}$ be the $i$th observation and $Y_i^{\text{aug}} \in \mathbb{R}^{\overline{m}}$ be the augmented vector of $Y_i$ and missing entries, we can write $Y_i = E_i^T Y_i^{\text{aug}}$ and $\text{Cov}(Y_i) = E_i^T \Sigma E_i$, where $E_i \in \mathbb{R}^{\overline{m} \times m_i}$ is the submatrix of the $\overline{m} \times \overline{m}$ identity matrix with the $j$th column included if the $j$th element of $Y_i^{\text{aug}}$ is observed, $j = 1, \dots, \overline{m}$. Assuming that the mean of $Y_i$ is only $X_i\theta$ for covariates $X_i \in \mathbb{R}^{m_i \times p}$ and sparse coefficients $\theta \in \mathbb{R}^p$ with $p > n$, the model of interest can be written as $Y_i = X_i\theta + \varepsilon_i$, $\varepsilon_i \overset{\text{ind}}{\sim} \text{N}_{m_i}(0, E_i^T \Sigma E_i)$, $i = 1, \dots, n$. The model belongs to the class described by (1) with $\xi_{\eta,i} = 0_{m_i}$ and $\Delta_{\eta,i} = E_i^T \Sigma E_i$ for $\eta = \Sigma$.

**Example 2** (Multivariate measurement error models)**.** Suppose that a scalar response variable $Y_i^* \in \mathbb{R}$ is connected to fixed covariates $X_i^* \in \mathbb{R}^p$ with $p > n$ and random covariates $Z_i \in \mathbb{R}^q$ with fixed $q \geq 1$, through the following linear additive relationship: $Y_i^* = \alpha + X_i^{*T}\theta + Z_i^T\beta + \varepsilon_i^*$, $Z_i \overset{\text{iid}}{\sim} \text{N}_q(\mu, \Sigma)$, $\varepsilon_i^* \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,

$i = 1, \dots, n$. While $X_i^*$ is fully observed without noise, we observe a surrogate $W_i$ of $Z_i$ as $W_i = Z_i + \tau_i$, $\tau_i \overset{\text{iid}}{\sim} \mathrm{N}_q(0, \Psi)$, where to ensure identifiability, $\Psi$ is assumed to be known. This type of model is called a measurement error model or an errors-in-variables model; see Fuller [13] and Carroll et al. [6] for a complete overview. By direct calculations, the joint distribution of $(Y_i^*, W_i)$ is given by

$$\begin{pmatrix} Y_i^* \\ W_i \end{pmatrix} \overset{\text{ind}}{\sim} \mathrm{N}_{q+1}\left( \begin{pmatrix} \alpha + X_i^{*T}\theta + \mu^T\beta \\ \mu \end{pmatrix}, \begin{pmatrix} \beta^T\Sigma\beta + \sigma^2 & \beta^T\Sigma \\ \Sigma\beta & \Sigma + \Psi \end{pmatrix} \right).$$

By writing $Y_i = (Y_i^*, W_i^T)^T \in \mathbb{R}^{q+1}$, $X_i = (X_i^*, 0_{p\times q})^T \in \mathbb{R}^{(q+1)\times p}$, $\xi_{\eta,i} = (\alpha + \mu^T\beta, \mu^T)^T \in \mathbb{R}^{q+1}$, and $\Delta_{\eta,i} = \left(\begin{smallmatrix} \beta^T\Sigma\beta+\sigma^2 & \beta^T\Sigma \\ \Sigma\beta & \Sigma+\Psi \end{smallmatrix}\right) \in \mathbb{R}^{(q+1)\times(q+1)}$ with $\eta = (\alpha, \beta, \mu, \sigma^2, \Sigma)$, the model is of form (1) with $m_i = q + 1$.

**Example 3** (Parametric correlation structure)**.** For $m_i \geq 1$, $i = 1, \dots, n$, suppose that we have a response variable $Y_i \in \mathbb{R}^{m_i}$ and covariates $X_i \in \mathbb{R}^{m_i \times p}$ with $p > n$. We consider a standard regression model given by $Y_i = X_i\theta + \varepsilon_i$, $\varepsilon_i \overset{\text{ind}}{\sim} \mathrm{N}_{m_i}(0, \Sigma_i)$, $i = 1, \dots, n$, but $m_i$ is considered to be possibly increasing. For a known parametric correlation structure $G_i$ and a fixed dimensional Euclidean parameter $\alpha$, we model the covariance matrix as $\Sigma_i = \sigma^2 G_i(\alpha)$ using a variance parameter $\sigma^2$ and a correlation matrix $G_i(\alpha) \in \mathbb{R}^{m_i \times m_i}$. Examples of $G_i$ include first order autoregressive and moving average correlation matrices. The model belongs to (1) by writing $\xi_{\eta,i} = 0_{m_i}$ and $\Delta_{\eta,i} = \sigma^2 G_i(\alpha)$ with $\eta = (\alpha, \sigma^2)$.

**Example 4** (Mixed effects models)**.** For $m_i \geq 1$, $i = 1, \dots, n$, consider a response variable $Y_i \in \mathbb{R}^{m_i}$ and covariates $X_i \in \mathbb{R}^{m_i \times p}$ with $p > n$ and $Z_i \in \mathbb{R}^{m_i \times q}$ with fixed $q \geq 1$. A mixed effect model given by $Y_i = X_i\theta + Z_i b_i + \varepsilon_i^*$, $b_i \overset{\text{iid}}{\sim} \mathrm{N}_q(0, \Psi)$, $\varepsilon_i^* \overset{\text{ind}}{\sim} \mathrm{N}_{m_i}(0, \sigma^2 I_{m_i})$, $i = 1, \dots, n$, where $\Psi \in \mathbb{R}^{q\times q}$ is a positive definite matrix. Then the marginal law of $Y_i$ is given by $Y_i = X_i\theta + \varepsilon_i$, $\varepsilon_i \overset{\text{ind}}{\sim} \mathrm{N}_{m_i}(0, \sigma^2 I_{m_i} + Z_i\Psi Z_i^T)$. We assume that $\sigma^2$ is known. The model belongs to (1) by letting $\xi_{\eta,i} = 0_{m_i}$ and $\Delta_{\eta,i} = \sigma^2 I_{m_i} + Z_i\Psi Z_i^T$ with $\eta = \Psi$.

**Example 5** (Graphical structure with a sparse precision matrix)**.** For a response variable $Y_i \in \mathbb{R}^{\overline{m}}$ and covariates $X_i \in \mathbb{R}^{\overline{m}\times p}$ with increasing $\overline{m} \geq 1$ and $p > n$, consider a model given by $Y_i = X_i\theta + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}_{\overline{m}}(0, \Omega^{-1})$, $i = 1, \dots, n$, where $\theta$ is a sparse coefficient vector and the precision matrix $\Omega \in \mathbb{R}^{\overline{m}\times\overline{m}}$ is a positive definite matrix. Along with $\theta$, we also impose sparsity on the off-diagonal entries of $\Omega$, which accounts for a graphical structure between observations. More precisely, if an off-diagonal entry is zero, it implies the conditional independence between the two concerned entries of $\varepsilon_i$ given the remaining ones, and we suppose that most off-diagonal entries are actually zero, even though we do not know their locations. The model is then seen to be a special case of (1) by letting $\xi_{\eta,i} = 0_{\overline{m}}$ and $\Delta_{\eta,i} = \Omega^{-1}$ with $\eta = \Omega$.

**Example 6** (Nonparametric heteroskedastic regression models)**.** For a response variable $Y_i \in \mathbb{R}$ and a row vector of covariates $X_i \in \mathbb{R}^{1\times p}$, a linear regression model with a nonparametric heteroskedastic error is given by $Y_i = X_i\theta + \varepsilon_i$, $\varepsilon_i \overset{\text{ind}}{\sim} \mathrm{N}(0, v(z_i))$, $i = 1, \dots, n$, where $\theta$ is a sparse coefficient vector, $v : [0,1] \mapsto (0, \infty)$

is a univariate variance function, and $z_i \in [0,1]$ is a one-dimensional variable associated with the $i$th observation that controls the variance of $Y_i$ through the variance function $v$.Then the model belongs to (1) by letting $\xi_{\eta,i} = 0$ and $\Delta_{\eta,i} = v(z_i)$ with $\eta = v$.

**Example 7** (Partial linear models)**.** Consider a partial linear model given by $Y_i = X_i\theta + g(z_i) + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma^2)$, $i = 1,\dots,n$, where $Y_i \in \mathbb{R}$ is a response variable, $X_i \in \mathbb{R}^{1\times p}$ is a row vector of covariates with $p > n$, $\theta \in \mathbb{R}^p$ is a sparse coefficient vector, $g : [0,1] \mapsto \mathbb{R}$ is a univariate function, and $z_i \in [0,1]$ is a scalar predictor. This model is expressed in form (1) by writing $\xi_{\eta,i} = g(z_i)$ and $\Delta_{\eta,i} = \sigma^2$ with $\eta = (g,\sigma^2)$.

### *1.2. Outline*

The rest of this paper is organized as follows. In Section 2, some notations are introduced and a prior distribution on sparse regression coefficients is specified. Sections 3–4 provide our main results on the posterior contraction, the Bernstein-von Mises phenomenon, and selection consistency of the posterior distribution. In Section 5, our general theorems are applied to the examples considered above to derive the posterior asymptotic properties in each specific example. All technical proofs are provided in Appendix.

## 2. Setup, notations, and prior specification

### *2.1. Notation*

Here we describe the notations we use throughout this paper. For a vector $\theta = (\theta_j) \in \mathbb{R}^p$ and a set $S \subset \{1,\dots,p\}$ of indices, we write $S_\theta = \{j : \theta_j \neq 0\}$ to denote the support of $\theta$, $s := |S|$ (or $s_\theta := |S_\theta|$) to denote the cardinality of $S$ (or $S_\theta$), and $\theta_S = \{\theta_j : j \in S\}$ and $\theta_{S^c} = \{\theta_j : j \notin S\}$ to separate components of $\theta$ using $S$. In particular, the support of the true parameter $\theta_0$ and its cardinality are written as $S_0$ and $s_0 := |S_0|$, respectively. The notation $\|\theta\|_q = (\sum_j |\theta_j|^q)^{1/q}$, $1 \le q < \infty$, stands for the $\ell_q$-norm and $\|\theta\|_\infty = \max_j |\theta_j|$ denotes the maximum norm. We write $\rho_{\min}(A)$ and $\rho_{\max}(A)$ for the minimum and maximum eigenvalues of a square matrix $A$, respectively. For a matrix $X = (\!(x_{ij})\!)$, let $\|X\|_{\text{sp}} = \rho_{\max}^{1/2}(X^TX)$ stand for the spectral norm and $\|X\|_{\text{F}} = (\sum_{i,j} x_{ij}^2)^{1/2}$ stand for the Frobenius norm of $X$. We also define a matrix norm $\|X\|_* = \max_j \|X_{\cdot j}\|_2$ for $X_{\cdot j}$ the $j$th column of $X$, which is used for compatibility conditions. The column space of $X$ is denoted by $\text{span}(X)$. For further convenience, we write $\varsigma_{\min}(X) = \rho_{\min}^{1/2}(X^TX)$ for the minimum singular value of $X$. The notation $X_S$ means the submatrix of $X$ with columns chosen by $S$. For sequences $a_n$ and $b_n$, $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) stands for $a_n \le Cb_n$ for some constant $C > 0$ independent of $n$, and $a_n \asymp b_n$ means $a_n \lesssim b_n \lesssim a_n$. These inequalities are also used for relations involving constant sequences.

For given parameters $\theta$ and $\eta$, we write the joint density as $p_{\theta,\eta} = \prod_{i=1}^n p_{\theta,\eta,i}$ for $p_{\theta,\eta,i}$ the density of the $i$th observation vector $Y_i$. In particular, the true joint density is expressed as $p_0 = \prod_{i=1}^n p_{0,i}$ for $p_{0,i} := p_{\theta_0,\eta_0,i}$ with the true parameters $\theta_0$ and $\eta_0$. The notation $\mathbb{E}_0$ denotes the expectation operator with the true density $p_0$. For two probability measures $P$ and $Q$, let $\|P-Q\|_{\mathrm{TV}}$ denote the total variation between $P$ and $Q$. For two $n$-variate densities $f := \prod_{i=1}^n f_i$ and $g := \prod_{i=1}^n g_i$ of independent variables, denote the average Rényi divergence (of order 1/2) by $R_n(f,g) = -n^{-1}\sum_{i=1}^n \log \int \sqrt{f_i g_i}$.

For any $\eta_1, \eta_2 \in \mathbb{H}$, we define $d_n^2(\eta_1,\eta_2) = d_{A,n}^2(\eta_1,\eta_2) + d_{B,n}^2(\eta_1,\eta_2)$ for the two squared pseudo-metrics:

$$d_{A,n}^2(\eta_1,\eta_2) = \frac{1}{n}\sum_{i=1}^n \|\xi_{\eta_1,i} - \xi_{\eta_2,i}\|_2^2, \quad d_{B,n}^2(\eta_1,\eta_2) = \frac{1}{n}\sum_{i=1}^n \|\Delta_{\eta_1,i} - \Delta_{\eta_2,i}\|_{\mathrm{F}}^2.$$

For compatibility conditions, the uniform compatibility number $\phi_1$ and the smallest scaled singular value $\phi_2$ are defined as

$$\phi_1(s) = \inf_{\theta:1\le|S_\theta|\le s} \frac{\|X\theta\|_2 |S_\theta|^{1/2}}{\|X\|_* \|\theta\|_1}, \quad \phi_2(s) = \inf_{\theta:1\le|S_\theta|\le s} \frac{\|X\theta\|_2}{\|X\|_* \|\theta\|_2}.$$

We write $Y^{(n)} = (Y_1^T, \dots, Y_n^T)^T$ for the observation vector, $n_* = \sum_{i=1}^n m_i$ for the dimension of $Y^{(n)}$, and $\Theta = \mathbb{R}^p$ for the parameter space of $\theta$. Lastly, for a (pseudo-)metric space $(\mathcal{F}, d)$, let $N(\epsilon, \mathcal{F}, d)$ denote the $\epsilon$-covering number, the minimal number of $\epsilon$-balls that cover $\mathcal{F}$.

### 2.2. *Prior for the high-dimensional coefficients*

In this subsection, we specify a prior distribution for the high-dimensional regression coefficients $\theta$. A prior for $\eta$ should satisfy the conditions required for the main results, so its specific characterization is deferred to Section 3. On the other hand, the prior for $\theta$ specified here is always good for our purposes and satisfies all requirements.

We first select a dimension $s$ from a prior $\pi_p$, and then randomly choose $S \subset \{1,\dots,p\}$ for given $s$. A nonzero part $\theta_S$ of $\theta$ is then selected from a prior $g_S$ on $\mathbb{R}^s$ while $\theta_{S^c}$ is fixed to zero. The resulting prior specification for $(S,\theta)$ is formulated as

$$(S,\theta) \mapsto \frac{\pi_p(s)}{\binom{p}{s}} g_S(\theta_S)\delta_0(\theta_{S^c}), \tag{2}$$

where $\delta_0$ is the Dirac measure at zero on $\mathbb{R}^{p-s}$ with suppressed dimensionality. For the prior $\pi_p$ on the model dimensions, we consider a prior satisfying the following: for some constants $A_1, A_2, A_3, A_4 > 0$,

$$A_1 p^{-A_3}\pi_p(s-1) \le \pi_p(s) \le A_2 p^{-A_4}\pi_p(s-1), \quad s = 1,\dots,p. \tag{3}$$

Examples of priors satisfying (3) can be found in Castillo and van der Vaart [9] and Castillo et al. [8]. For the prior $g_S$, the $s$-fold product of the exponential power density is considered, where the regularization parameter is allowed to vary with $p$ and $\|X\|_*$, i.e.,

$$g_S(\theta_S) = \prod_{j\in S} \frac{\lambda}{2} \exp\left(-\lambda|\theta_j|\right), \quad \frac{\|X\|_*}{L_1 p^{L_2}} \le \lambda \le \frac{L_3\|X\|_*}{\sqrt{n}}, \tag{4}$$

for some constants $L_1, L_2, L_3 > 0$. The order of $\lambda$ is important in that it determines the boundedness requirement of the true signal $\theta_0$ (see condition (C3) below). A particularly interesting case is obtained when $\lambda$ is set to the lower bound $\|X\|_*/(L_1 p^{L_2})$. Then the boundedness condition becomes very mild by choosing $L_2$ sufficiently large. When $\lambda$ is set to the upper bound, the boundedness condition is still reasonably mild. However, it can actually be relaxed if the true signal is known to be small enough, though we do not pursue this generalization in this study. In Section 4, we shall see that values of $\lambda$ that do not increase too fast are in fact necessary for a distributional approximation and selection consistency.

**Remark 1.** Since some size restriction on $\theta_0$ will be made unlike Castillo et al. [8], we note that the use of the Laplace density is not necessary and other prior distributions may also be used for $\theta$. For example, normal densities can be used for $g_S$ to exploit semi-conjugacy. However, if its precision parameter is fixed independent of $n$, a normal prior requires a stronger restriction on the true signal than (C3) below. To achieve the nearly optimal posterior contraction, other densities with similar tail properties should also work with appropriate modifications for the true signal size (see, e.g., Jeong and Ghosal [19]). Instead of the spike-and-slab prior in (2) and (3), a class of continuous shrinkage priors may also be used at the expense of substantial modifications in the technical details [28]. In this paper, we only consider the prior in (2)–(4).

## 3. Posterior contraction rates

The prior for a nuisance parameter $\eta$ should be chosen to complete the prior specification. Once we assign the prior for the full parameters, the posterior distribution $\Pi(\cdot \mid Y^{(n)})$ is defined by Bayes' rule. How the prior for $\eta$ is chosen is crucial to obtain desirable asymptotic properties of the posterior distribution. In this subsection, we shall examine such conditions on the prior distribution for a nuisance parameter and study the posterior contraction rates for both $\theta$ and $\eta$.

The prior for $\eta$ is put on a subspace $\mathcal{H} \subset \mathbb{H}$. In many instances, we take $\mathcal{H} = \mathbb{H}$, especially when a nuisance parameter is finite dimensional, but the flexibility of a subspace may be beneficial in infinite-dimensional situations. We need to choose $\mathcal{H}$ to satisfy certain conditions.

(C1) There exists a nondecreasing sequence $a_n = o(n)$ such that

$$a_n \max_{1\le i\le n} \|\Delta_{\eta',i} - \Delta_{\eta_0,i}\|_{\mathrm{F}}^2 =: e_n \to 0, \quad \text{for some } \eta' \in \mathcal{H},$$

$$\max_{1\le i\le n}\|\Delta_{\eta_1,i}-\Delta_{\eta_2,i}\|_{\rm F}^2\le a_n d_{B,n}^2(\eta_1,\eta_2),\quad \eta_1,\eta_2\in\mathcal{H}.$$

(C2) For some sequence $\bar\epsilon_n$ such that $a_n\bar\epsilon_n^2\to 0$ and $n\bar\epsilon_n^2\to\infty$ with $a_n$ satisfying (C1),

$$\log\Pi\left(\eta\in\mathcal{H}:d_n(\eta,\eta_0)\le\bar\epsilon_n\right)\gtrsim -n\bar\epsilon_n^2.$$

The first condition of (C1) implies that we have a good approximation to the true parameter value in the parameter set $\mathcal{H}$. This holds trivially if there exists $\eta'\in\mathcal{H}$ such that $\Delta_{\eta',i}=\Delta_{\eta_0,i}$ for every $i\le n$, which is obviously true if $\eta_0\in\mathcal{H}$. The second condition of (C1) means that in $\mathcal{H}$, the maximum Frobenius norm of the difference between covariance matrices can be controlled by the average Frobenius norm multiplied by the sequence $a_n$. Clearly, this holds with $a_n=1$ if $\Delta_{\eta,i}$ is the same for every $i\le n$. By the triangle inequality, we see that (C1) implies that

$$\max_{1\le i\le n}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\rm F}^2\lesssim e_n+a_n d_{B,n}^2(\eta,\eta_0),\quad \eta\in\mathcal{H},\tag{5}$$

which is used throughout the paper. Condition (C2) is typically called the prior concentration condition, which requires a prior to put sufficient mass around the true parameter $\eta_0$, measured by the pseudo-metric $d_n$. As in other infinite-dimensional situations, such a closeness is translated into the closeness in terms of the Kullback-Leibler divergence and variation (see Lemma 1 in Appendix for more details).

As noted in Section 1, the true parameters should be restricted to certain norm-bounded subset of the parameter space. This is clarified as follows.

(C3) The true signal satisfies $\|\theta_0\|_\infty\lesssim\lambda^{-1}\log p$.

(C4) The eigenvalues of the true covariance matrix satisfy

$$1\lesssim\min_{1\le i\le n}\rho_{\min}(\Delta_{\eta_0,i})\le\max_{1\le i\le n}\rho_{\max}(\Delta_{\eta_0,i})\lesssim 1.$$

Condition (C3) is required to apply the general strategy for posterior contraction to our modeling framework containing nuisance parameters. More specifically, the condition is imposed such that the prior assigns sufficient mass on a Kullback-Leibler neighborhood of $\theta_0$. If nuisance parameters are not present, one can directly handle the model and such a restriction may be removed [e.g., 8, 14]. One may refer to Song and Liang [28], Ning et al. [25], and Bai et al. [2] for conditions similar to ours, where a variance parameter stands for a nuisance parameter. Still, the condition is mild if $\lambda$ is chosen to decrease at an appropriate order. In particular, if $\lambda$ is matched to the lower bound $1/(L_1p^{L_2})$, the condition becomes $\|\theta_0\|_\infty\lesssim(p^{L_2}\log p)/\|X\|_*$ which is very mild if $L_2$ is sufficiently large. Even if the upper bound $L_3\|X\|_*/\sqrt{n}$ is chosen, the condition is not restrictive as the right hand side of the condition can be made nondecreasing as long as $\|X\|_*$ is increasing at a suitable order. Condition (C4) implies that the eigenvalues of the true covariance matrix are bounded below and above. The lower and upper bounds are required for a lot of technical details, including the construction of an exponentially powerful test in Lemma 2 in Appendix.

**Remark 2.** Condition (C3) is actually stronger than what it needs to be, but is adopted for the ease of interpretation. For Theorem 3 below to hold, it suffices if we have $\lambda\|\theta_0\|_1 \le (s_0 \log p) \vee n\bar{\epsilon}_n^2$ for $\bar{\epsilon}_n$ satisfying (C2). For the optimal posterior contraction in Theorem 4 below, a slightly stronger bound is needed: $\lambda\|\theta_0\|_1 \le s_0 \log p$ (see Lemma 6 and its proof in Appendix).

### *3.1. Rényi posterior contraction and recovery*

The goal of this subsection is to study posterior contraction of $\theta$ relative to the $\ell_1$- and $\ell_2$-metrics. To do so, we derive the posterior contraction rate with respect to the average Rényi divergence $R_n(f,g)$, and then the rates for $\theta$ relative to more concrete metrics will be recovered from the Rényi contraction.

To proceed, we first need to examine a dimensionality property of the support of $\theta$. The following theorem shows that the posterior distribution is concentrated on models of relatively small sizes.

**Theorem 1** (Dimension)**.** *Suppose that* (C1)–(C4) *are satisfied. Then for* $s_\star :$ $= s_0 \vee (n\bar{\epsilon}_n^2/\log p)$*, there exists a constant* $K_1$ *such that*

$$\mathbb{E}_0\Pi\left(\theta : s_\theta > K_1 s_\star \,\middle|\, Y^{(n)}\right) \to 0.$$

Compared to the literature [e.g., 8, 23, 3], the rate in Theorem 1 is floored by the extra term $n\bar{\epsilon}_n^2/\log p$. This arises from the presence of a nuisance parameter in the model formulation. To minimize its impact, a prior on $\eta$ should be chosen such that (C2) holds for as small $\bar{\epsilon}_n$ as possible; a suitable choice induces the (nearly) optimal contraction rate.

Using the basic results in Theorem 1, the next theorem obtains the rate at which the posterior distribution contracts at the truth with respect to the average Rényi divergence. The theorem requires additional assumptions on a prior.

(C5) For $s_\star := s_0 \vee (n\bar{\epsilon}_n^2/\log p)$ with $\bar{\epsilon}_n$ satisfying (C2), a sufficiently large $B > 0$, and some sequences $\gamma_n$ and $\epsilon_n \ge \sqrt{s_\star \log(p \vee \overline{m} \vee \gamma_n)/n}$ satisfying $\epsilon_n^2/\overline{m} \to 0$, there exists a subset $\mathcal{H}_n \subset \mathcal{H}$ such that

$$\min_{1\le i\le n} \inf_{\eta\in\mathcal{H}_n} \rho_{\min}(\Delta_{\eta,i}) \ge \frac{1}{\gamma_n}, \tag{6}$$

$$\log N\left(\frac{1}{6\overline{m}\gamma_n n^{3/2}}, \mathcal{H}_n, d_n\right) \lesssim n\epsilon_n^2, \tag{7}$$

$$e^{Bs_\star \log p}\Pi(\mathcal{H} \setminus \mathcal{H}_n) \to 0. \tag{8}$$

The above conditions are related to the classical ones in the literature (e.g., see Theorem 2.1. of Ghosal et al. [15]). Condition (6) requires that for every $i \le n$, the minimum eigenvalue of $\Delta_{\eta,i}$ is not too small on a sieve $\mathcal{H}_n$. Although $\gamma_n$ can be any positive sequence, a sequence increasing exponentially fast makes the entropy in (7) too large, resulting in a suboptimal rate $\epsilon_n$. If $\gamma_n$ can be chosen to be smaller than $p$ and $\overline{m}$, then this does not lead to any deterioration of the rate in $\epsilon_n$. The entropy condition (7) is actually stronger than needed. Scrutinizing the proof of the theorem, one can see that the entropy appearing in the theorem is obtained using pieces that are smaller than those giving the exponentially powerful test in Lemma 2 in Appendix. However, the covering number with those pieces looks more complicated and the form in (7) suffices for all examples in the present paper. Lastly, condition (8) implies that the outside of a sieve $\mathcal{H}_n$ should possess sufficiently small prior mass to kill the factor $s_\star \log p$ arising from the lower bound of the denominator of the posterior distribution. In fact, conditions similar to (C2), (7) and (8) are also required for the prior of $\theta$. By reading the proof, it is easy to see that the prior (2) explicitly satisfies the analogous conditions on an appropriately chosen sieve.

**Theorem 2** (Contraction rate, Rényi)**.** *Suppose that* (C1)–(C5) *are satisfied. Then there exists a constant* $K_2$ *such that*

$$\mathbb{E}_0\Pi\left((\theta,\eta) : R_n(p_{\theta,\eta}, p_0) > K_2\epsilon_n^2 \,\middle|\, Y^{(n)}\right) \to 0.$$

We want to sharpen the rate $\epsilon_n \geq \sqrt{s_\star \log(p \vee \overline{m} \vee \gamma_n)/n}$ as much as possible. In most instances, $\gamma_n$ can be chosen such that $\log \gamma_n \lesssim \log p$. This is trivially satisfied if $\gamma_n$ is some polynomial in $n$ as in the examples in this paper. If $p$ is known to increase much faster than $n$, e.g., $\log p \asymp n^c$ for some $c \in (0,1)$, then $\gamma_n$ need not be a polynomial in $n$ and the condition can be met more easily with a sequence that grows even faster. Note also that we typically have $\log \overline{m} \lesssim \log p$ in most cases. These postulates lead to $\epsilon_n \geq \sqrt{(s_\star \log p)/n}$. Indeed, it is often possible to choose $\epsilon_n = \sqrt{(s_\star \log p)/n}$, which is commonly guaranteed by choosing an appropriate sieve $\mathcal{H}_n$ and a prior. The condition will be made precise in (C5*) below for recovery and we only consider the situation that $\epsilon_n = \sqrt{(s_\star \log p)/n}$ in what follows.

Although Theorem 2 provides the basic results for posterior contraction, it does not give precise interpretations for the parameters $\theta$ and $\eta$ themselves, because of the abstruse expression of the average Rényi divergence. The contraction rates with respect to more concrete metrics are recovered under some additional conditions. Under the additional assumption $a_n\epsilon_n^2 \to 0$, it can be shown that Theorem 1 and Theorem 2 explicitly imply that for the set

$$\mathcal{A}_n = \Bigg\{(\theta,\eta) \in \Theta \times \mathcal{H} : s_\theta \leq K_1 s_\star, \\ \frac{1}{n}\sum_{i=1}^n \|X_i(\theta-\theta_0) + \xi_{\eta,i} - \xi_{\eta_0,i}\|_2^2 + d_{B,n}^2(\eta,\eta_0) \leq M_1\epsilon_n^2\Bigg\},$$

with a sufficiently large constant $M_1$, the posterior mass of $\mathcal{A}_n$ goes to one in probability (see the proof of Theorem 3). To complete the recovery, we need to separate the sum of squares of the mean into $\|X(\theta-\theta_0)\|_2$ and $nd_{A,n}^2(\eta,\eta_0)$, which requires an additional condition. The conditions required for the recovery are clarified as follows.

(C5*) While $\log \overline{m} \lesssim \log p$, (C5) holds for $\gamma_n$ and $\epsilon_n = \sqrt{(s_\star \log p)/n}$ such that $\log \gamma_n \lesssim \log p$ and $a_n\epsilon_n^2 \to 0$ with $a_n$ satisfying (C1).

(C6) For $s_\star$ satisfying (C5*), there exists $\eta_* \in \mathbb{H}$ such that

$$\liminf_{n\geq 1} \inf_{(\theta,\eta)\in\mathcal{A}_n} \frac{\sum_{i=1}^n (\theta-\theta_0)^T X_i^T (\xi_{\eta,i} - \xi_{\eta_*,i})}{\|X(\theta-\theta_0)\|_2^2 + nd_{A,n}^2(\eta,\eta_*)} > -\frac{1}{2},$$

$$d_{A,n}(\eta_*,\eta_0) \lesssim \sqrt{\frac{s_\star \log p}{n}},$$

where $\epsilon_n$ in $\mathcal{A}_n$ satisfies $\epsilon_n = \sqrt{(s_\star \log p)/n}$.

By expanding the quadratic term for the mean in $\mathcal{A}_n$, one can see that the separation is possible if (C6) is satisfied. Clearly, (C6) is trivially satisfied if the model has only $X\theta$ for its mean, in which we take $\xi_{\eta,i} - \xi_{\eta_*,i} = \xi_{\eta_*,i} - \xi_{\eta_0,i} = 0$ for every $i \leq n$. In many cases where there exists $\eta' \in \mathcal{H}$ such that $d_{A,n}(\eta',\eta_0) = 0$, we can often take $\eta_* = \eta'$ for the second inequality of (C6) to hold automatically.

The following theorem shows that the posterior distribution of $\theta$ and $\eta$ contracts at their respective true values at some rates, relative to more easily comprehensible metrics than the average Rényi divergence. In the expressions, if $K_1 s_\star + s_0 < 1$, the compatibility numbers should be understood be equal to 1 for interpretation.

**Theorem 3** (Recovery)**.** *Suppose that* (C1)–(C4), (C5*)*, and* (C6) *are satisfied. Then, there exists a constant* $K_3$ *such that*

$$\begin{aligned}
\mathbb{E}_0\Pi\left(\theta : \|\theta-\theta_0\|_1 > \frac{K_3 s_\star \sqrt{\log p}}{\phi_1(K_1 s_\star + s_0)\|X\|_*} \,\middle|\, Y^{(n)}\right) &\to 0,\\
\mathbb{E}_0\Pi\left(\theta : \|\theta-\theta_0\|_2 > \frac{K_3 \sqrt{s_\star \log p}}{\phi_2(K_1 s_\star + s_0)\|X\|_*} \,\middle|\, Y^{(n)}\right) &\to 0,\\
\mathbb{E}_0\Pi\left(\theta : \|X(\theta-\theta_0)\|_2 > K_3\sqrt{s_\star \log p} \,\middle|\, Y^{(n)}\right) &\to 0,\\
\mathbb{E}_0\Pi\left(\eta : d_n(\eta,\eta_0) > K_3\sqrt{\frac{s_\star \log p}{n}} \,\middle|\, Y^{(n)}\right) &\to 0.
\end{aligned} \tag{9}$$

The thresholds for contraction depend upon the compatibility conditions, which make their implication somewhat vague. As $K_1 s_\star + s_0$ is much smaller than $n_*$, it is not unreasonable to assume that $\phi_1(K_1 s_\star + s_0)$ and $\phi_2(K_1 s_\star + s_0)$ are bounded away from zero, whence the compatibility number is removed from the rates. We refer to Example 7 of Castillo et al. [8] for more discussion. In the next subsection, we will see that one of these restrictions is actually necessary for shape approximation or selection consistency.

**Remark 3.** The separation condition (C6) can be left as an assumption to be satisfied, but can also be verified by a stronger condition on the design matrix without resorting to the values of the parameters. Suppose that for some integer $q \geq 1$, there exists a matrix $Z_i \in \mathbb{R}^{m_i \times q}$ such that $\xi_{\eta,i} = Z_i h(\eta)$ for every $\eta \in \mathcal{H}$, with some map $h : \mathcal{H} \mapsto \mathbb{R}^q$. Since we can write $\xi_{\eta,i} - \xi_{\eta_*,i} = Z_i(h(\eta) - h(\eta_*))$ for any $\eta, \eta_* \in \mathcal{H}$, the Cauchy-Schwarz inequality indicates that the first inequality of (C6) is implied by

$$\liminf_{n \geq 1} \inf_{(\theta,\eta) \in \Theta \times \mathcal{H}: s_\theta \leq K_1 s_\star} \frac{(\theta - \theta_0)^T X^T Z(h(\eta) - h(\eta_*))}{\|X(\theta - \theta_0)\|_2 \|Z(h(\eta) - h(\eta_*))\|_2} > -1,$$

for $Z = (Z_1^T, \ldots, Z_n^T)^T$. The left hand side is always between $-1$ and 1 by the Cauchy-Schwarz inequality, and is exactly equal to $-1$ or 1 if and only if the two vectors are linearly dependent. A sufficient condition for the preceding display is thus $\min\{\varsigma_{\min}([X_S, Z]) : s \leq K_1 s_\star + s_0\} \gtrsim 1$ since the linear dependence cannot happen under such a condition due to the inequality $s_{\theta-\theta_0} \leq s_\theta + s_0 \leq K_1 s_\star + s_0$ for $\theta$ such that $s_\theta \leq K_1 s_\star$. This sufficient condition is not restrictive at all if $q = o(n)$ as we already have $K_1 s_\star + s_0 = o(n)$. Since there typically exists $\eta_* \in \mathcal{H}$ satisfying the second inequality of (C6) as long as $\mathcal{H}$ provides a good approximation for the true parameter $\eta_0$, condition (C6) can be easily satisfied if the sufficient condition is met.

Notwithstanding the lack of formal study of minimax rates with additional complications, we still want to match our rates for $\theta$ with those in simple linear regression, which we call the "optimal" rates. In this sense, Theorem 3 only provides the suboptimal rates for $\theta$ if $s_0 = o(s_\star)$. Although the theorem gives the optimal results if $s_0 \log p \gtrsim n\bar{\epsilon}_n^2$, it is practically hard to check this condition as $s_0$ is unknown. If $s_0$ is known to be nonzero, the desired conclusion is trivially achieved as soon as $n\bar{\epsilon}_n^2/\log p \lesssim 1$. The following corollary, however, shows that the optimal rates are still available even if $s_0 = 0$, with restrictions on $\bar{\epsilon}_n$ and the prior.

**Corollary 1** (Optimality under restriction). *For $\bar{\epsilon}_n$ satisfying the conditions for Theorem 3, we have the following assertions.*

(a) *Assume that $n\bar{\epsilon}_n^2/\log p \to 0$. Then, Theorems 1 and 3 hold for $s_\star$ replaced by $s_0$.*

(b) *Assume that $n\bar{\epsilon}_n^2/\log p \lesssim 1$. Then, Theorems 1 and 3 hold for $s_\star$ replaced by $s_0$ if either $A_4$ in (3) is chosen large enough or $s_0 > 0$.*

The corollary is useful in limited situations, especially when a parametric rate is available for a nuisance parameter. Even if $n\bar{\epsilon}_n^2 = \log n$, we need to further assume that $\log n = o(\log p)$, i.e., the ultra high-dimensional setup, to conclude that (a) holds, while we can always apply (b) because $\log n \lesssim \log p$. Although assertion (b) holds for any $s_0 \geq 0$ if $A_4$ is chosen sufficiently large, its specific threshold is not directly available. Indeed, by carefully reading the proof of Theorem 1 together with Lemma 1 in Appendix, one can see that the threshold depends on unknown constant bounds for the eigenvalues of the true covariance matrix in (C4). Still, (b) holds for any $A_4 > 0$ if $s_0 > 0$. We believe that the assumption $s_0 > 0$ is very mild, and hence simply apply

(b) with this assumption to conclude the optimal contraction for models with finite dimensional nuisance parameters. The optimal rates can still be achieved for any $s_0 \geq 0$ by verifying the conditions in the following subsection. With finite dimensional nuisance parameters, we do not pursue this direction as it seems an overkill considering the mildness of the assumption $s_0 > 0$, though those conditions are actually required for the Bernstein-von Mises theorem and selection consistency in Section 4.

In semiparametric situations with high- or infinite-dimensional nuisance parameters, none of (a) and (b) generally works unless $p$ increases sufficiently fast. Still, the optimal rates can be achieved under stronger conditions using the semiparametric theory, as the following subsection provides.

### *3.2. Optimal posterior contraction for θ*

Recall that only suboptimal rates may be available from Theorem 3 if $s_0 \log p \lesssim n\bar{\epsilon}_n^2$. In many semiparametric situations, however, it is often possible to obtain parametric rates for finite dimensional parameters under stronger conditions, even when there are infinite-dimensional nuisance parameters in a model [4, 7]. It has also been shown that a similar argument holds in some high-dimensional semiparametric regression models [10]. Therefore, it is naturally of interest to examine under what conditions we can replace $s_\star$ by $s_0$ in the rates for $\theta$, even if $s_0 \log p \lesssim n\bar{\epsilon}_n^2$. Similar to other semiparametric settings [4, 10], this can be established by the semiparametric theory, but requires stronger conditions than those in traditional fixed dimensional parametric cases because of the high-dimensions of the parameters in our setup.

To proceed, some additional conditions are required for technical reasons, which are made for the size of $\bar{\epsilon}_n$ as the optimal rates are automatically attained if $s_0 \log p \gtrsim n\bar{\epsilon}_n^2$. Still, in a practical sense, the conditions almost always need to be verified to reach the optimal rates, since only oracle rates are generally available and we do not know which term is greater.

In what follows, we write $\bar{s}_\star := n\bar{\epsilon}_n^2/\log p$ for $\bar{\epsilon}_n$ satisfying the conditions of Theorem 3 through the definition of $\epsilon_n$. We first assume the following condition on the uniform compatibility number.

(C7) For a sufficiently large $M$, the uniform compatibility number $\phi_1(M\bar{s}_\star + s_0)$ is bounded away from zero.

This condition is weaker than assuming that the smallest scaled singular value $\phi_2(M\bar{s}_\star + s_0)$ is bounded away from zero, as we have $\phi_1(s) \geq \phi_2(s)$ for any $s > 0$ by the Cauchy-Schwarz inequality. We will also resort on a slightly stronger condition with respect to $\phi_1$ for a distributional approximation in the following section. In this sense, our condition is weaker than those for Theorem 4 of Castillo et al. [8]. Condition (C7) is not restrictive as (C5*) requires $s_\star = o(n)$; we again refer to Example 7 of Castillo et al. [8].

To precisely describe other conditions, hereafter we use the following additional notations. We write

$$\tilde{X} = \begin{pmatrix} \Delta_{\eta_0,1}^{-1/2} X_1 \\ \vdots \\ \Delta_{\eta_0,n}^{-1/2} X_n \end{pmatrix} \in \mathbb{R}^{n_* \times p}, \quad \tilde{\xi}_\eta = \begin{pmatrix} \Delta_{\eta_0,1}^{-1/2} \xi_{\eta,1} \\ \vdots \\ \Delta_{\eta_0,n}^{-1/2} \xi_{\eta,n} \end{pmatrix} \in \mathbb{R}^{n_*},$$

and $\tilde{\Delta}_\eta$ to denote the collection of $\Delta_{\eta,i}$ for $i = 1, \dots, n$. In particular, $\tilde{X}_S \in \mathbb{R}^{n_* \times |S|}$ denotes the submatrix of $\tilde{X}$ with columns chosen by an index set $S$. We also define the following neighborhoods of the true parameters: for $\bar{s}_\star$ and $\bar{\epsilon}_n$ satisfying (C5*), and sufficiently large constants $\tilde{M}_1$ and $\tilde{M}_2$,

$$\begin{aligned} \widetilde{\Theta}_n &= \left\{ \theta \in \Theta : s_\theta \le K_1 \bar{s}_\star, \ \|X(\theta - \theta_0)\|_2 \le \tilde{M}_1 \sqrt{n} \bar{\epsilon}_n \right\}, \\ \widetilde{\mathcal{H}}_n &= \left\{ \eta \in \mathcal{H} : d_n(\eta, \eta_0) \le \tilde{M}_2 \bar{\epsilon}_n \right\}. \end{aligned} \tag{10}$$

Combined by other conditions, Theorem 3 implies that the posterior probabilities of these neighborhoods tend to one in probability if $s_0 \log p \lesssim n\bar{\epsilon}_n^2$. We need some bounding conditions on these neighborhoods, which will be specified below.

Let $\Phi(\eta) = (\tilde{\xi}_\eta, \tilde{\Delta}_\eta)$ for any given $\eta \in \mathcal{H}$. For a given $\theta$, we choose a bijective map $\eta \mapsto \tilde{\eta}_n(\theta, \eta) : \mathcal{H} \mapsto \mathcal{H}$ such that $\Phi(\tilde{\eta}_n(\theta, \eta)) = (\tilde{\xi}_\eta + H\tilde{X}(\theta - \theta_0), \tilde{\Delta}_\eta)$ for some orthogonal projection $H$ which may depend on the true parameter values, but not on $\theta$ and $\eta$. The projection $H$ plays a key role here and for a distributional approximation in the following section, and thus should be appropriately chosen to satisfy the followings.

(C8) The orthogonal projection $H$ satisfies

$$\frac{1}{(s_0 \vee 1) \log p} \sup_{\eta \in \widetilde{\mathcal{H}}_n} \|(I - H)(\tilde{\xi}_\eta - \tilde{\xi}_{\eta_0})\|_2^2 \to 0,$$

$$\min_{S: s \le K_1 \bar{s}_\star} \inf_{v \in \mathbb{R}^s : \|v\|_2 = 1} \frac{\|(I - H)\tilde{X}_S v\|_2}{\|\tilde{X}_S v\|_2} \gtrsim 1.$$

(C9) The conditional law $\Pi_{n,\theta}$ of $\tilde{\eta}_n(\theta, \eta)$ given $\theta$, induced by the prior, is absolutely continuous relative to its distribution $\Pi_{n,\theta_0}$ at $\theta = \theta_0$ (which is the same as the prior for $\eta$), and the Radon-Nikodym derivative $d\Pi_{n,\theta}/d\Pi_{n,\theta_0}$ satisfies

$$\sup_{\theta \in \widetilde{\Theta}_n} \sup_{\eta \in \widetilde{\mathcal{H}}_n} \left| \log \frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}}(\eta) \right| \lesssim 1.$$

By reading the proof, one can see that Theorem 4 below is based on the approximate likelihood ratio. The first condition of (C8) is required to control the remainder of an approximation. The second condition of (C8) implies that $\|u\|_2 \lesssim \|(I - H)u\|_2 \le \|u\|_2$ for every $u \in \operatorname{span}(\tilde{X}_S)$ with $S$ such that $s \le K_1 \bar{s}_\star$, as the second inequality trivially holds by the fact that $I - H$ is an orthogonal projection. The use of the shifting map $\eta \mapsto \tilde{\eta}_n(\theta, \eta)$ is justified by the condition (C9), which implies that a shift in certain directions does not substantially affect the prior on $\eta$. This is related in spirit to the absolute continuity condition in the semiparametric Bernstein-von Mises theorem (see, for example, Theorem 12. 8 of Ghosal and van der Vaart [17]). We will see that a distributional approximation also requires similar, but stronger conditions.

Lastly, the complexity of the neighborhood $\widetilde{\mathcal{H}}_n$ should also be controlled. Specifically, we make the following condition.

(C10) For $a_n$ and $e_n$ satisfying (C1) and a sufficiently large $C > 0$,

$$\sqrt{\frac{n\bar{\epsilon}_n^2(e_n + a_n\bar{\epsilon}_n^2)}{(s_0 \vee 1)\log p}} + \sqrt{a_n}\int_0^{C\bar{\epsilon}_n} \sqrt{\log N(\delta, \widetilde{\mathcal{H}}_n, d_{B,n})}d\delta \to 0.$$

(C11) The parameter space $\mathcal{H}$ is separable with the pseudo-metric $d_{B,n}$.

Similar to (C8), these conditions are required to control the remainder of an approximation. The integral term comes from the expected supremum of a separable Gaussian process, exploiting the Gaussian likelihood of the model and the separability of $\widetilde{\mathcal{H}}_n$ with the standard deviation metric. Condition (C11) is crucial for this reason. Since we usually put a prior on $\eta$ in an explicit way, condition (C11) is rarely violated in practice. One may see a connection between the first term of (C10) and the conditions for Corollary 1. The former easily tends to zero even if $n\bar{\epsilon}_n^2/\log p$ is increasing, due to the extra term $\bar{\epsilon}_n$ which commonly tends to zero in a polynomial order. Note also that the term $s_0 \vee 1$ appears in (C8) and (C10). Although this gives sharper bounds, the conditions often need to be verified with $s_0 \vee 1$ replaced by 1 as $s_0$ is unknown.

Under the conditions specified above, we obtain the following theorem for the contraction rates for $\theta$ which do not depend on $\bar{\epsilon}_n$. The compatibility numbers below should be understood to be 1 if $s_0 = 0$.

**Theorem 4** (Optimal posterior contraction)**.** *Suppose that* (C1)–(C4)*,* (C5*)*, and* (C6)–(C11) *are satisfied. Then, there exist constants $K_4$ and $K_5$ such that*

$$\begin{aligned}
\mathbb{E}_0\Pi\left(\theta : s_\theta > K_4 s_0 \,\middle|\, Y^{(n)}\right) &\to 0,\\
\mathbb{E}_0\Pi\left(\theta : \|\theta-\theta_0\|_1 > \frac{K_5 s_0\sqrt{\log p}}{\phi_1((K_4+1)s_0)\|X\|_*} \,\middle|\, Y^{(n)}\right) &\to 0,\\
\mathbb{E}_0\Pi\left(\theta : \|\theta-\theta_0\|_2 > \frac{K_5 \sqrt{s_0\log p}}{\phi_2((K_4+1)s_0)\|X\|_*} \,\middle|\, Y^{(n)}\right) &\to 0,\\
\mathbb{E}_0\Pi\left(\theta : \|X(\theta-\theta_0)\|_2 > K_5\sqrt{s_0\log p} \,\middle|\, Y^{(n)}\right) &\to 0.
\end{aligned} \tag{11}$$

Similar to the paragraph followed by Theorem 3, the compatibility numbers are easily bounded away from zero so that they can be removed from the expressions. These are actually weaker than before as $s_0 \le s_\star$. The simplified rates are then available for ease of interpretation.

**Remark 4.** In regression models where no additional mean part $\xi_{\eta,i}$ exists, conditions (C8) and (C9) are trivially satisfied by choosing the zero matrix for $H$. This is also true for (C8*) and (C9*) specified in the next section.

**Remark 5.** Suppose that there exists a matrix $Z_i \in \mathbb{R}^{m_i \times q}$ such that $\xi_{\eta,i} = Z_i h(\eta)$ for every $\eta \in \mathcal{H}$ with some map $h : \mathcal{H} \mapsto \mathbb{R}^q$. Then, a general strategy to choose $H$ is to set $H = \tilde{Z}(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T$ for $\tilde{Z} = (Z_1^T \Delta_{\eta_0,1}^{-1/2}, \dots, Z_n^T \Delta_{\eta_0,n}^{-1/2})^T$. In this case, by the triangle inequality, the first condition of (C8) is satisfied if there exists $\eta_* \in \mathcal{H}$ such that $nd_{A,n}^2(\eta_*, \eta_0)/(s_0 \log p) \to 0$. For (C8*) in the next section, this is replaced by $(s_\star^2 \log p) nd_{A,n}^2(\eta_*, \eta_0) \to 0$. These are trivially the case if there exists $\eta' \in \mathcal{H}$ such that $d_{A,n}(\eta', \eta_0) = 0$. Also similar to Remark 3, a sufficient condition for the second line of (C8) is $\min\{\varsigma_{\min}([X_S, Z]) : s \le K_1 \bar{s}_\star\} \gtrsim 1$ as pre-multiplication of a positive definite matrix by $X_S$ and $Z$ is an isomorphism. This is also sufficient for (C8*) in the next section with $\bar{s}_\star$ replaced by $s_\star$.

**Remark 6.** In many instances, for every $\delta > 0$ and $\zeta_n > 0$, we typically have

$$\log N\left(\delta, \{\eta \in \mathcal{H} : d_{B,n}(\eta, \eta_0) \le \zeta_n\}, d_{B,n}\right) \le 0 \vee r_n \log\left(\frac{b_n \zeta_n}{\delta}\right),$$

for some sequences $r_n$ and $b_n$, especially when the part of $\eta$ involved with $d_{B,n}$ is an $r_n$-dimensional Euclidean parameter. Note that $\int_0^{C\zeta_n} \sqrt{0 \vee r_n \log(b_n \zeta_n/\delta)}\, d\delta$ is equal to

$$\begin{aligned}\int_0^{(C\wedge b_n)\zeta_n} & \sqrt{r_n \log\left(\frac{b_n\zeta_n}{\delta}\right)}\, d\delta \\ &= (C \wedge b_n)\zeta_n \sqrt{r_n \log\left(\frac{b_n}{C \wedge b_n}\right)} + b_n \zeta_n \sqrt{r_n} \int_{\sqrt{\log(b_n/(C\wedge b_n))}}^{\infty} e^{-t^2} dt.\end{aligned}$$

If $b_n$ is increasing, the right hand side is bounded by a multiple of $\zeta_n\sqrt{r_n \log b_n}$ by the tail probability of a normal distribution, while it is bounded by a multiple of $\zeta_n b_n \sqrt{r_n}$ for nonincreasing $b_n$. This simplification is useful to verify (C10) in many applications, and can also be used for (C10*) in the next section.

## 4. Bernstein-von Mises and selection consistency

An extremely important question is whether the true support $S_0$ is recovered with probability tending to one, which is the property called selection consistency. We will show this based on a distributional approximation to the posterior distribution. Combined with selection consistency, the shape approximation also leads to the product of a point mass and a normal distribution, which we call the Bernstein-von Mises theorem. This reduced approximate distribution enables us to correctly quantify the remaining uncertainty of the parameter through the posterior distribution.

### 4.1. Shape approximation to the posterior distribution

It is worth noting that selection consistency can often be verified without a distributional approximation. For example, in sparse linear regression with scalar unknown variance $\sigma^2$, Song and Liang [28] deployed the marginal likelihood of the model support which can be obtained by integrating out $\theta$ and $\sigma^2$ from the likelihood using the inverse gamma kernel. In our general formulation, however, this approach is hard to implement due to the arbitrary structure of a nuisance parameter $\eta$. Indeed, the approach is not directly available even for a parametric covariance matrix with dimension $\overline{m} \geq 2$. In this sense, using a shape approximation could be a natural solution to the problem, which may require some extra conditions on the parameter space and on the priors for $\theta$ and $\eta$.

Recall that the results in Section 3.2 are based on the semiparametric theory. In this section we will need very similar conditions as before, but the requirements are generally stronger, as the remainder of an approximation should be strictly manipulated. Since the setup is high-dimensional, our conditions are even more restrictive than those for semiparametric models with a fixed dimensional parametric segment [e.g., 7]. One may refer to Section 3.3 of Chae et al. [10] for a relevant discussion.

Throughout this section, we only consider $s_\star$ that satisfies the conditions of Theorem 3. First of all, we make a modification of (C7). The following condition is slightly stronger than (C7), but is still not too restrictive as (C5*) requires $s_\star = o(n)$.

(C7*) Condition (C7) is satisfied with $\bar{s}_\star$ replaced by $s_\star$.

The assumption on the prior for $\theta$ is made only through the regularization parameter $\lambda$. As in Castillo et al. [8], $\lambda$ should not increase too fast and should satisfy $\lambda s_\star \sqrt{\log p}/\|X\|_* \to 0$. In our setup, the range of $\lambda$ induces a sufficient condition for this: $s_\star^2 \log p = o(n)$. Since this is weaker than the one that will be made later in this section, the "small lambda regime" is automatically met by a stronger condition for the entire procedure for a distributional approximation (see (C10*) below and the following paragraph).

For sufficiently large constants $\hat{M}_1$ and $\hat{M}_2$, we now define the neighborhoods,

$$\begin{aligned}
\widehat{\Theta}_n &= \left\{\theta \in \Theta : s_\theta \leq K_1 s_\star,\ \|\theta - \theta_0\|_1 \leq \hat{M}_1 s_\star \sqrt{\log p}/\|X\|_* \right\},\\
\widehat{\mathcal{H}}_n &= \left\{\eta \in \mathcal{H} : d_{A,n}(\eta,\eta_0) \leq \hat{M}_2 s_\star \sqrt{\frac{\log p}{n}}, d_{B,n}(\eta,\eta_0) \leq \hat{M}_2 \sqrt{\frac{s_\star \log p}{n}} \right\}.
\end{aligned} \tag{12}$$

Note that $\widehat{\Theta}_n$ is defined with an $\ell_1$-ball, which makes it contract more slowly than $\widetilde{\Theta}_n$ in (10) under (C7*). This is due to technical reasons that for a distributional approximation, the $\ell_1$-ball should be directly manipulated in the complement of $\widehat{\Theta}_n$. The neighborhood $\widehat{\mathcal{H}}_n$ is also increased to be matched with $\widehat{\Theta}_n$. We leave more details on this to the reader; refer to the proof of Theorem 5 below.

As in Section 3.2, we choose a bijective map $\eta \mapsto \tilde{\eta}_n(\theta,\eta)$ which gives rise to $\Phi(\tilde{\eta}_n(\theta,\eta)) = (\tilde{\xi}_\eta + H\tilde{X}(\theta-\theta_0), \tilde{\Delta}_\eta)$ for some orthogonal projection $H$. Again, the orthogonal projection $H$ should be carefully chosen to satisfy some boundedness conditions. The conditions are similar to, but stronger than those in Section 3.2. This is not only because of the increased neighborhoods $\widehat{\Theta}_n$ and $\widehat{\mathcal{H}}_n$, but also because the remainder of an approximation should be bounded on their complements. We precisely make the required conditions below.

(C8$^*$) The orthogonal projection $H$ satisfies

$$s_\star^2 \log p \sup_{\eta\in\widehat{\mathcal{H}}_n} \|(I-H)(\tilde{\xi}_\eta - \tilde{\xi}_{\eta_0})\|_2^2 \to 0,$$

$$\min_{S:s\le K_1 s_\star} \inf_{v\in\mathbb{R}^s:\|v\|_2=1} \frac{\|(I-H)\tilde{X}_S v\|_2}{\|\tilde{X}_S v\|_2} \gtrsim 1.$$

(C9$^*$) The conditional law $\Pi_{n,\theta}$ of $\tilde{\eta}_n(\theta,\eta)$ given $\theta$, induced by the prior, is absolutely continuous relative to its distribution $\Pi_{n,\theta_0}$ at $\theta=\theta_0$, and the Radon-Nikodym derivative $d\Pi_{n,\theta}/d\Pi_{n,\theta_0}$ satisfies

$$\sup_{\theta\in\widehat{\Theta}_n}\sup_{\eta\in\widehat{\mathcal{H}}_n}\left|\log\frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}}(\eta)\right| \to 0.$$

(C10$^*$) For $a_n$ and $e_n$ satisfying (C1) and a sufficiently large $C>0$,

$$s_\star \log p \left\{ s_\star \sqrt{e_n + \frac{a_n s_\star \log p}{n}} \right. \\ \left. + \sqrt{a_n}\int_0^{C\sqrt{(s_\star \log p)/n}} \sqrt{\log N\left(\delta, \widehat{\mathcal{H}}_n, d_{B,n}\right)} d\delta \right\} \to 0.$$

Conditions (C8*)–(C10*) are required for similar reasons as in Section 3.2. We mention that (C10*) is a sufficient condition for the small lambda regime, since its necessary condition is $s_\star^5 \log^3 p = o(n)$ that is stronger than $s_\star^2\log p = o(n)$. This necessary condition for (C10*) is often a sufficient condition in many finite dimensional models.

We define the standardized vector,

$$U = \begin{pmatrix} \Delta_{\eta_0,1}^{-1/2}(Y_1 - X_1\theta_0 - \xi_{\eta_0,1}) \\ \vdots \\ \Delta_{\eta_0,n}^{-1/2}(Y_n - X_n\theta_0 - \xi_{\eta_0,n}) \end{pmatrix} \in \mathbb{R}^{n_*}.$$

Under the assumptions above, the posterior distribution of $\theta$ is approximated by $\Pi^\infty$ given by

$$\Pi^\infty(\theta\in\cdot\,|\,Y^{(n)}) = \sum_{S:s\le K_1 s_\star} \hat{w}_S \left(\mathcal{N}^S_{\hat{\theta}_S, \tilde{X}_S^T(I-H)\tilde{X}_S} \otimes \delta_0^{S^c}\right)(\theta\in\cdot), \qquad (13)$$

where $\mathcal{N}^S_{\mu,\Omega}$ is the Gaussian measure with mean $\mu \in \mathbb{R}^s$ and precision $\Omega \in \mathbb{R}^{s\times s}$ on the coordinate $S$, $\delta_0^{S^c}$ is the Dirac measure at zero on $S^c$, $\hat\theta_S$ is the least squares solution $\hat\theta_S = (\tilde X_S^T(I-H)\tilde X_S)^{-1}\tilde X_S^T(I-H)(U+\tilde X\theta_0)$, and the weights $\hat w_S$ satisfy

$$\hat w_S \propto \frac{\pi_p(s)}{\binom{p}{s}}\left(\frac{\lambda}{2}\right)^s (2\pi)^{s/2}\det\left(\tilde X_S^T(I-H)\tilde X_S\right)^{-1/2}\exp\left\{\frac{1}{2}\|(I-H)\tilde X_S\hat\theta_S\|_2^2\right\}.$$

Another way to express $\Pi^\infty$, for any measurable $\mathcal{B}\subset\mathbb{R}^p$, is

$$\Pi^\infty(\theta\in\mathcal{B}\,|\,Y^{(n)}) = \frac{\sum_{S:s\le K_1 s_\star}\pi_p(s)\binom{p}{s}^{-1}(\lambda/2)^s\int_{\mathcal{B}}\Lambda_n^\star(\theta)d\{\mathcal{L}(\theta_S)\otimes\delta_0(\theta_{S^c})\}}{\sum_{S:s\le K_1 s_\star}\pi_p(s)\binom{p}{s}^{-1}(\lambda/2)^s\int_{\mathbb{R}^p}\Lambda_n^\star(\theta)d\{\mathcal{L}(\theta_S)\otimes\delta_0(\theta_{S^c})\}},$$

where $\mathcal{L}$ denotes the Lebesgue measure and

$$\Lambda_n^\star(\theta) = \exp\left\{-\frac{1}{2}\|(I-H)\tilde X(\theta-\theta_0)\|_2^2 + U^T(I-H)\tilde X(\theta-\theta_0)\right\}. \tag{14}$$

It can be easily checked that both the expressions are equivalent. The results are summarized in the following theorem.

**Theorem 5** (Distributional approximation)**.** *Suppose that* (C1)–(C4)*,* (C5*)*,* (C6)*,* (C7*)–(C10*)*, and* (C11) *are satisfied for some orthogonal projection $H$. Then*

$$\mathbb{E}_0\left\|\Pi(\theta\in\cdot\,|\,Y^{(n)}) - \Pi^\infty(\theta\in\cdot\,|\,Y^{(n)})\right\|_{\mathrm{TV}} \to 0. \tag{15}$$

### *4.2. Model selection consistency*

The shape approximation to the posterior distribution facilitates obtaining the next theorem which shows that the posterior distribution is concentrated on subsets of the true support with probability tending to one. The result is then used as the basis of selection consistency. Similar to the literature, the theorem requires an additional condition on the prior as follows.

(C12) The prior satisfies $A_4 > 1$ and $s_\star \lesssim p^a$ for $a < A_4 - 1$.

**Theorem 6** (Selection, no supersets)**.** *Suppose that* (C1)–(C4)*,* (C5*)*,* (C6)*,* (C7*)–(C10*)*, and* (C11)–(C12) *are satisfied for some orthogonal projection $H$. Then*

$$\mathbb{E}_0\Pi\left(\theta : S_\theta \supset S_0, S_\theta \ne S_0\,|\,Y^{(n)}\right) \to 0. \tag{16}$$

Since coefficients that are too close to zero cannot be identified by any selection strategy, some threshold for the true nonzero coefficients is needed for detection. The requirement of a threshold is a fundamental limitation in high-dimensional setups. We make the following threshold, the so-called beta-min condition. The condition is made in view of the third assertion of Theorem 4. The second assertion can also be used to make a similar threshold, but we only consider the given one below as it is generally weaker.

(C13) The true parameter satisfies

$$\min_{\theta_{0,j}\neq 0} |\theta_{0,j}| > \frac{K_5\sqrt{s_0 \log p}}{\phi_2((K_4+1)s_0)\|X\|_*}.$$

Since Theorem 3 implies that the posterior distribution of the support of $\theta$ includes that of the true support with probability tending to one, selection consistency is an easy consequence of Theorem 6 under the beta-min condition (C13). Moreover, this improves the distributional approximation in (15) so that the posterior distribution can be approximated by a single component of the mixture; that is, the Bernstein-von Mises theorem holds for the parameter component $\theta_{S_0}$. The arguments here are summarized in the following two corollaries, whose proofs are straightforward and thus are omitted.

**Corollary 2** (Selection consistency)**.** *Suppose that* (C1)–(C4)*,* (C5*)*,* (C6)*,* (C7*)–(C10*)*, and* (C11)–(C13) *are satisfied for some orthogonal projection* $H$*. Then*

$$\mathbb{E}_0\Pi\left(\theta : S_\theta \neq S_0 \,|\, Y^{(n)}\right) \to 0. \tag{17}$$

**Corollary 3** (Bernstein-von Mises)**.** *Suppose that* (C1)–(C4)*,* (C5*)*,* (C6)*,* (C7*)–(C10*)*, and* (C11)–(C13) *are satisfied for some orthogonal projection* $H$*. Then*

$$\mathbb{E}_0\left\|\Pi(\theta\in\cdot\,|\,Y^{(n)}) - \left(\mathcal{N}^S_{\hat\theta_{S_0},\tilde X_{S_0}^T(I-H)\tilde X_{S_0}} \otimes \delta_0^{S_0^c}\right)(\theta\in\cdot)\right\|_{\mathrm{TV}} \to 0. \tag{18}$$

Corollary 3 enables us to quantify the remaining uncertainty of the parameter through the posterior distribution. Specifically, we can construct credible sets for the individual components of $\theta_0$ as in Castillo et al. [8]. It is easy to see that by the definition of $\hat\theta_{S_0}$, its $j$th component has a normal distribution, whose mean is the $j$th element of $\theta_{S_0}$ and variance is the $j$th diagonal element of $(\tilde X_{S_0}^T(I-H)\tilde X_{S_0})^{-1}$. Correct uncertainty quantification is thus guaranteed by the weak convergence.

## 5. Applications

In this section, we apply the main results established in this study to the examples considered in Section 1.1. The main objective is to obtain nearly optimal posterior contraction rates and selection consistency via shape approximation to the posterior distribution with the Bernstein-von Mises phenomenon.

To use Corollary 1 for the optimal posterior contraction when $n\bar{\epsilon}_n^2 = \log n$, we simply assume that $s_0 > 0$ for all examples in this section, although Theorem 4 can also be applied under stronger conditions. The assumption $s_0 > 0$ is extremely mild rather than considering the ultra high-dimensional case, i.e., $\log n = o(\log p)$. A large enough $A_4$ is also sufficient instead of the assumption $s_0 > 0$, but we do not pursue this direction as a specific threshold is not available. We check the conditions of Theorem 4 only for more complicated models where $n\bar{\epsilon}_n^2 > \log n$.

### 5.1. Multiple response models with missing components

We first apply the main results to Example 1. To recover posterior contraction of $\Sigma$ from the primitive results, it is necessary to assume that every entry of the response is jointly observed sufficiently many times. To be more specific, let $e_{ij}$ be 1 if the $j$th entry of $Y_i^{\text{aug}}$ is observed and be zero otherwise. The contraction rate of the $(j,k)$th element of $\Sigma$ is directly determined by the order of $n^{-1}\sum_{i=i}^n e_{ij}e_{ik}$. The ideal case is when this quantity is bounded away from zero, that is, the entries are jointly observed at a rate proportional to $n$. Then the recovery is possible without any loss of information. If $n^{-1}\sum_{i=1}^n e_{ij}e_{ik}$ decays to zero, then the optimal recovery is not attainable, but consistent estimation may still be possible with slower rates. With an inverse Wishart prior on $\Sigma$, the following theorem studies the posterior asymptotic properties of the given model.

**Theorem 7.** *Assume that $s_0 > 0$, $1 \lesssim \rho_{\min}(\Sigma_0) \le \rho_{\max}(\Sigma_0) \lesssim 1$, $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$, and $\min_{j,k} n^{-1}\sum_{i=1}^n e_{ij}e_{ik} \gtrsim c_n^{-1}$ for some nondecreasing $c_n$ such that $c_n s_0 \log p = o(n)$. Then the following assertions hold.*

(a) *The optimal posterior contraction rates for $\theta$ in (11) are obtained.*
(b) *The posterior contraction rate for $\Sigma$ is $\sqrt{c_n(s_0\log p)/n}$ with respect to the Frobenius norm.*

*Assume further that $c_n(s_0^2 \vee \log c_n)(s_0\log p)^3 = o(n)$ and $\phi_1(Ds_0) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

(c) *For $H \in \mathbb{R}^{n_*\times n_*}$ the zero matrix, the distributional approximation in (15) holds.*
(d) *If $A_4 > 1$ and $s_0 \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (16) holds.*
(e) *Under the beta-min condition as well as the conditions for (d), the selection consistency in (17) and the Bernstein-von Mises theorem in (18) hold.*

### 5.2. Multivariate measurement error models

We now consider Example 2. For convenience we write $Y^* = (Y_1^*, \dots Y_n^*)^T \in \mathbb{R}^n$, $W = (W_1^T, \dots, W_n^T)^T \in \mathbb{R}^{nq}$, and $X^* = (X_1^*, \dots, X_n^*)^T \in \mathbb{R}^{n\times p}$ in what follows.

In this subsection, we use the symbol $\otimes$ for the Kronecker product of matrices. For priors of the nuisance parameters, normal prior distributions are assigned for the location parameters ($\alpha$, $\beta$, and $\mu$) and an inverse gamma and inverse Wishart prior are used for the scale parameters ($\sigma^2$ and $\Sigma$). The next theorem shows posterior asymptotic properties of the model. In particular, specific forms of their mean and variance for shape approximation are provided considering the modeling structure.

**Theorem 8.** *Assume that $s_0 > 0$, $s_0 \log p = o(n)$, $|\alpha_0| \vee \|\beta_0\|_\infty \vee \|\mu_0\|_\infty \lesssim 1$, $1 \lesssim \sigma_0^2 \lesssim 1$, $1 \lesssim \rho_{\min}(\Sigma_0) \leq \rho_{\max}(\Sigma_0) \lesssim 1$, $\|\theta_0\|_\infty \lesssim \lambda^{-1} \log p$, and $\min\{\varsigma_{\min}([X_S^*, 1_n]) : s \leq Ds_0\} \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

- (a) *The optimal posterior contraction rates for $\theta$ in (11) are obtained.*
- (b) *The contraction rates for $\alpha$, $\beta$, $\mu$, and $\sigma^2$ are $\sqrt{(s_0 \log p)/n}$ relative to the $\ell_2$-norms. The same rate is also obtained for $\Sigma$ with respect to the Frobenius norm.*

*Assume further that $s_0^5 \log^3 p = o(n)$ and $\phi_1(Ds_0) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

- (c) *The distributional approximation in (15) holds with the mean vector*

$$\begin{aligned}\hat{\theta}_S =& (X_S^{*T} H^* X_S^*)^{-1} X_S^{*T} \Big\{ H^* \Big[ \big(Y^* - (\alpha_0 + \mu_0^T \beta_0) 1_n\big) \\ & - \big(I_n \otimes (\beta_0^T \Sigma_0 (\Sigma_0 + \Psi)^{-1})\big) \left(W - 1_n \otimes \mu_0\right) \Big] \Big\}\end{aligned}$$

  *and the covariance matrix $(\sigma_0^2 + \beta_0^T \Sigma_0 (\Sigma_0 + \Psi)^{-1} \Psi \beta_0)(X_S^{*T} H^* X_S^*)^{-1}$ for $H^* = I_n - n^{-1} 1_n 1_n^T$.*
- (d) *If $A_4 > 1$ and $s_0 \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (16) holds.*
- (e) *Under the beta-min condition as well as the conditions for* (d)*, the selection consistency in (17) and the Bernstein-von Mises theorem in (18) hold.*

We note that the marginal law of $W_i$ is given by $W_i \sim \mathrm{N}(\mu, \Sigma + \Psi)$. This gives a hope that the rates for $\mu$ and $\Sigma$ may actually be improved up to the parametric rate $n^{-1/2}$ (possibly up to some logarithmic factors). However, other parameters are connected to the high-dimensional coefficients $\theta$, so such a parametric rate may not be obtained for them.

### 5.3. Parametric correlation structure

Next, our main results are applied to Example 3. A correlation matrix $G_i(\alpha)$ should be chosen so that the conditions in the main theorems can be satisfied. Here we consider a compound-symmetric, a first order autoregressive, and a first order moving average correlation matrices: for $\alpha \in (b_1, b_2)$ with fixed boundaries $b_1$ and $b_2$ of the range, respectively, $\{G_i^{\mathrm{CS}}(\alpha)\}_{j,k} = \mathbb{1}(j = k) + \alpha\mathbb{1}(j \neq k)$, $\{G_i^{\mathrm{AR}}(\alpha)\}_{j,k} = \alpha^{|j-k|}$, and $\{G_i^{\mathrm{MA}}(\alpha)\}_{j,k} = \mathbb{1}(j = k) + \alpha\mathbb{1}(|j-k| = 1)$. The range is chosen so that the corresponding correlation matrix can be positive definite, i.e., $(b_1, b_2) = (0, 1)$ for $G_i^{\mathrm{CS}}(\alpha)$, $(b_1, b_2) = (-1, 1)$ for $G_i^{\mathrm{AR}}(\alpha)$, and $(b_1, b_2) = (-1/2, 1/2)$ for $G_i^{\mathrm{MA}}(\alpha)$. Again, an inverse gamma prior is assigned to $\sigma^2$. For a prior on $\alpha$, we consider a density

$$\Pi(d\alpha) \propto \exp\left\{-\frac{1}{(\alpha - b_1)^{c_1}(b_2 - \alpha)^{c_2}}\right\}, \quad \alpha \in (b_1, b_2),$$

for some $c_1, c_2 > 0$ such that $\Pi(\alpha < t) \lesssim \exp(-(t - b_1)^{-c_1})$ for $t > b_1$ close to $b_1$ and $\Pi(\alpha > t) \lesssim \exp(-(b_2 - t)^{-c_2})$ for $t < b_2$ close to $b_2$.

**Theorem 9.** *Assume that $s_0 > 0$, $s_0 \log p = o(n)$, $\overline{m}n \asymp n_*$, $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$, $\sigma_0^2 \asymp 1$, $\alpha_0 \in [b_1 + \epsilon, b_2 - \epsilon]$ for some fixed $\epsilon > 0$. Suppose further that $\overline{m} \lesssim 1$ for the compound-symmetric correlation matrix and $\log \overline{m} \lesssim \log p$ for the autoregressive and moving average correlation matrices. Then the following assertions hold.*

(a) *For any correlation matrix discussed above, the optimal posterior contraction rates for $\theta$ in* (11) *are obtained.*

(b) *For the autoregressive and moving average correlation matrices, the posterior contraction rates for $\sigma^2$ and $\alpha$ are $\sqrt{(s_0 \log p)/(\overline{m}n)}$ with respect to the $\ell_2$-norms. For the compound-symmetric correlation matrix, their contraction rates are $\sqrt{(s_0 \log p)/n}$ relative to the $\ell_2$-norm.*

*Assume further that $s_0^5 \log^3 p = o(n)$ and $\phi_1(Ds_0) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

(c) *For $H \in \mathbb{R}^{n_* \times n_*}$ the zero matrix, the distributional approximation in* (15) *holds.*

(d) *If $A_4 > 1$ and $s_\star \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in* (16) *holds.*

(e) *Under the beta-min condition as well as the conditions for* (d)*, the selection consistency in* (17) *and the Bernstein-von Mises theorem in* (18) *hold.*

As for the prior for $\alpha$, the property that the tail probabilities decay to zero exponentially fast near both zero and one is crucial for the optimal posterior contraction rates. It should be noted that many common probability distributions with compact supports may not be enough for this purpose (e.g., beta distributions).

The main difference between this example and those in the preceding subsections is that we consider possibly increasing $m_i$ here. Although we have the same form of contraction rates for $\theta$ as in previous examples, the implication is not the same due to a different order of $\|X\|_*$. For increasing $m_i$, it is expected to have $\|X\|_* \asymp \sqrt{n_*}$, which is commonly the case in regression settings. This is reduced to $\|X\|_* \asymp \sqrt{n}$ for the cases with fixed $m_i$, and hence increasing $m_i$

may help get faster rates. While the increasing dimensionality of $m_i$ is often a benefit for contraction properties of $\theta$, this may or may not be the case for the nuisance parameters since it depends on the dimensionality of $\eta$. In the example in this subsection, the dimension of the nuisance parameters is fixed although $m_i$ can increase, which makes their posterior contraction rates faster than those with fixed $m_i$. However, this may not be true if $\eta$ is increasing dimensional. For example, see the example in Section 5.5.

### 5.4. Mixed effects models

For the mixed effects models with sparse regression coefficients in Example 4, we assume that the maximum of $\|Z_i\|_{\mathrm{sp}}$ is bounded, which is particularly mild if $\overline{m}$ is bounded. We also assume that $\sum_{i=1}^n \mathbb{1}(m_i \geq q) \asymp n$ and $\min_i\{\varsigma_{\min}(Z_i) : m_i \geq q\} \gtrsim 1$, that is, $m_i$ is likely to be larger than $q$ with fixed probability and $Z_i$ is a full rank. These conditions are required for (C1) to hold. We put an inverse Wishart prior on $\Psi$ as in other examples. The following theorem shows that the posterior asymptotic properties of the mixed effects models.

**Theorem 10.** *Assume that $s_0 > 0$, $s_0 \log p = o(n)$, $1 \lesssim \rho_{\min}(\Psi_0) \leq \rho_{\max}(\Psi_0) \lesssim 1$, $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$, $\sum_{i=1}^n \mathbb{1}(m_i \geq q) \asymp n$, $\min_i\{\varsigma_{\min}(Z_i) : m_i \geq q\} \gtrsim 1$, and $\max_i\|Z_i\|_{\mathrm{sp}} \lesssim 1$. Then the following assertions hold.*

(a) *The optimal posterior contraction rates for $\theta$ in (11) are obtained.*
(b) *The posterior contraction rate for $\Psi$ is $\sqrt{(s_0 \log p)/n}$ with respect to the Frobenius norm.*

*Assume further that $s_0^5 \log^3 p = o(n)$ and $\phi_1(Ds_0) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

(c) *For $H \in \mathbb{R}^{n_* \times n_*}$ the zero matrix, the distributional approximation in (15) holds.*
(d) *If $A_4 > 1$ and $s_0 \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (16) holds.*
(e) *Under the beta-min condition as well as the conditions for (d), the selection consistency in (17) and the Bernstein-von Mises theorem in (18) hold.*

Note that we assume that $\sigma^2$ is known, which is actually unnecessary at the modeling stage. The assumption was made to find a sequence $a_n$ satisfying (C1) with ease. This can be relaxed only with stronger assumptions on $Z_i$. For example, if $q = 1$ and $Z_i$ is an all-one vector, then the model is equivalent to that with a compound-symmetric correlation matrix in Section 5.3 with some reparameterization, in which $\sigma^2$ can be treated as unknown.

### 5.5. Graphical structure with a sparse precision matrix

For the graphical structure models in Example 5, we define an edge-inclusion indicator $\Upsilon = \{\upsilon_{jk} : 1 \leq j \leq k \leq \overline{m}\}$ such that $\upsilon_{jk} = 1$ if $\omega_{jk} \neq 0$ and $\upsilon_{jk} = 0$

otherwise, where $\omega_{jk}$ is the $(j,k)$th element of $\Omega$. We put a prior with a density $f_1$ on $(0,\infty)$ to the nonzero off-diagonal entries and a prior with a density $f_2$ on $\mathbb{R}$ to the diagonal entries of $\Omega$, such that the support is truncated to a matrix space with restricted eigenvalues and entries. For the edge-inclusion indicator, we use a binomial prior with probability $\varpi$ when $|\Upsilon| := \sum_{j,k} \upsilon_{jk}$ is given, and assign a prior to $|\Upsilon|$ such that $\log \Pi(|\Upsilon| \le \bar{r}) \lesssim -\bar{r}\log\bar{r}$. The prior specification is summarized as

$$\begin{aligned}\Pi(\Omega|\Upsilon) &\propto \prod_{j,k:\upsilon_{jk}=1} f_1(\omega_{jk}) \prod_{j=1}^{\overline{m}} f_2(\omega_{jj}) \mathbb{1}_{\mathcal{M}_0^+(L)}(\Omega),\\ \Pi(\Upsilon) &\propto \varpi^{\bar{r}}(1-\varpi)^{\binom{\overline{m}}{2}-\bar{r}}\Pi(|\Upsilon|=\bar{r}), \quad \log\Pi(|\Upsilon|\le\bar{r}) \lesssim -\bar{r}\log\bar{r},\end{aligned}$$

where $\mathcal{M}_0^+(L)$ is a collection of $\overline{m}\times\overline{m}$ positive definite matrices for a sufficiently large $L$, in which eigenvalues are between $[L^{-1}, L]$ and entries are also bounded by $L$ in absolute value.

**Theorem 11.** *Let $s_\star = s_0 \vee \bar{s}_\star$ for $\bar{s}_\star = (\overline{m}+d)(\log n)/\log p$. Assume that $s_0 > 0$, $s_0 \log p = o(n)$, $\overline{m}\log n = o(n)$, $|\Upsilon_0| \le d$ for some $d$ such that $d\log n = o(n)$, $\Omega_0 \in \mathcal{M}_0^+(cL)$ for some $0 < c < 1$, and $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$. Then the following assertions hold.*

(a) *The posterior contraction rates for $\theta$ are given by* (9). *If $\bar{s}_\star \lesssim 1$, the optimal rates in* (11) *are obtained.*
(b) *The posterior contraction rate of $\Omega$ is $\sqrt{(s_0\log p \vee (\overline{m}+d)\log n)/n}$ with respect to the Frobenius norm.*

*If further $(\bar{s}_\star \vee \overline{m}^2)\bar{s}_\star \log p = o(n)$ and $\phi_1(D\bar{s}_\star) \gtrsim 1$ for a sufficiently large $D$, then the following assertion holds.*

(c) *The optimal posterior contraction rates for $\theta$ in* (11) *are obtained even if $\bar{s}_\star \to \infty$.*

*Assume further that $(s_\star \vee \overline{m})^2(s_\star \log p)^3 = o(n)$ and $\phi_1(Ds_\star) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

(d) *For $H \in \mathbb{R}^{n_*\times n_*}$ the zero matrix, the distributional approximation in* (15) *holds.*
(e) *If $A_4 > 1$ and $s_\star \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in* (16) *holds.*
(f) *Under the beta-min condition as well as the conditions for* (e)*, the selection consistency in* (17) *and the Bernstein-von Mises theorem in* (18) *hold.*

Note that increasing $\overline{m}$ is likely to improve the $\ell_2$-norm contraction rate for $\theta$ as we expect that $\|X\|_* \asymp \sqrt{\overline{m}n}$. In particular, the improvement is clearly the case if $d \lesssim \overline{m}$ and $\phi_2(Ds_\star) \gtrsim 1$ for a sufficiently large $D$. However, as pointed out in Section 5.3, this is not the case for $\Omega$ as its dimension is also increasing.

If we assume that $\log n \lesssim \log \overline{m}$, then the term $\sqrt{(\overline{m}+d)(\log n)/n}$ arising from the sparse precision matrix $\Omega$ becomes $\sqrt{(\overline{m}+d)(\log\overline{m})/n}$. The latter is comparable to the frequentist convergence rate of the graphical lasso in Rothman et al. [27]. Therefore, our rate is deemed to be optimal considering the additional complication due to the mean term involving sparse regression coefficients.

### 5.6. Nonparametric heteroskedastic regression models

Next, we use the main results for Example 6. For a bounded, convex subset $\mathcal{X} \subset \mathbb{R}$, define the $\alpha$-Hölder class $\mathfrak{C}^\alpha(\mathcal{X})$ as the collection of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_{\mathfrak{C}^\alpha} < \infty$, where

$$\|f\|_{\mathfrak{C}^\alpha} = \max_{0 \le k \le \lfloor \alpha \rfloor} \sup_{x \in \mathcal{X}} |f^{(k)}(x)| + \sup_{x,y \in \mathcal{X}: x \neq y} \frac{|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)|}{|x-y|^{\alpha - \lfloor \alpha \rfloor}},$$

with the $k$th derivative $f^{(k)}$ of $f$ and $\lfloor \alpha \rfloor$ the largest integer that is strictly smaller than $\alpha$. Let the true function $v_0$ belong to $\mathfrak{C}^\alpha[0,1]$ with assumption that $v_0$ is strictly positive. While $\alpha > 1/2$ suffices for the basic posterior contraction, we will see that the optimal posterior contraction for $\theta$ requires $\alpha > 1$. The stronger condition $\alpha > 2$ is even needed for the Bernstein-von Mises theorem and the selection consistency, but all these conditions are mild if the true function is sufficiently smooth.

We put a prior on $g$ through B-splines. The function is expressed as a linear combination of $J$-dimensional B-spline basis terms $B_J$ of order $q \ge \alpha$, i.e., $v_\beta(z) = \beta^T B_J(z)$, while an inverse Gaussian prior distribution is independently assigned to each entry of $\beta$. For any measurable function $f : [0,1] \mapsto \mathbb{R}$, we let $\|f\|_\infty = \sup_{z \in [0,1]} |f(z)|$ and $\|f\|_{2,n} = (n^{-1} \sum_{i=1}^n |f(z_i)|^2)^{1/2}$ denote the sup-norm and empirical $L_2$-norm, respectively. To deploy the properties of B-splines, we assume that $z_i$ are sufficiently regularly distributed on $[0,1]$.

**Theorem 12.** *The true function $v_0$ is assumed to be strictly positive on $[0,1]$ and belong to $\mathfrak{C}^\alpha[0,1]$ with $\alpha > 1/2$. We choose $J \asymp (n/\log n)^{1/(2\alpha+1)}$. Let $s_\star = s_0 \vee \bar{s}_\star$ for $\bar{s}_\star = (\log n)^{2\alpha/(2\alpha+1)} n^{1/(2\alpha+1)} / \log p$ and assume that $s_0 > 0$, $J s_0 \log p = o(n)$, and $\|\theta_0\|_\infty \lesssim \lambda^{-1} \log p$. Then the following assertions hold.*

(a) *The posterior contraction rates for $\theta$ are given by (9). If $\bar{s}_\star \lesssim 1$, the optimal rates in (11) are obtained.*
(b) *The posterior contraction rate for $v$ is $\sqrt{(s_0 \log p)/n} \vee (\log n/n)^{\alpha/(2\alpha+1)}$ with respect to the $\|\cdot\|_{2,n}$-norm.*

*If further $\alpha > 1$ and $\phi_1(D\bar{s}_\star) \gtrsim 1$ for a sufficiently large $D$, then the following assertion holds.*

(c) *The optimal posterior contraction rates for $\theta$ in (11) are obtained even if $\bar{s}_\star \to \infty$.*

*Assume further that $\alpha > 2$, $J(s_\star^2 \vee J)(s_\star \log p)^3 = o(n)$ and $\phi_1(D s_\star) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

(d) *The distributional approximation in (15) holds with $H$ the $n \times n$ zero matrix.*

(e) *If $A_4 > 1$ and $s_\star \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in* (16) *holds.*

(f) *Under the beta-min condition as well as the conditions for* (e)*, the selection consistency in* (17) *and the Bernstein-von Mises theorem in* (18) *hold.*

An inverse Gaussian prior is used due to the property that its tail probabilities at both zero and infinity decay to zero exponentially fast. The exponentially decaying tail probabilities in both directions are essential to obtain the optimal contraction rate. Note that standard choices such as gamma and inverse gamma distributions do not satisfy this property.

By investigating the proof, it can be seen that the condition $\alpha > 1/2$ is required to satisfy condition (C1) for posterior contraction, so this condition is not avoidable in applying the main theorems. Unlike Theorem 13 below, assertion (c) does not require any further boundedness condition. This is because the restriction $\alpha > 1$ makes the required bound tend to zero. For the Bernstein-von Mises theorem and the selection consistency, it can be seen that $\alpha > 2$ is necessary for the condition $J(s_\star^2 \vee J)(s_\star \log p)^3 = o(n)$ but not sufficient. Although the requirement $\alpha > 2$ is implied by the latter condition, we specify this in the statement due to its importance. We refer to the proof of Theorem 12 for more details.

### *5.7. Partial linear models*

Lastly, we consider Example 7. We assume that the true function $g_0$ belongs to $\mathfrak{C}^\alpha[0,1]$ for with $\alpha > 0$. Any $\alpha > 0$ suffices for the basic posterior contraction, but stronger restrictions are required for further assertions as in Theorem 12. We put a prior on $g$ through $J$-dimensional B-spline basis terms of order $q \geq a$, i.e., $g_\beta(z) = \beta^T B_J(z)$. With a given $J$, we define the design matrix $W_J = (B_J(z_1), \ldots, B_J(z_n))^T \in \mathbb{R}^{n\times J}$. The standard normal prior is independently assigned to each component of $\beta$ and an inverse gamma prior is assigned to $\sigma^2$. Similar to Section 5.6, we assume that $z_i$ are sufficiently regularly distributed on $[0,1]$.

**Theorem 13.** *The true function is assumed to satisfy $g_0 \in \mathfrak{C}^\alpha[0,1]$ with $\alpha > 0$. We choose $J \asymp (n/\log n)^{1/(2\bar\alpha+1)}$ for some $\bar\alpha \leq \alpha$. Let $s_\star = s_0 \vee \bar s_\star$ for $\bar s_\star = (\log n)^{2\bar\alpha/(2\bar\alpha+1)} n^{1/(2\bar\alpha+1)}/\log p$ and assume that $s_0 > 0$, $s_0 \log p = o(n)$, $\sigma_0^2 \asymp 1$, $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$, and $\min\{\varsigma_{\min}([X_S, W_J]) : s \leq D s_\star\} \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.*

(a) *The posterior contraction rates for $\theta$ are given by* (9)*. If $\bar s_\star \lesssim 1$, the optimal rates in* (11) *are obtained.*

(b) *The contraction rates for $g$ and $\sigma^2$ are $\sqrt{(s_0 \log p)/n} \vee (\log n/n)^{\bar\alpha/(2\bar\alpha+1)}$ with respect to the $\|\cdot\|_{2,n}$- and $\ell_2$-norms, respectively.*

*If further $1/2 \leq \bar\alpha < \alpha$, $(\log n)^{2(\alpha\wedge 2\bar\alpha)/(2\bar\alpha+1)} n^{(-2(\alpha\wedge 2\bar\alpha)+2\bar\alpha+1)/(2\bar\alpha+1)} = o(\log p)$, and $\phi_1(D\bar s_\star) \gtrsim 1$ for a sufficiently large $D$, then the following assertion holds.*

(c) *The optimal posterior contraction rates for $\theta$ in* (11) *are obtained even if* $\bar{s}_\star \to \infty$.

*Assume that* $1 < \bar{\alpha} < \alpha - 1/2$, $(s_\star^2 \log p)(\log n)^{2\alpha/(2\bar{\alpha}+1)} n^{(2(\bar{\alpha}-\alpha)+1)/(2\bar{\alpha}+1)} = o(1)$, $s_\star^5 \log^3 p = o(n)$, *and* $\phi_1(Ds_\star) \gtrsim 1$ *for a sufficiently large* $D$. *Then the following assertions hold.*

(d) *The distributional approximation in* (15) *holds for the projection matrix* $H = W_J(W_J^T W_J)^{-1} W_J^T$.

(e) *If* $A_4 > 1$ *and* $s_\star \lesssim p^a$ *for* $a < A_4 - 1$, *then the no-superset result in* (16) *holds.*

(f) *Under the beta-min condition as well as the conditions for* (e), *the selection consistency in* (17) *and the Bernstein-von Mises theorem in* (18) *hold.*

Here we elaborate more on the choices of the number $J$ of basis terms. For assertions (a)–(b), $J$ can be chosen such that $\bar{\alpha} = \alpha$ which gives rise to the optimal rates for the nuisance parameters. This choice, however, does not satisfy (C8) and (C8*), and hence we need a better approximation for $\|(I - H)\tilde{\xi}_{\eta_0}\|_2$ with some $\bar{\alpha} < \alpha$ to strictly control the remaining bias. For example, if $\bar{\alpha} = \alpha$, the bondedness condition for (c) is reduced to $\bar{s}_\star = o(1)$, which gives the optimal contraction for $\theta$ by (a). Therefore, to incorporate the case that $\bar{s}_\star \to \infty$, there is a need to consider some appropriate $\bar{\alpha}$ that is strictly smaller than $\alpha$. For the Bernstein-von Mises theorem and the selection consistency, the required restriction becomes even stronger such that $\bar{\alpha} < \alpha - 1/2$.

## Appendix A: Proofs for the main results

In this section, we provide proofs of the main theorems. We first describe the additional notations used for the proofs. For a matrix $X$, we write $\rho_1(X) \geq \rho_2(X) \geq \cdots$ for the eigenvalues of $X$ in decreasing order. The notation $\Lambda_n(\theta, \eta) = \prod_{i=1}^n (p_{\theta,\eta,i}/p_{0,i})(Y_i)$ stands for the likelihood ratio of $p_{\theta,\eta}$ and $p_0$. Let $\mathbb{E}_{\theta,\eta}$ denote the expectation operator with the density $p_{\theta,\eta}$ and let $\mathbb{P}_0$ denote the probability operator with the true density. For two densities $f$ and $g$, let $K(f, g) = \int f \log(f/g)$ and $V(f, g) = \int f |\log(f/g) - K(f, g)|^2$ stand for the Kullback-Leibler divergence and variation, respectively. Using some constants $\underline{\rho}_0, \overline{\rho}_0 > 0$, we rewrite (C4) as $\underline{\rho}_0 \leq \min_i \rho_{\min}(\Delta_{\eta_0,i}) \leq \max_i \rho_{\max}(\Delta_{\eta_0,i}) \leq \overline{\rho}_0$ for clarity.

### A.1. Proof of Theorem 1

We first state a lemma showing that the denominator of the posterior distribution is bounded below by a factor with probability tending to one, which will be used to prove the main theorems.

**Lemma 1.** *Suppose that* (C1)–(C4) *are satisfied. Then there exists a constant* $K_0$ *such that*

$$\mathbb{P}_0\left( \int_{\Theta \times \mathcal{H}} \Lambda_n(\theta, \eta) d\Pi(\theta, \eta) \geq \pi_p(s_0) e^{-K_0(s_0 \log p + n\bar{\epsilon}_n^2)} \right) \to 1. \qquad (19)$$

*Proof.* We define the Kullback-Leibler-type neighborhood $\mathcal{B}_n = \{(\theta,\eta) \in \Theta\times\mathcal{H} : \sum_{i=1}^n K(p_{0,i}, p_{\theta,\eta,i}) \le C_1 n\bar\epsilon_n^2, \sum_{i=1}^n V(p_{0,i}, p_{\theta,\eta,i}) \le C_1 n\bar\epsilon_n^2\}$ for a sufficiently large $C_1$. Then Lemma 10 of Ghosal and van der Vaart [16] implies that for any $C > 0$,

$$\mathbb{P}_0\left(\int_{\mathcal{B}_n} \Lambda_n(\theta,\eta) d\Pi(\theta,\eta) \le e^{-(1+C)C_1 n\bar\epsilon_n^2}\Pi(\mathcal{B}_n)\right) \le \frac{1}{C^2 C_1 n\bar\epsilon_n^2}. \qquad (20)$$

Hence, it suffices to show that $\Pi(\mathcal{B}_n)$ is bounded below as in the lemma. By Lemma 9, the Kullback-Leibler divergence and variation of the $i$th observation are given by

$$\begin{aligned} K(p_{0,i}, p_{\theta,\eta,i}) &= \frac{1}{2}\bigg\{ -\sum_{k=1}^{m_i}\log\rho_{i,k}^* - \sum_{k=1}^{m_i}(1-\rho_{i,k}^*) \\ &\qquad + \|\Delta_{\eta,i}^{-1/2}(X_i(\theta-\theta_0) + \xi_{\eta,i} - \xi_{\eta_0,i})\|_2^2\bigg\}, \\ V(p_{0,i}, p_{\theta,\eta,i}) &= \frac{1}{2}\sum_{k=1}^{m_i}(1-\rho_{i,k}^*)^2 + \|\Delta_{\eta_0,i}^{1/2}\Delta_{\eta,i}^{-1}(X_i(\theta-\theta_0) + \xi_{\eta,i} - \xi_{\eta_0,i})\|_2^2, \end{aligned}$$

where $\rho_{i,k}^*$, $k = 1,\dots,m_i$, are the eigenvalues of $\Delta_{\eta_0,i}^{1/2}\Delta_{\eta,i}^{-1}\Delta_{\eta_0,i}^{1/2}$. For $\mathcal{I}_{n,\delta} = \{1 \le i \le n : \sum_{k=1}^{m_i}(1-\rho_{i,k}^*)^2 \ge \delta\}$ with small $\delta > 0$ and $|\mathcal{I}_{n,\delta}|$ the cardinality of $\mathcal{I}_{n,\delta}$, we see that on $\mathcal{B}_n$,

$$a_n\bar\epsilon_n^2 \gtrsim \frac{a_n}{n}\sum_{i=1}^n\sum_{k=1}^{m_i}(1-\rho_{i,k}^*)^2 \ge \frac{a_n\delta|\mathcal{I}_{n,\delta}|}{n} + \frac{a_n}{n}\sum_{i\notin\mathcal{I}_{n,\delta}}\sum_{k=1}^{m_i}(1-\rho_{i,k}^*)^2. \qquad (21)$$

Since every $i \notin \mathcal{I}_{n,\delta}$ satisfies $\sum_{k=1}^{m_i}(1-\rho_{i,k}^*)^2 < \delta$ for small $\delta > 0$, observe that

$$\sum_{i\notin\mathcal{I}_{n,\delta}}\sum_{k=1}^{m_i}(1-\rho_{i,k}^*)^2 \gtrsim \sum_{i\notin\mathcal{I}_{n,\delta}}\sum_{k=1}^{m_i}(1-1/\rho_{i,k}^*)^2 \ge \frac{1}{\overline{\rho}_0^2}\sum_{i\notin\mathcal{I}_{n,\delta}}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2,$$

where the first inequality follows by the relation $|1-x| \asymp |1-x^{-1}|$ as $x \to 1$ and the second inequality holds by (i) of Lemma 10 in Appendix. Since $a_n|\mathcal{I}_{n,\delta}|/n \lesssim a_n\bar\epsilon_n^2$ by (21), it follows using (5) that for some constants $C_2, C_3 > 0$,

$$\begin{aligned} \frac{a_n}{n}\sum_{i\notin\mathcal{I}_{n,\delta}}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2 &\ge a_n d_{B,n}^2(\eta,\eta_0) - \frac{a_n|\mathcal{I}_{n,\delta}|}{n}\max_{1\le i\le n}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2 \\ &\ge (C_2 - C_3 a_n\bar\epsilon_n^2)\max_{1\le i\le n}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2 - e_n. \end{aligned}$$

Combining this with (21), we conclude that $a_n\bar\epsilon_n^2 + e_n \gtrsim \max_i\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2$ on $\mathcal{B}_n$, which implies that $\max_{i,k}|1-\rho_{i,k}^*|$ is small for all sufficiently large $n$, by (i) of Lemma 10 and the inequality $|1-x| \asymp |1-x^{-1}|$ as $x \to 1$. Hence, $\log\rho_{i,k}^*$ can be expanded in the powers of $(1-\rho_{i,k}^*)$ to get $-\log\rho_{i,k}^* - (1-\rho_{i,k}^*) \sim (1-\rho_{i,k}^*)^2/2$

for every $i$ and $k$. Furthermore, since $\max_{i,k}|1-\rho^*_{i,k}|$ is sufficiently small, we obtain that $\sum_{k=1}^{m_i}(1-\rho^*_{i,k})^2 \lesssim \sum_{k=1}^{m_i}(1-1/\rho^*_{i,k})^2 \lesssim \|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2$ by (i) of Lemma 10, and that $\|\Delta_{\eta,i}^{-1}\|_{\mathrm{sp}} \lesssim \|\Delta_{\eta_0,i}^{1/2}\Delta_{\eta,i}^{-1}\Delta_{\eta_0,i}^{1/2}\|_{\mathrm{sp}} \lesssim 1$ by the restriction on the eigenvalues of $\Delta_{\eta_0,i}$. Combining these results, it follows that on $\mathcal{B}_n$, both $n^{-1}\sum_{i=1}^n K(p_{0,i},p_{\theta,\eta,i})$ and $n^{-1}\sum_{i=1}^n V(p_{0,i},p_{\theta,\eta,i})$ are bounded above by a constant multiple of $n^{-1}\|X(\theta-\theta_0)\|_2^2+d_n^2(\eta,\eta_0)$. Hence, $C_1$ can be chosen sufficiently large such that

$$\begin{aligned}\Pi(\mathcal{B}_n) &\ge \Pi\left\{(\theta,\eta)\in\Theta\times\mathcal{H}: n^{-1}\|X\|_*^2\|\theta-\theta_0\|_1^2+d_n^2(\eta,\eta_0)\le 2\bar\epsilon_n^2\right\}\\ &\ge \Pi\left\{\theta\in\Theta: n^{-1}\|X\|_*^2\|\theta-\theta_0\|_1^2\le\bar\epsilon_n^2\right\}\Pi\left\{\eta\in\mathcal{H}: d_n^2(\eta,\eta_0)\le\bar\epsilon_n^2\right\},\end{aligned} \tag{22}$$

by the inequality $\|X\theta\|_2\le\sum_{j=1}^p|\theta_j|\|X_{\cdot j}\|_2\le\|X\|_*\|\theta\|_1$. The logarithm of the second term on the rightmost side is bounded below by a constant multiple of $-n\bar\epsilon_n^2$ by (C2). To find the lower bound for the first term, we shall first work with the case $s_0\ge 1$, and then show that the same lower bound is obtained even when $s_0=0$.

Now, assume that $s_0\ge 1$ and let $\Theta_{0,n}=\{\theta_{S_0}\in\mathbb{R}^{s_0}: n^{-1/2}\|X\|_*\|\theta_{S_0}-\theta_{0,S_0}\|_1\le\epsilon\}$ for $\epsilon>0$ to be chosen later. Then

$$\begin{aligned}&\Pi\{\theta\in\Theta: n^{-1/2}\|X\|_*\|\theta-\theta_0\|_1\le\epsilon\}\\ &\quad\ge\frac{\pi_p(s_0)}{\binom{p}{s_0}}\int_{\Theta_{0,n}} g_{S_0}(\theta_{S_0})d\theta_{S_0}\\ &\quad\ge\frac{\pi_p(s_0)}{\binom{p}{s_0}}e^{-\lambda\|\theta_0\|_1}\int_{\Theta_{0,n}} g_{S_0}(\theta_{S_0}-\theta_{0,S_0})d\theta_{S_0}\end{aligned} \tag{23}$$

by the inequality $g_{S_0}(\theta_{S_0})\ge e^{-\lambda\|\theta_0\|_1}g_{S_0}(\theta_{S_0}-\theta_{0,S_0})$. Using the relation (6.2) of Castillo et al. [8] and the assumption on the prior in (4), the integral on the rightmost side satisfies

$$\begin{aligned}\int_{\Theta_{0,n}} g_{S_0}(\theta_{S_0}-\theta_{0,S_0})d\theta_{S_0} &\ge e^{-\lambda\epsilon\sqrt{n}/\|X\|_*}\frac{(\lambda\epsilon\sqrt{n}/\|X\|_*)^{s_0}}{s_0!}\\ &\ge e^{-L_3\epsilon}\frac{(\epsilon\sqrt{n}/L_1p^{L_2})^{s_0}}{s_0!},\end{aligned} \tag{24}$$

for $s_0>0$, and thus the rightmost side of (23) is bounded below by

$$\pi_p(s_0)(\epsilon\sqrt{n})^{s_0}\exp\left\{-\lambda\|\theta_0\|_1-L_3\epsilon-(L_1+1)s_0\log p-s_0\log L_1\right\},$$

by the inequality $\binom{p}{s_0}s_0!\le p^{s_0}$. Choosing $\epsilon=\bar\epsilon_n$, the first term on the rightmost side of (22) satisfies

$$\begin{aligned}&\Pi\left\{\theta\in\Theta: n^{-1}\|X\|_*^2\|\theta-\theta_0\|_1^2\le\bar\epsilon_n^2\right\}\\ &\quad\ge\pi_p(s_0)(n\bar\epsilon_n^2)^{s_0/2}\exp\left\{-\lambda\|\theta_0\|_1-L_3\bar\epsilon_n-(L_1+1)s_0\log p-s_0\log L_1\right\}.\end{aligned}$$

Note that $n\bar{\epsilon}_n^2 > 1$ and $s_0 + \bar{\epsilon}_n + s_0 \log p \lesssim s_0 \log p$ if $s_0 > 0$, and thus the last display implies that there exists a constant $C_4 > 0$ such that

$$\Pi(\mathcal{B}_n) \geq \pi_p(s_0) \exp\left\{ -C_4 (\lambda \|\theta_0\|_1 + s_0 \log p + n\bar{\epsilon}_n^2) \right\}.$$

If $s_0 = 0$, the first term of (22) is clearly bounded below by $\pi_p(0)$, so that the same lower bound for $\Pi(\mathcal{B}_n)$ in the last display is also obtained since we have $\lambda \|\theta_0\|_1 + s_0 \log p = 0$. Finally, the lemma follows from (20). □

*Proof of Theorem 1.* For the set $\mathcal{B} = \{(\theta, \eta) : s_\theta > \bar{s}\}$ with any integer $\bar{s} \geq s_0$, we see that $\Pi(\mathcal{B})$ is equal to

$$\sum_{s=\bar{s}+1}^{p} \pi_p(s) \leq \pi_p(s_0) \sum_{s=\bar{s}+1}^{p} \left( \frac{A_2}{p^{A_4}} \right)^{s - s_0} \leq \pi_p(s_0) \left( \frac{A_2}{p^{A_4}} \right)^{\bar{s}+1-s_0} \sum_{j=0}^{\infty} \left( \frac{A_2}{p^{A_4}} \right)^{j}.$$

Let $\mathcal{E}_n$ be the event in (19). Since $\Lambda_n(\theta, \eta)$ is nonnegative, by Fubini's theorem and Lemma 1,

$$\begin{aligned} \mathbb{E}_0 \Pi(\mathcal{B} \,|\, Y^{(n)}) \mathbb{1}_{\mathcal{E}_n} &= \mathbb{E}_0 \left[ \frac{\int_{\mathcal{B}} \Lambda_n(\theta, \eta) d\Pi(\theta, \eta)}{\int \Lambda_n(\theta, \eta) d\Pi(\theta, \eta)} \mathbb{1}_{\mathcal{E}_n} \right] \\ &\leq \pi_p(s_0)^{-1} \exp\{ C_1 (s_0 \log p + n\bar{\epsilon}_n^2) \} \Pi(\mathcal{B}) \\ &\lesssim \exp\left\{ (\bar{s} + 1 - s_0)(\log A_2 - A_4 \log p) + 2C_1 s_\star \log p \right\}, \end{aligned} \tag{25}$$

for some constant $C_1$ and sufficiently large $p$. For a sufficiently large constant $C_2$, choose the largest integer that is smaller than $C_2 s_\star$ for $\bar{s}$. Replacing $\bar{s} + 1$ by $C_2 s_\star$ in the last display, it is easy to see that the rightmost side goes to zero. The proof is complete since $\mathbb{P}_0(\mathcal{E}_n^c) \to 0$ by Lemma 1. □

### *A.2. Proof of Theorems 2–3 and Corollary 1*

The following lemma shows that a small piece of the alternative centered at any $(\theta_1, \eta_1) \in \Theta \times \mathcal{H}$ are locally testable with exponentially small errors, provided that the center is sufficiently separated from the truth with respect to the average Rényi divergence. Theorem 2 for posterior contraction relative to the average Rényi divergence will then be proved by showing that the number of those pieces is controlled by the target rate. We write $p_1$ for the density with $(\theta_1, \eta_1)$, and $\mathbb{E}_1$ and $\mathbb{P}_1$ for the expectation and probability with $p_1$, respectively.

**Lemma 2.** *For a given sequence $\gamma_n' > 0$, a sequence $a_n$ satisfying* (C1), *given $(\theta_1, \eta_1) \in \Theta \times \mathcal{H}$ such that $R_n(p_0, p_1) \geq \delta_n^2$ with $\delta_n = o(\sqrt{\overline{m}})$, define*

$$\begin{aligned} \mathcal{F}_{1,n} = \Bigg\{ (\theta, \eta) \in \Theta \times \mathcal{H} : \frac{1}{n} \sum_{i=1}^{n} \| X_i(\theta - \theta_1) + \xi_{\eta,i} - \xi_{\eta_1,i} \|_2^2 \leq \frac{\delta_n^2}{16 \gamma_n'}, \\ d_{B,n}(\eta, \eta_1) \leq \frac{\delta_n^2}{2\overline{m} \gamma_n' \sqrt{a_n}}, \max_{1 \leq i \leq n} \| \Delta_{\eta,i}^{-1} \|_{\mathrm{sp}} \leq \gamma_n' \Bigg\}. \end{aligned} \tag{26}$$

*Then under (C1), there exists a test $\bar\varphi_n$ such that*

$$\mathbb{E}_0\bar\varphi_n \le e^{-n\delta_n^2}, \qquad \sup_{(\theta,\eta)\in\mathcal{F}_{1,n}} \mathbb{E}_{\theta,\eta}(1-\bar\varphi_n) \le e^{-n\delta_n^2/16}.$$

*Proof.* For given $(\theta_1,\eta_1)\in\Theta\times\mathcal{H}$ such that $R_n(p_0,p_1)\ge\delta_n^2$, consider the most powerful test $\bar\varphi_n = \mathbb{1}_{\{\Lambda_n(\theta_1,\eta_1)\ge 1\}}$ given by the Neyman-Pearson lemma. It is then easy to see that

$$\begin{aligned} \mathbb{E}_0\bar\varphi_n &= \mathbb{P}_0\left(\sqrt{\Lambda_n(\theta_1,\eta_1)}\ge 1\right) \le \int\sqrt{p_0p_1} \le e^{-n\delta_n^2}, \\ \mathbb{E}_1(1-\bar\varphi_n) &= \mathbb{P}_1\left(\sqrt{\Lambda_n(\theta_1,\eta_1)}\le 1\right) \le \int\sqrt{p_0p_1} \le e^{-n\delta_n^2}. \end{aligned} \tag{27}$$

The first inequality of the lemma is a direct consequence of the first line of the preceding display. For the second inequality of the lemma, note that by the Cauchy-Schwarz inequality, we have

$$\{\mathbb{E}_{\theta,\eta}(1-\bar\varphi_n)\}^2 \le \mathbb{E}_1(1-\bar\varphi_n)\,\mathbb{E}_1((p_{\theta,\eta}/p_1)(Y^{(n)}))^2.$$

Thus, by the second line of (27), it suffices to show $\mathbb{E}_1((p_{\theta,\eta}/p_1)(Y^{(n)}))^2 \le e^{7n\delta_n^2/8}$ for every $(\theta,\eta)\in\mathcal{F}_{1,n}$. Defining $\Delta_{\eta,i}^* = \Delta_{\eta,i}^{-1/2}\Delta_{\eta_1,i}\Delta_{\eta,i}^{-1/2}$, observe that

$$\begin{aligned} \max_{1\le i\le n}\|\Delta_{\eta,i}^*-I\|_{\mathrm{sp}} &\le \max_{1\le i\le n}\|\Delta_{\eta,i}^{-1}\|_{\mathrm{sp}}\|\Delta_{\eta,i}-\Delta_{\eta_1,i}\|_{\mathrm{sp}} \\ &\le \max_{1\le i\le n}\|\Delta_{\eta,i}^{-1}\|_{\mathrm{sp}}\sqrt{a_n}\,d_{B,n}(\eta,\eta_1) \le \frac{\delta_n^2}{2\overline{m}}, \end{aligned}$$

on the set $\mathcal{F}_{1,n}$, where the second inequality is due to (C1). Since the leftmost side of the display is further bounded below by $\max_i|\rho_k(\Delta_{\eta,i}^*)-1|$ for every $k\le m_i$, we have that

$$1-\frac{\delta_n^2}{2\overline{m}} \le \min_{1\le i\le n}\rho_{\min}(\Delta_{\eta,i}^*) \le \max_{1\le i\le n}\rho_{\max}(\Delta_{\eta,i}^*) \le 1+\frac{\delta_n^2}{2\overline{m}}. \tag{28}$$

Since $\delta_n^2/\overline{m}\to 0$ and $\rho_k(2\Delta_{\eta,i}^*-I) = 2\rho_k(\Delta_{\eta,i}^*)-1$ for every $k\le m_i$, (28) implies that $2\Delta_{\eta,i}^*-I$ is nonsingular for every $i\le n$, and hence on $\mathcal{F}_{1,n}$, it can be shown that $\mathbb{E}_1((p_{\theta,\eta}/p_1)(Y^{(n)}))^2$ can be written as being equal to

$$\begin{aligned} &\prod_{i=1}^n\left\{\det(\Delta_{\eta,i}^*)^{1/2}\det(2I-\Delta_{\eta,i}^{*-1})^{-1/2}\right\} \\ &\times\exp\left\{\sum_{i=1}^n\|(2\Delta_{\eta,i}^*-I)^{-1/2}\Delta_{\eta,i}^{-1/2}(X_i(\theta-\theta_1)+\xi_{\eta,i}-\xi_{\eta_1,i})\|_2^2\right\}. \end{aligned} \tag{29}$$

To bound this, note that $\det(\Delta_{\eta,i}^*)^{1/2}\det(2I-\Delta_{\eta,i}^{*-1})^{-1/2}$ is equal to

$$\prod_{k=1}^{m_i}\left\{\frac{\rho_k(\Delta_{\eta,i}^*)}{2-\rho_k^{-1}(\Delta_{\eta,i}^*)}\right\}^{1/2} \le \left(\frac{1-\delta_n^4/4\overline{m}^2}{1-\delta_n^2/\overline{m}}\right)^{m_i/2} \le \left(1+\frac{3\delta_n^2}{2\overline{m}}\right)^{m_i/2} \le e^{3\delta_n^2/4}, \tag{30}$$

where the first inequality holds by (28), the second inequality holds by the inequality $(1-x^2)/(1-2x) \leq 1+3x$ for small $x > 0$, and the last inequality holds by the inequality $x+1 \leq e^x$. Now, for every $(\theta, \eta) \in \mathcal{F}_{1,n}$, observe that the exponent in (29) is bounded above by

$$\max_{1\leq i\leq n} \|(2\Delta^*_{\eta,i} - I)^{-1}\|_{\rm sp} \max_{1\leq i\leq n} \|\Delta^{-1}_{\eta,i}\|_{\rm sp} \sum_{i=1}^n \|X_i(\theta-\theta_1) + \xi_{\eta,i} - \xi_{\eta_1,i}\|_2^2 \leq \frac{n\delta_n^2}{8},$$

since $\max_i \|(2\Delta^*_{\eta,i} - I)^{-1}\|_{\rm sp} \leq 2$ for large $n$. Combined with (29) and (30), the display completes the proof. □

*Proof of Theorem 2.* Let $\Theta_n = \{\theta \in \Theta : s_\theta \leq K_1 s_\star\}$ and $R_n^\star(\theta,\eta) = R_n(p_{\theta,\eta}, p_0)$. Then for every $\epsilon > 0$,

$$\begin{aligned} &\mathbb{E}_0 \Pi\left((\theta,\eta) \in \Theta \times \mathcal{H} : \sqrt{R_n^\star(\theta,\eta)} > \epsilon \,|\, Y^{(n)}\right) \\ &\quad\leq \mathbb{E}_0 \Pi\left((\theta,\eta) \in \Theta_n \times \mathcal{H} : \sqrt{R_n^\star(\theta,\eta)} > \epsilon \,|\, Y^{(n)}\right) + \mathbb{E}_0 \Pi\left(\Theta_n^c \,|\, Y^{(n)}\right), \end{aligned} \tag{31}$$

where the second term on the right hand side goes to zero by Theorem 1. Hence, it suffices to show that the first term goes to zero for $\epsilon > 0$ chosen to be the threshold in the theorem. Now, let $\Theta_n^* = \{\theta \in \Theta : s_\theta \leq K_1 s_\star, \|\theta\|_\infty \leq p^{L_2+2}/\|X\|_*\}$ and define $\mathcal{F}_{1,n}$ as in (26) with $\gamma_n' = \gamma_n$ and $\delta_n = \epsilon_n$. Then Lemma 2 implies that small pieces of the alternative densities can be tested with exponentially small errors as long as the center is $\epsilon_n$-separated from the true parameter values relative to the average Rényi divergence. To complete the proof, we shall show that the minimal number $N_n^*$ of those small pieces that are needed to cover $\Theta_n^* \times \mathcal{H}_n$ is controlled appropriately in terms of $\epsilon_n$, and that the prior mass of $\Theta_n \setminus \Theta_n^*$ and $\mathcal{H} \setminus \mathcal{H}_n$ decreases fast enough to balance the denominator of the posterior distribution. (For more discussion on a construction of a test using metric entropies, see Section D.2 and Section D.3 of Ghosal and van der Vaart [17].)

Note that for every $\theta, \theta' \in \Theta$ and $\eta, \eta' \in \mathcal{H}$,

$$\frac{1}{n}\sum_{i=1}^n \|X_i(\theta-\theta') + \xi_{\eta,i} - \xi_{\eta',i}\|_2^2 \leq 2\left\{\frac{p^2}{n}\|X\|_*^2 \|\theta-\theta'\|_\infty^2 + d_{A,n}^2(\eta,\eta')\right\},$$

by the inequality $\|X(\theta-\theta')\|_2 \leq \|X\|_* \|\theta-\theta'\|_1 \leq p\|X\|_*\|\theta-\theta'\|_\infty$ and the Cauchy-Schwarz inequality. Since $a_n < n$ and $\epsilon_n^2 > n^{-1}$, it is easy to see that we have $\mathcal{F}_{1,n} \supset \mathcal{F}_{1,n}'$ for

$$\begin{aligned} \mathcal{F}_{1,n}' = \Bigg\{(\theta,\eta) \in \Theta \times \mathcal{H} : \frac{p^2}{n}\|X\|_*^2 \|\theta-\theta_1\|_\infty^2 + d_n^2(\eta,\eta_1) \leq \frac{1}{32\overline{m}^2\gamma_n^2 n^3}, \\ \max_{1\leq i\leq n} \|\Delta^{-1}_{\eta,i}\|_{\rm sp} \leq \gamma_n\Bigg\}, \end{aligned}$$

with the same $(\theta_1, \eta_1)$ used to define $\mathcal{F}_{1,n}$. Hence, $\log N_n^*$ is bounded above by

$$\log N\left(\frac{1}{6\overline{m}\gamma_n np\|X\|_*}, \Theta_n^*, \|\cdot\|_\infty\right) + \log N\left(\frac{1}{6\overline{m}\gamma_n n^{3/2}}, \mathcal{H}_n, d_n\right). \tag{32}$$

Note that for any small $\delta > 0$,

$$N(\delta, \Theta_n^*, \|\cdot\|_\infty) \leq \binom{p}{\lfloor K_1 s_\star \rfloor} \left(\frac{3p^{L_2+2}}{\delta\|X\|_*}\right)^{\lfloor K_1 s_\star \rfloor} \leq \left(\frac{3p^{L_2+3}}{\delta\|X\|_*}\right)^{K_1 s_\star},$$

and thus we obtain

$$\log N\left(\frac{1}{6\overline{m}\gamma_n np\|X\|_*}, \Theta_n^*, \|\cdot\|_\infty\right) \lesssim s_\star(\log \overline{m} + \log\gamma_n + \log p) \lesssim n\epsilon_n^2.$$

Using the last display and the entropy condition (7), the right hand side of (32) is bounded above by a constant multiple of $n\epsilon_n^2$. Hence, by Lemma D.3 of Ghosal and van der Vaart [17], for every $\epsilon > \epsilon_n$, there exists a test $\varphi_n$ such that for some $C_1 > 0$, $\mathbb{E}_0\varphi_n \leq 2\exp(C_1 n\epsilon_n^2 - n\epsilon^2)$ and $\mathbb{E}_{\theta,\eta}(1-\varphi_n) \leq \exp(-n\epsilon^2/16)$ for every $(\theta,\eta) \in \Theta_n^* \times \mathcal{H}_n$ such that $\sqrt{R_n^\star(\theta,\eta)} > \epsilon$. Note that under condition (3) on the prior distribution, we have $-\log\pi_p(s_0) \lesssim s_0\log p - \log\pi_p(0) \lesssim s_\star \log p$ since $\pi_p(0)$ is bounded away from zero. Hence, for $\mathcal{E}_n$ the event in (19) and some constant $C_2 > 0$, the first term on the right hand side of (31) is bounded by

$$\begin{aligned}
&\mathbb{E}_0\Pi\left((\theta,\eta) \in \Theta_n \times \mathcal{H} : \sqrt{R_n^\star(\theta,\eta)} > \epsilon \,|\, Y^{(n)}\right) \mathbb{1}_{\mathcal{E}_n}(1-\varphi_n) + \mathbb{E}_0(\varphi_n + \mathbb{1}_{\mathcal{E}_n^c}) \\
&\quad \leq \left\{ \sup_{(\theta,\eta)\in\Theta_n^*\times\mathcal{H}_n : R_n^\star(\theta,\eta) > \epsilon^2} \mathbb{E}_{\theta,\eta}(1-\varphi_n) + \Pi(\Theta_n \backslash \Theta_n^*) + \Pi(\mathcal{H}\backslash\mathcal{H}_n) \right\} e^{C_2 s_\star \log p} \\
&\qquad + \mathbb{E}_0\varphi_n + \mathbb{P}_0\mathcal{E}_n^c,
\end{aligned}$$

where the term $\mathbb{P}_0\mathcal{E}_n^c$ converges to zero by Lemma 1. Choosing $\epsilon = C_3\epsilon_n$ for a sufficiently large $C_3$, we have

$$\mathbb{E}_0\varphi_n \to 0, \qquad \sup_{(\theta,\eta)\in\Theta_n^*\times\mathcal{H}_n : R_n^\star(\theta,\eta) > \epsilon^2} \mathbb{E}_{\theta,\eta}(1-\varphi_n) e^{C_2 s_\star \log p} \to 0.$$

Furthermore, $\Pi(\mathcal{H} \setminus \mathcal{H}_n)e^{C_2 s_\star \log p}$ goes to zero by condition (8). Now, to show that $\Pi(\Theta_n \setminus \Theta_n^*)$ goes to zero exponentially fast, observe that

$$\begin{aligned}
\Pi(\Theta_n \setminus \Theta_n^*) &= \Pi\left\{\theta \in \Theta : s_\theta \leq K_1 s_\star, \|\theta\|_\infty > p^{L_2+2}/\|X\|_*\right\} \\
&= \sum_{S: s \leq K_1 s_\star} \frac{\pi_p(s)}{\binom{p}{s}} \int_{\{\theta_S : \|\theta_S\|_\infty > p^{L_2+2}/\|X\|_*\}} g_S(\theta_S) d\theta_S \\
&\leq \sum_{S: s \leq K_1 s_\star} \frac{(A_2 p^{-A_4})^s}{\binom{p}{s}} \int_{\{\theta_S : \|\theta_S\|_\infty > p^{L_2+2}/\|X\|_*\}} g_S(\theta_S) d\theta_S.
\end{aligned}$$

by the inequality $\pi_p(s) \leq (A_2 p^{-A_4})^s \pi_p(0)$ for every $S$. Since the tail probability of the Laplace distribution is given by $\int_{|x|>t} 2^{-1}\lambda e^{-\lambda|x|}dx = \exp(-\lambda t)$ for every $t > 0$, the rightmost side of the last display is bounded above by a constant multiple of

$$\sum_{s=1}^{K_1 s_\star} s e^{-\lambda p^{L_2+2}/\|X\|_*} \left(\frac{A_2}{p^{A_4}}\right)^s \lesssim s_\star e^{-\lambda p^{L_2+2}/\|X\|_*}.$$

Since $\lambda p^{L_2+2}/\|X\|_* \gtrsim p^2$ by (4), the right hand side is bounded by $e^{-C_4 p^2}$ for some $C_4 > 0$, and thus $\Pi(\Theta_n \setminus \Theta_n^*)e^{C_2 s_\star \log p}$ goes to zero since $s_\star \log p = o(p^2)$. Finally, we conclude that the left hand side of (31) goes to zero with $\epsilon = C_3\epsilon_n$. □

*Proof of Theorem 3.* By Theorem 2, we obtain the contraction rate of the posterior distribution with respect to the average Rényi divergence $R_n(p_{\theta,\eta}, p_0)$ between $p_{\theta,\eta}$ and $p_0$ given by

$$\begin{aligned}R_n(p_{\theta,\eta}, p_0) = &-\frac{1}{n}\sum_{i=1}^n \log\left\{\frac{(\det\Delta_{\eta,i})^{1/4}(\det\Delta_{\eta_0,i})^{1/4}}{\det((\Delta_{\eta,i}+\Delta_{\eta_0,i})/2)^{1/2}}\right\}\\ &+\frac{1}{4n}\sum_{i=1}^n\|(\Delta_{\eta,i}+\Delta_{\eta_0,i})^{-1}(X_i(\theta-\theta_0)+\xi_{\eta,i}-\xi_{\eta_0,i})\|_2^2.\end{aligned}$$

Define

$$g^2(\Delta_{\eta,i},\Delta_{\eta_0,i}) = 1 - \frac{(\det\Delta_{\eta,i})^{1/4}(\det\Delta_{\eta_0,i})^{1/4}}{\det((\Delta_{\eta,i}+\Delta_{\eta_0,i})/2)^{1/2}}. \tag{33}$$

Then Theorem 2 implies that by the last display,

$$\epsilon_n^2 \gtrsim -\frac{1}{n}\sum_{i=1}^n \log(1-g^2(\Delta_{\eta,i},\Delta_{\eta_0,i})) \geq \frac{1}{n}\sum_{i=1}^n g^2(\Delta_{\eta,i},\Delta_{\eta_0,i}), \tag{34}$$

where the second inequality holds by the inequality $\log x \leq x-1$. Note that by combining (i) and (ii) of Lemma 10 in Appendix, we obtain $g^2(\Delta_{\eta,i},\Delta_{\eta_0,i}) \gtrsim \|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\rm F}^2$ if the left hand side is small. Thus, using the same approach in the proof of Lemma 1, (34) is further bounded below by

$$\begin{aligned}&C_1 d_{B,n}^2(\eta,\eta_0) - C_2\epsilon_n^2 \max_{1\leq i\leq n}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\rm F}^2\\ &\quad\geq (C_1 - C_3 a_n\epsilon_n^2)d_{B,n}^2(\eta,\eta_0) - C_3 e_n\epsilon_n^2,\end{aligned} \tag{35}$$

for some constants $C_1, C_2, C_3 > 0$. Since $C_1 - C_3 a_n\epsilon_n^2$ is bounded away from zero and $e_n$ is decreasing, (34) and (35) imply that $\epsilon_n \gtrsim d_{B,n}(\eta,\eta_0)$. Now, it is easy to see that by (5),

$$\begin{aligned}\max_{1\leq i\leq n}\|\Delta_{\eta,i}+\Delta_{\eta_0,i}\|_{\rm sp}^2 &\leq 2\max_{1\leq i\leq n}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\rm sp}^2 + 8\max_{1\leq i\leq n}\|\Delta_{\eta_0,i}\|_{\rm sp}^2\\ &\lesssim e_n + a_n d_{B,n}^2(\eta,\eta_0)+1,\end{aligned}$$

which is bounded since $e_n + a_n\epsilon_n^2 = o(1)$. Hence, we see that for $\eta_*$ satisfying (C6), $n^{-1}\|X(\theta-\theta_0)\|_2^2 + d_{A,n}^2(\eta,\eta_0)$ is bounded by a constant multiple of

$$\begin{aligned}&\frac{1}{n}\|X(\theta-\theta_0)\|_2^2 + d_{A,n}^2(\eta,\eta_*) + d_{A,n}^2(\eta_*,\eta_0)\\ &\quad\lesssim \frac{1}{n}\sum_{i=1}^n\|X_i(\theta-\theta_0)+\xi_{\eta,i}-\xi_{\eta_*,i}\|_2^2 + d_{A,n}^2(\eta_*,\eta_0)\end{aligned}$$

$$\lesssim \frac{1}{n}\sum_{i=1}^{n}\|(\Delta_{\eta,i}+\Delta_{\eta_0,i})^{-1}(X_i(\theta-\theta_0)+\xi_{\eta,i}-\xi_{\eta_0,i})\|_2^2+d_{A,n}^2(\eta_*,\eta_0).$$

The display implies that $\|X(\theta-\theta_0)\|_2^2+nd_{A,n}^2(\eta,\eta_0)\lesssim n\epsilon_n^2$ by Theorem 2 and (C6). Combining the results verifies the third and fourth assertions of the theorem. For the remainder, observe that $s_{\theta-\theta_0}\le s_\theta+s_0\le K_1s_\star+s_0\lesssim s_\star$ for $\theta$ such that $s_\theta\le K_1s_\star$. Therefore by Theorem 1, the first and the second assertions readily follow from the definitions of $\phi_1$ and $\phi_2$. □

*Proof of Corollary 1.* We first verify the assertion (a). If $s_0>0$ the assertion is trivial. If $s_0=0$, the condition $n\bar\epsilon_n^2/\log p\to 0$ implies that $s_\star\to 0$, and hence Theorem 1 holds with $s_\star=0$. Since this means that $\theta=\theta_0=0$ if $s_0=0$, we can plug in $s_0$ for $s_\star$ in Theorem 3.

Similarly, the assertion (b) trivially holds if $s_0>0$ and we only need to verify the case $s_0=0$. By reading the proof of Theorem 1, one can see that (25) goes to zero for large enough $A_4$ if $s_0=0$. This completes the proof. □

### A.3. Proof of Theorem 4

To prove Theorem 4, we first provide preliminary results. Some of these will also be used to prove Theorems 5–6.

**Lemma 3.** *Suppose that* (C1), (C2), (C7), (C8) *and* (C10) *are satisfied for some orthogonal projection $H$. Then, for $\Lambda_n^*(\theta,\eta)=(p_{\theta,\eta}/p_{\theta_0,\tilde\eta_n(\theta,\eta)})(Y^{(n)})$ and $\Lambda_n^\star(\theta)$ in (14) with the corresponding $H$, there exists a positive sequence $\delta_n\to 0$ such that for any $\theta$ with $s_\theta\le K_1\bar s_\star$,*

$$\begin{aligned}\mathbb{P}_0\Bigg(&\sup_{\eta\in\widetilde{\mathcal{H}}_n}|\log\Lambda_n^*(\theta,\eta)-\log\Lambda_n^\star(\theta)|\\ &\le\delta_n\left\{\|X(\theta-\theta_0)\|_2\sqrt{(s_\theta+s_0)\log p}+\|X(\theta-\theta_0)\|_2^2\right\}\Bigg)\to 1.\end{aligned}\tag{36}$$

*Proof.* If $s_\theta=s_0=0$, the left hand side in the probability operator is zero, and the assertion trivially holds. We thus only consider the case $s_\theta+s_0>0$ below.

By Markov's inequality, it suffices to show that there exists a positive sequence $\delta_n'=o(\delta_n)$ such that

$$\begin{aligned}&\mathbb{E}_0\sup_{\eta\in\widetilde{\mathcal{H}}_n}|\log\Lambda_n^*(\theta,\eta)-\log\Lambda_n^\star(\theta)|\\ &\quad\le\delta_n'\left\{\|X(\theta-\theta_0)\|_2\sqrt{(s_\theta+s_0)\log p}+\|X(\theta-\theta_0)\|_2^2\right\}.\end{aligned}\tag{37}$$

Let $\Delta_\eta^\star\in\mathbb{R}^{n_*\times n_*}$ be the block-diagonal matrix formed by stacking $\Delta_{\eta_0,i}^{1/2}\Delta_{\eta,i}^{-1}\Delta_{\eta_0,i}^{1/2}$, $i=1,\dots,n$, and observe that

$$\log\Lambda_n^*(\theta,\eta)=-\frac{1}{2}\|\Delta_\eta^{\star1/2}(I-H)\tilde X(\theta-\theta_0)\|_2^2$$

$$+ (\theta-\theta_0)^T\tilde{X}^T(I-H)\Delta_\eta^\star\{U-(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0})-H\tilde{X}(\theta-\theta_0)\}.$$

The left hand side of (37) is thus bounded by the sum of the following terms:

$$\sup_{\eta\in\widetilde{\mathcal{H}}_n}\left|(\theta-\theta_0)^T\tilde{X}^T(I-H)(I-\Delta_\eta^\star)(I-H)\tilde{X}(\theta-\theta_0)\right|, \tag{38}$$

$$\sup_{\eta\in\widetilde{\mathcal{H}}_n}\left|(\theta-\theta_0)^T\tilde{X}^T(I-H)\Delta_\eta^\star(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0}+H\tilde{X}(\theta-\theta_0))\right|, \tag{39}$$

$$\mathbb{E}_0\sup_{\eta\in\widetilde{\mathcal{H}}_n}\left|(\theta-\theta_0)^T\tilde{X}^T(I-H)(I-\Delta_\eta^\star)U\right|. \tag{40}$$

First, observe that (38) is bounded above by a constant multiple of

$$\sup_{\eta\in\widetilde{\mathcal{H}}_n}\|I-\Delta_\eta^\star\|_{\mathrm{sp}}\|\tilde{X}(\theta-\theta_0)\|_2^2 \lesssim \|X(\theta-\theta_0)\|_2^2\sup_{\eta\in\widetilde{\mathcal{H}}_n}\max_{1\le i\le n}\|\Delta_{\eta,i}^{-1}-\Delta_{\eta_0,i}^{-1}\|_{\mathrm{F}}. \tag{41}$$

Using (i) of Lemma 10 and the inequality $|1-x|\asymp|1-x^{-1}|$ as $x\to1$, we obtain that for $\rho_{i,k}^*=\rho_k(\Delta_{\eta_0,i}^{1/2}\Delta_{\eta,i}^{-1}\Delta_{\eta_0,i}^{1/2})$,

$$\|\Delta_{\eta,i}^{-1}-\Delta_{\eta_0,i}^{-1}\|_{\mathrm{F}}^2 \lesssim \sum_{k=1}^{m_i}\left(1-\rho_{i,k}^*\right)^2 \lesssim \sum_{k=1}^{m_i}\left(1-1/\rho_{i,k}^*\right)^2 \lesssim \|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2, \tag{42}$$

provided that the rightmost side is sufficiently small. Because $\max_i\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\mathrm{F}}^2 \le e_n+a_nd_{B,n}^2(\eta,\eta_0) \lesssim e_n+a_n\bar{\epsilon}_n^2$ on $\widetilde{\mathcal{H}}_n$, (42) holds. This implies that for all sufficiently large $n$, the right hand side of (41) is bounded above by a constant multiple of

$$\|X(\theta-\theta_0)\|_2^2\sup_{\eta\in\widetilde{\mathcal{H}}_n}\sqrt{e_n+a_nd_{B,n}^2(\eta,\eta_0)} \lesssim \|X(\theta-\theta_0)\|_2^2\sqrt{e_n+a_n\bar{\epsilon}_n^2},$$

where $e_n+a_n\bar{\epsilon}_n^2=o(1)$ due to (C1) and (C2).

Next, (39) is equal to

$$\sup_{\eta\in\widetilde{\mathcal{H}}_n}\left|(\theta-\theta_0)^T\tilde{X}^T(I-H)\left\{(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0})-(I-\Delta_\eta^\star)(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0}+H\tilde{X}(\theta-\theta_0))\right\}\right|.$$

By the triangle inequality, the display is bounded by a constant multiple of

$$\begin{aligned}&\|X(\theta-\theta_0)\|_2\sup_{\eta\in\widetilde{\mathcal{H}}_n}\|(I-H)(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0})\|_2\\ &+\sup_{\eta\in\widetilde{\mathcal{H}}_n}\left\{\|X(\theta-\theta_0)\|_2^2+\|X(\theta-\theta_0)\|_2\sqrt{n}d_{A,n}(\eta,\eta_0)\right\}\max_{1\le i\le n}\|\Delta_{\eta,i}^{-1}-\Delta_{\eta_0,i}^{-1}\|_{\mathrm{sp}}.\end{aligned} \tag{43}$$

Using the same approach used in (42), the second term is further bounded above by a constant multiple of

$$\|X(\theta-\theta_0)\|_2^2\sqrt{e_n+a_n\bar{\epsilon}_n^2}+\|X(\theta-\theta_0)\|_2\sqrt{n\bar{\epsilon}_n^2(e_n+a_n\bar{\epsilon}_n^2)}.$$

Therefore, by (C8) and (C10), (43) is bounded by

$$\delta_n'\{\|X(\theta-\theta_0)\|_2\sqrt{(s_0\vee 1)\log p}+\|X(\theta-\theta_0)\|_2^2\}$$

for some $\delta_n'\to 0$. This is not more than the right hand side of (37) if $s_\theta+s_0>0$.

Note also that (40) is bounded by

$$\begin{aligned}&\|\theta-\theta_0\|_1\,\mathbb{E}_0\sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{X}^T(I-H)(I-\Delta_\eta^\star)U\|_\infty\\&\quad\le\frac{\sqrt{s_\theta+s_0}\|X(\theta-\theta_0)\|_2}{\phi_1(s_\theta+s_0)\|X\|_*}\mathbb{E}_0\sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{X}^T(I-H)(I-\Delta_\eta^\star)U\|_\infty.\end{aligned}$$

We have that $\phi_1(s_\theta+s_0)\ge\phi_1(K_1\bar{s}_\star+s_0)\gtrsim 1$ by condition (C7). By Lemma 4 below, one can see that

$$\begin{aligned}&\mathbb{E}_0\sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{X}^T(I-H)(I-\Delta_\eta^\star)U\|_\infty\\&\quad\lesssim\|X\|_*\sqrt{\log p}\left\{\sqrt{e_n+a_n\bar{\epsilon}_n^2}+\sqrt{a_n}\int_0^{C_3\bar{\epsilon}_n}\sqrt{\log N(\delta,\widetilde{\mathcal{H}}_n,d_{B,n})}d\delta\right\},\end{aligned}\tag{44}$$

for some $C_3>0$. The term in the braces goes to zero by (C10). Combining the bounds, we easily see that there exists $\delta_n'\to 0$ satisfying (37). The assertion holds by choosing $\delta_n=\sqrt{\delta_n'}$. □

**Lemma 4.** *Consider a neighborhood $\mathcal{H}_n^*=\{\eta\in\mathcal{H}:d_{B,n}(\eta,\eta_0)\le\zeta_n\}$ with any given $\zeta_n=o(a_n^{-1/2})$ for $a_n$ satisfying (C1). Then, for any orthogonal projection $P$ and a sufficiently large $C>0$, we have that under (C1),*

$$\begin{aligned}&\mathbb{E}_0\sup_{\eta\in\mathcal{H}_n^*}\|\tilde{X}^TP(I-\Delta_\eta^\star)U\|_\infty\\&\quad\lesssim\|X\|_*\sqrt{\log p}\left\{\sqrt{e_n+a_n\zeta_n^2}+\sqrt{a_n}\int_0^{C\zeta_n}\sqrt{\log N\left(\delta,\mathcal{H}_n^*,d_{B,n}\right)}d\delta\right\},\end{aligned}$$

*where $\Delta_\eta^\star\in\mathbb{R}^{n_*\times n_*}$ is the block-diagonal matrix formed by stacking the matrices $\Delta_{\eta_0,i}^{1/2}\Delta_{\eta,i}^{-1}\Delta_{\eta_0,i}^{1/2}$, $i=1,\dots,n$.*

*Proof.* Let $W_{\eta,j}=\tilde{X}_{\cdot j}^TP(I-\Delta_\eta^\star)U$ for $\tilde{X}_{\cdot j}\in\mathbb{R}^{n_*}$ the $j$th column of $\tilde{X}$. Then, by Lemma 2.2.2 of van der Vaart and Wellner [29] applied with $\psi(x)=e^{x^2}-1$, the expectation in the lemma is equal to

$$\mathbb{E}_0\max_{1\le j\le p}\sup_{\eta\in\mathcal{H}_n^*}|W_{\eta,j}|\le\left\|\max_{1\le j\le p}\sup_{\eta\in\mathcal{H}_n^*}|W_{\eta,j}|\right\|_\psi\lesssim\sqrt{\log p}\max_{1\le j\le p}\left\|\sup_{\eta\in\mathcal{H}_n^*}|W_{\eta,j}|\right\|_\psi,\tag{45}$$

where $\|\cdot\|_\psi$ is the Orlicz norm for $\psi$. For any $\eta_1,\eta_2\in\mathcal{H}_n^*$, define the standard deviation pseudo-metric between $W_{\eta_1,j}$ and $W_{\eta_2,j}$ as

$$d_{\sigma,j}(\eta_1,\eta_2)\coloneqq\sqrt{\mathrm{Var}(W_{\eta_1,j}-W_{\eta_2,j})}=\|(\Delta_{\eta_1}^\star-\Delta_{\eta_2}^\star)P\tilde{X}_{\cdot j}\|_2.$$

Using the tail bound for normal distributions and Lemma 2.2.1 of van der Vaart and Wellner [29], we see that $\|W_{\eta_1,j} - W_{\eta_2,j}\|_\psi \lesssim d_{\sigma,j}(\eta_1,\eta_2)$ for every $\eta_1, \eta_2 \in \mathcal{H}_n^*$. We shall show that $\mathcal{H}_n^*$ is a separable pseudo-metric space with $d_{\sigma,j}$ for every $j \le p$. Then, under the true model $\mathbb{P}_0$, we see that $\{W_{\eta,j} : \eta \in \mathcal{H}_n^*\}$ is a separable Gaussian process for $d_{\sigma,j}$. Hence, by Corollary 2.2.5 of van der Vaart and Wellner [29], for any fixed $\eta' \in \mathcal{H}_n^*$,

$$\left\| \sup_{\eta \in \mathcal{H}_n^*} |W_{\eta,j}| \right\|_\psi \lesssim \|W_{\eta',j}\|_\psi + \int_0^{\mathrm{diam}_j(\mathcal{H}_n^*)} \sqrt{\log N(\epsilon/2, \mathcal{H}_n^*, d_{\sigma,j})}\, d\epsilon, \tag{46}$$

where $\mathrm{diam}_j(\mathcal{H}_n^*) = \sup\{d_{\sigma,j}(\eta_1,\eta_2) : \eta_1,\eta_2 \in \mathcal{H}_n^*\}$. It is clear that $W_{\eta',j}$ possesses a normal distribution with mean zero and variance $\|(I - \Delta_{\eta'}^\star)P\tilde{X}_{\cdot j}\|_2^2$.

Using Lemma 2.2.1 of van der Vaart and Wellner [29] again, we see that

$$\begin{aligned} \|W_{\eta',j}\|_\psi &\lesssim \|(I - \Delta_{\eta'}^\star)P\tilde{X}_{\cdot j}\|_2 \\ &\lesssim \max_{1\le i\le n} \|\Delta_{\eta',i}^{-1} - \Delta_{\eta_0,i}^{-1}\|_2 \|X_{\cdot j}\|_2 \\ &\lesssim \|X\|_* \sqrt{e_n + a_n\zeta_n^2}, \end{aligned} \tag{47}$$

for every $\eta' \in \mathcal{H}_n^*$. Here the last inequality holds by using (42) and the fact that $\max_i \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|_{\mathrm{F}}^2 \le e_n + a_n d_{B,n}^2(\eta,\eta_0) \lesssim e_n + a_n\zeta_n^2 = o(1)$ on $\mathcal{H}_n^*$, under (C1).

Next, to further bound the second term in (46), note that for every $\eta_1, \eta_2 \in \mathcal{H}_n^*$,

$$a_n\zeta_n^2 \gtrsim \sum_{k=1}^2 2a_n d_{B,n}^2(\eta_k,\eta_0) \ge a_n d_{B,n}^2(\eta_1,\eta_2) \ge \max_{1\le i\le n} \|\Delta_{\eta_1,i} - \Delta_{\eta_2,i}\|_{\mathrm{F}}^2,$$

which is further bounded below by

$$\min_{1\le i\le n} \rho_{\min}^2(\Delta_{\eta_2,i}) \max_{1\le i\le n} \sum_{k=1}^{m_i} \left\{1 - 1/\rho_k(\Delta_{\eta_2,i}^{1/2}\Delta_{\eta_1,i}^{-1}\Delta_{\eta_2,i}^{1/2})\right\}^2,$$

using (i) of Lemma 10. In the last display, we see that $\min_i \rho_{\min}(\Delta_{\eta_2,i})$ is bounded away from zero since

$$\max_{1\le i\le n} \|\Delta_{\eta_2,i}^{-1}\|_{\mathrm{sp}} \le \max_{1\le i\le n} \|\Delta_{\eta_2,i}^{-1} - \Delta_{\eta_0,i}^{-1}\|_{\mathrm{sp}} + \max_{1\le i\le n} \|\Delta_{\eta_0,i}^{-1}\|_{\mathrm{sp}} \lesssim \sqrt{e_n + a_n\zeta_n^2} + 1,$$

and hence every eigenvalue $\rho_k(\Delta_{\eta_2,i}^{1/2}\Delta_{\eta_1,i}^{-1}\Delta_{\eta_2,i}^{1/2})$ is bounded below and above by a multiple of its reciprocal, as $a_n\zeta_n^2 \to 0$. This implies that $a_n\zeta_n^2$ is further bounded below by a constant multiple of

$$\begin{aligned} &\max_{1\le i\le n} \sum_{k=1}^{m_i} \left\{1 - \rho_k(\Delta_{\eta_2,i}^{1/2}\Delta_{\eta_1,i}^{-1}\Delta_{\eta_2,i}^{1/2})\right\}^2 \\ &\quad \ge \min_{1\le i\le n} \rho_{\min}^2(\Delta_{\eta_2,i}) \max_{1\le i\le n} \|\Delta_{\eta_1,i}^{-1} - \Delta_{\eta_2,i}^{-1}\|_{\mathrm{F}}^2. \end{aligned}$$

By the definition of $d_{\sigma,j}$ and the preceding displays, we thus obtain

$$\begin{aligned} d_{\sigma,j}(\eta_1,\eta_2) &\le \|\Delta^\star_{\eta_1} - \Delta^\star_{\eta_2}\|_{\mathrm{sp}} \|\tilde{X}_{\cdot j}\|_2 \\ &\lesssim \|X_{\cdot j}\|_2 \max_{1\le i\le n} \|\Delta^{-1}_{\eta_1,i} - \Delta^{-1}_{\eta_2,i}\|_{\mathrm{sp}} \\ &\lesssim \|X_{\cdot j}\|_2 \sqrt{a_n}\, d_{B,n}(\eta_1,\eta_2), \end{aligned} \tag{48}$$

for every $\eta_1,\eta_2 \in \mathcal{H}^*_n$. Hence, using that $\mathrm{diam}_j(\mathcal{H}^*_n) \lesssim \|X_{\cdot j}\|_2 \zeta_n \sqrt{a_n}$, we can bound the second term in (46) above by a constant multiple of

$$\int_0^{C_1\|X_{\cdot j}\|_2\zeta_n\sqrt{a_n}} \sqrt{\log N\left(\epsilon/C_2\|X_{\cdot j}\|_2\sqrt{a_n}, \mathcal{H}^*_n, d_{B,n}\right)}d\epsilon,$$

for some $C_1, C_2 > 0$. This can be further bounded by replacing $\|X_{\cdot j}\|_2$ in the display by $\|X\|_*$. Then, using (45), (46), and (47), and by the substitution $\delta = \epsilon/(C_2\|X\|_*\sqrt{a_n})$ for the last display, we bound (45) above by a constant multiple of

$$\|X\|_*\sqrt{\log p}\left\{\sqrt{e_n + a_n\zeta_n^2} + \sqrt{a_n}\int_0^{C_3\zeta_n} \sqrt{\log N\left(\delta, \mathcal{H}^*_n, d_{B,n}\right)}d\delta\right\},$$

for some $C_3 > 0$.

To complete the proof, it remains to show that $\mathcal{H}^*_n$ is a separable pseudo-metric space with $d_{\sigma,j}$ for every $j \le p$. By (48), we see that $d_{\sigma,j}(\eta_1,\eta_2) \lesssim \|X\|_*\sqrt{a_n}\, d_{B,n}(\eta_1,\eta_2)$ for every $\eta_1,\eta_2 \in \mathcal{H}^*_n$. This implies that $\mathcal{H}^*_n$ is separable with $d_{\sigma,j}$ since $\mathcal{H}$ is separable with $d_{B,n}$. □

**Lemma 5.** *For any orthogonal projection $P$,*

$$\mathbb{P}_0\left(\|\tilde{X}^T P U\|_\infty > 2\underline{\rho}_0^{-1/2}\sqrt{\log p}\|X\|_*\right) \le \frac{2}{p}.$$

*Proof.* Note first that $\tilde{X}^T_{\cdot j} P U$ has a normal distribution with mean zero and variance $\|P\tilde{X}_{\cdot j}\|_2^2$, and hence we have

$$\mathbb{P}_0\left(\|\tilde{X}^T P U\|_\infty > t \max_{1\le j\le p}\|P\tilde{X}_{\cdot j}\|_2\right) \le 2pe^{-t^2/2}, \quad t > 0,$$

by the tail probabilities of normal distributions. By choosing $t = 2\sqrt{\log p}$ and using the inequality $\|P\tilde{X}_{\cdot j}\|_2 \le \|\tilde{X}_{\cdot j}\|_2 \le \underline{\rho}_0^{-1/2}\|X\|_*$ for every $j \le p$, we verify the assertion. □

**Lemma 6.** *If* (C7) *and* (C10) *are satisfied and $s_0 \log p \lesssim n\bar{\epsilon}_n^2$, there exists a constant $K_0' > 0$ such that*

$$\mathbb{P}_0\left(\inf_{\eta\in\tilde{\mathcal{H}}_n} \int \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(Y^{(n)})d\Pi(\theta) \ge e^{-K_0'(1+s_0\log p)}\right) \to 1. \tag{49}$$

*Proof.* Let $\Theta_n^* = \{\theta \in \Theta : s_\theta = s_0, \|X(\theta-\theta_0)\|_2^2 \leq 1\}$. Restricting the integral to this set, the left hand side of the inequality in (49) is bounded below by

$$\begin{aligned}\inf_{\eta\in\widetilde{\mathcal{H}}_n} \int_{\Theta_n^*} \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(Y^{(n)})d\Pi(\theta) &\geq \int_{\Theta_n^*} \inf_{\eta\in\widetilde{\mathcal{H}}_n} \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(Y^{(n)})d\Pi(\theta) \\ &= \int_{\Theta_n^*} \exp\left(\inf_{\eta\in\widetilde{\mathcal{H}}_n} \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(Y^{(n)})\right) d\Pi(\theta).\end{aligned} \tag{50}$$

The exponent is equal to

$$\begin{aligned}&\inf_{\eta\in\widetilde{\mathcal{H}}_n}\left\{(\theta-\theta_0)^T\tilde{X}^T\Delta_\eta^\star(U-\tilde{\xi}_\eta+\tilde{\xi}_{\eta_0}) - \frac{1}{2}\|\Delta_\eta^{\star 1/2}\tilde{X}(\theta-\theta_0)\|_2^2\right\} \\ &\quad\gtrsim -\|\theta-\theta_0\|_1 \sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{X}^T\Delta_\eta^\star U\|_\infty \\ &\qquad - \|X(\theta-\theta_0)\|_2 \sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{\xi}_\eta - \tilde{\xi}_{\eta_0}\|_2 - \|X(\theta-\theta_0)\|_2^2,\end{aligned} \tag{51}$$

since $\|\Delta_\eta^\star\|_{\mathrm{sp}} \lesssim 1$ on $\widetilde{\mathcal{H}}_n$. We first consider the case $s_0 > 0$. Observe that $\sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{X}^T\Delta_\eta^\star U\|_\infty \leq \|\tilde{X}^T U\|_\infty + \sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{X}^T(I-\Delta_\eta^\star)U\|_\infty$, where the first term is bounded by a constant multiple of $\|X\|_*\sqrt{\log p}$ with $\mathbb{P}_0$-probability tending to one, due to Lemma 5. By Lemma 4 applied with $P = I$ together with (C10), the expected value of the second term is bounded by $\delta_n\|X\|_*\sqrt{\log p}$ for some $\delta_n \to 0$. Hence, for any $M_n \to \infty$,

$$\mathbb{P}_0\left(\sup_{\eta\in\widetilde{\mathcal{H}}_n}\|\tilde{X}^T(I-\Delta_\eta^\star)U\|_\infty \leq M_n\delta_n\|X\|_*\sqrt{\log p}\right) \to 1.$$

Consequently, taking a sufficiently slowly increasing $M_n$ for the above, (51) is bounded below by a constant multiple of

$$-\|X\|_*\|\theta-\theta_0\|_1\sqrt{\log p} - \|X(\theta-\theta_0)\|_2^2,$$

with $\mathbb{P}_0$-probability tending to one. Note that $\|X\|_*\|\theta-\theta_0\|_1 \leq \sqrt{s_\theta+s_0}\|X(\theta-\theta_0)\|_2/\phi_1(s_\theta+s_0)$ and $\phi_1(s_\theta+s_0) = \phi_1(2s_0) \gtrsim 1$ on $\Theta_n^*$ by (C7), if $s_0\log p \lesssim n\bar{\epsilon}_n^2$. The last display is thus bounded below by $-C_1 s_0\log p$ for some $C_1 > 0$, uniformly over $\theta \in \Theta_n^*$. Consequently, with $\mathbb{P}_0$-probability tending to one, (50) is bounded below by

$$e^{-C_1 s_0\log p}\Pi(\Theta_n^*) \geq \pi_p(s_0)e^{-C_2 s_0\log p},$$

for some $C_2 > 0$, where the inequality holds by (23) and (24) since $\lambda\|\theta_0\|_1 \leq s_0\log p$ by (C3). Since $-\log\pi_p(s_0) \lesssim s_0\log p$ if $s_0 > 0$, the display is further bounded below as in the assertion.

If $s_0 = 0$, (51) is equal to zero on $\Theta_n^*$, as this is a singleton set $\{\theta : \theta = 0\}$. This means that (50) is bounded below by $\pi_p(0)$, which is also bounded away from zero. This leads to the desired assertion. □

*Proof of Theorem 4.* The idea of our proof is similar in part to that of Theorem 3.5 in Chae et al. [10]. We only need to verify the first and fourth assertions. The second and third assertions then follow from the definitions of $\phi_1$ and $\phi_2$. Note also that we only need to consider the case $s_0 \log p \lesssim n\bar{\epsilon}_n^2$, as the assertions follow from Theorems 1 and 3 if $s_0 \log p \gtrsim n\bar{\epsilon}_n^2$.

Let $\mathcal{B}_n = \{\theta \in \Theta : s_\theta > K_4 s_0\} \cup \{\theta \in \Theta : \|X(\theta - \theta_0)\|_2^2 > K_5 s_0 \log p\}$. Also define $\widetilde{\mathcal{H}}_n'$ as $\widetilde{\mathcal{H}}_n$ but using a constant $\tilde{M}_2' \le \tilde{M}_2$ such that $\widetilde{\mathcal{H}}_n' \subset \widetilde{\mathcal{H}}_n$. Then, by Theorem 3, we have that

$$\begin{aligned}\mathbb{E}_0\Pi(\theta \in \mathcal{B}_n | Y^{(n)}) &\le \mathbb{E}_0\Pi(\theta \in \mathcal{B}_n \cap \widetilde{\Theta}_n, \eta \in \widetilde{\mathcal{H}}_n' | Y^{(n)}) + o(1) \\ &\le \mathbb{E}_0\Pi(\theta \in \mathcal{B}_n \cap \widetilde{\Theta}_n, \eta \in \widetilde{\mathcal{H}}_n' | Y^{(n)}, \eta \in \widetilde{\mathcal{H}}_n) + o(1).\end{aligned}$$

Let $\Omega$ be the event that is an intersection of the events in (36), (49), and the event $\{\|\tilde{X}^T(I - H)U\|_\infty \le 2\underline{\rho}_0^{-1/2}\sqrt{\log p}\|X\|_*\}$ whose probability goes to zero by Lemma 5. Since $\mathbb{P}_0(\Omega^c) \to 0$, it suffices to show that

$$\begin{aligned}&\mathbb{E}_0\Pi(\theta \in \mathcal{B}_n \cap \widetilde{\Theta}_n, \eta \in \widetilde{\mathcal{H}}_n' | Y^{(n)}, \eta \in \widetilde{\mathcal{H}}_n)\mathbb{1}_\Omega \\ &\quad = \mathbb{E}_0 \frac{\int_{\widetilde{\Theta}_n \cap \mathcal{B}_n} \int_{\widetilde{\mathcal{H}}_n'} p_{\theta,\eta}(Y^{(n)}) d\Pi(\eta) d\Pi(\theta)}{\int \int_{\widetilde{\mathcal{H}}_n} p_{\theta,\eta}(Y^{(n)}) d\Pi(\eta) d\Pi(\theta)}\end{aligned} \tag{52}$$

tends to zero. Observe that by Fubini's theorem, the denominator of the ratio is equal to

$$\begin{aligned}&\int_{\widetilde{\mathcal{H}}_n} \int \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(Y^{(n)}) d\Pi(\theta) p_{\theta_0,\eta} d\Pi(\eta) \\ &\quad \ge \left\{ \inf_{\eta \in \widetilde{\mathcal{H}}_n} \int \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(Y^{(n)}) d\Pi(\theta) \right\} \int_{\widetilde{\mathcal{H}}_n} p_{\theta_0,\eta}(Y^{(n)}) d\Pi(\eta).\end{aligned}$$

By Lemma 6, the term in the braces on the right hand side is further bounded below by $e^{-K_0'(1+s_0 \log p)}$ on the event $\Omega$. Note also that the numerator of the ratio in (52) is equal to

$$\begin{aligned}&\int_{\widetilde{\Theta}_n \cap \mathcal{B}_n} \int_{\widetilde{\mathcal{H}}_n'} \Lambda_n^*(\theta, \eta) p_{\theta_0, \tilde{\eta}_n(\theta,\eta)}(Y^{(n)}) d\Pi(\eta) d\Pi(\theta) \\ &\quad \le \left\{ \int_{\widetilde{\Theta}_n \cap \mathcal{B}_n} \Lambda_n^\star(\theta) \sup_{\eta \in \widetilde{\mathcal{H}}_n'} \frac{\Lambda_n^*(\theta, \eta)}{\Lambda_n^\star(\theta)} d\Pi(\theta) \right\} \sup_{\theta \in \widetilde{\Theta}_n \cap \mathcal{B}_n} \int_{\widetilde{\mathcal{H}}_n'} p_{\theta_0, \tilde{\eta}_n(\theta,\eta)}(Y^{(n)}) d\Pi(\eta).\end{aligned}$$

Combining the bounds, on the event $\Omega$, the ratio in (52) is bounded by

$$\begin{aligned}&e^{K_0'(1+s_0 \log p)} \sup_{\theta \in \widetilde{\Theta}_n \cap \mathcal{B}_n} \frac{\int_{\widetilde{\mathcal{H}}_n'} p_{\theta_0, \tilde{\eta}_n(\theta,\eta)}(Y^{(n)}) d\Pi(\eta)}{\int_{\widetilde{\mathcal{H}}_n} p_{\theta_0,\eta}(Y^{(n)}) d\Pi(\eta)} \\ &\quad \times \int_{\widetilde{\Theta}_n \cap \mathcal{B}_n} \Lambda_n^\star(\theta) \sup_{\eta \in \widetilde{\mathcal{H}}_n'} \frac{\Lambda_n^*(\theta, \eta)}{\Lambda_n^\star(\theta)} d\Pi(\theta).\end{aligned}$$

At the end of this proof, we will verify that

$$\sup_{\theta\in\widetilde{\Theta}_n\cap\mathcal{B}_n}\frac{\int_{\widetilde{\mathcal{H}}'_n} p_{\theta_0,\tilde{\eta}_n(\theta,\eta)}(Y^{(n)})d\Pi(\eta)}{\int_{\widetilde{\mathcal{H}}_n} p_{\theta_0,\eta}(Y^{(n)})d\Pi(\eta)}\lesssim 1, \tag{53}$$

with $\mathbb{P}_0$-probability tending to one. Assuming that this is true for now and letting $\Omega^*$ be the event satisfying (53), we see that (52) is bounded by

$$e^{K'_0(1+s_0\log p)}\mathbb{E}_0\int_{\widetilde{\Theta}_n\cap\mathcal{B}_n}\Lambda_n^\star(\theta)\sup_{\eta\in\widetilde{\mathcal{H}}'_n}\frac{\Lambda_n^*(\theta,\eta)}{\Lambda_n^\star(\theta)}d\Pi(\theta)\mathbb{1}_{\Omega\cap\Omega^*}+o(1).$$

To show that this tends to zero, for $\delta_n$ in Lemma 3, define $\mathcal{B}_{1,n}=\{\theta\in\widetilde{\Theta}_n : s_\theta > K_4 s_0, \|X(\theta-\theta_0)\|_2^2\le\delta_n^{-1/2}(s_\theta+s_0)\log p\}$, $\mathcal{B}_{2,n}=\{\theta\in\widetilde{\Theta}_n : s_\theta>K_4s_0, \|X(\theta-\theta_0)\|_2^2>\delta_n^{-1/2}(s_\theta+s_0)\log p\}$, and $\mathcal{B}_{3,n}=\{\theta\in\widetilde{\Theta}_n : s_\theta\le K_4s_0, \|X(\theta-\theta_0)\|_2^2 > K_5s_0\log p\}$ such that $\widetilde{\Theta}_n\cap\mathcal{B}_n=\cup_{k=1}^3\mathcal{B}_{k,n}$. Below we will show that

$$\begin{aligned}A(\mathcal{B}_{k,n})&\coloneqq e^{K'_0(1+s_0\log p)}\\&\qquad\times\mathbb{E}_0\int_{\mathcal{B}_{k,n}}\Lambda_n^\star(\theta)\sup_{\eta\in\mathcal{H}_n}\frac{\Lambda_n^*(\theta,\eta)}{\Lambda_n^\star(\theta)}d\Pi(\theta)\mathbb{1}_{\Omega\cap\Omega^*}\to 0,\quad k=1,2,3.\end{aligned}$$

Since $\mathbb{E}_0\Lambda_n^\star(\theta)=1$ by the moment generating function of normal distributions, we obtain that

$$\begin{aligned}A(\mathcal{B}_{1,n})&\le\mathbb{E}_0\int_{\mathcal{B}_{1,n}}\Lambda_n^\star(\theta)e^{K'_0(1+s_0\log p)+2\delta_n^{1/2}(s_\theta+s_0)\log p}d\Pi(\theta)\\&\le\pi_p(0)\sum_{s>K_4s_0}e^{K'_0(1+s_0\log p)+2\delta_n^{1/2}(s+s_0)\log p}\left(\frac{A_2}{p^{A_4}}\right)^{s-s_0}.\end{aligned}$$

If $s_0=0$, the rightmost side goes to zero for any $K_4>0$. If $s_0>0$, it still goes to zero for $K_4$ that is much larger than $K'_0$.

Note also that by conditions (C4), (C7) and (C8), we have that for some $C_1, C_2>0$ and any $\theta$,

$$\begin{aligned}\log\Lambda_n^\star(\theta)&=-\frac{1}{2}\|(I-H)\tilde{X}(\theta-\theta_0)\|_2^2+(\theta-\theta_0)^T\tilde{X}^T(I-H)U\\&\le -C_1\|X(\theta-\theta_0)\|_2^2+\|\theta-\theta_0\|_1\|\tilde{X}^T(I-H)U\|_\infty\\&\le -C_1\|X(\theta-\theta_0)\|_2^2+C_2\|X(\theta-\theta_0)\|_2\sqrt{(s_\theta+s_0)\log p},\end{aligned}\tag{54}$$

on the event $\Omega$. Hence by (36) and (54), for every $\theta\in\mathcal{B}_{2,n}$,

$$\log\left\{\Lambda_n^\star(\theta)\sup_{\eta\in\mathcal{H}_n}\frac{\Lambda_n^*(\theta,\eta)}{\Lambda_n^\star(\theta)}\right\}\le(C_2\delta_n^{1/4}+\delta_n+\delta_n^{5/4}-C_1)\|X(\theta-\theta_0)\|_2^2\le 0,$$

on the event $\Omega$. Therefore,

$$A(\mathcal{B}_{2,n})\le e^{K'_0(1+s_0\log p)}\int_{\mathcal{B}_{2,n}}d\Pi(\theta)+o(1)$$

$$\leq \pi_p(0) e^{K_0'(1+s_0 \log p)} \sum_{s > K_4 s_0} \left( \frac{A_2}{p^{A_4}} \right)^{s-s_0} + o(1).$$

This tends to zero if $K_4$ is sufficiently large.

If $s_0 = 0$, $\mathcal{B}_{3,n}$ is the empty set as it implies $\theta = \theta_0 = 0$. Hence it suffices to consider the case that $s_0 > 0$ below. By (36) and (54) again, there exists a constant $C_3 > 0$ such that for every $\theta \in \mathcal{B}_{3,n}$,

$$\begin{aligned} &\log \left\{ \Lambda_n^\star(\theta) \sup_{\eta \in \mathcal{H}_n} \frac{\Lambda_n^*(\theta,\eta)}{\Lambda_n^\star(\theta)} \right\} \\ &\quad \leq -C_1 \|X(\theta-\theta_0)\|_2^2 + \left\{ C_2 \sqrt{\frac{K_4+1}{K_5}} + \delta_n \left( 1 + \frac{1}{\sqrt{K_5}} \right) \right\} \|X(\theta-\theta_0)\|_2^2 \\ &\quad \leq -C_3 \|X(\theta-\theta_0)\|_2^2, \end{aligned}$$

on the event $\Omega$, where the last inequality holds by choosing $K_5$ much larger than $K_4$. Therefore,

$$\begin{aligned} A(\mathcal{B}_{3,n}) &\leq e^{K_0'(1+s_0\log p)} \int_{\mathcal{B}_{3,n}} e^{-C_3\|X(\theta-\theta_0)\|_2^2} d\Pi(\theta) \\ &\leq e^{K_0'(1+s_0\log p) - C_3 K_5 s_0 \log p}, \end{aligned}$$

which tends to zero for $K_5$ that is much larger than $K_0'$, if $s_0 > 0$.

It only remains to show (53). Since the map $\eta \mapsto \tilde\eta_n(\theta,\eta)$ is bijective for every fixed $\theta$, for the set defined by $\tilde\eta_n(\theta, \widetilde{\mathcal{H}}_n') = \{\tilde\eta_n(\theta,\eta) : \eta \in \widetilde{\mathcal{H}}_n'\}$ with given $\theta \in \widetilde\Theta_n$, we see that

$$\int_{\widetilde{\mathcal{H}}_n'} p_{\theta_0, \tilde\eta_n(\theta,\eta)}(Y^{(n)}) d\Pi(\eta) = \int_{\tilde\eta_n(\theta,\widetilde{\mathcal{H}}_n')} p_{\theta_0,\eta}(Y^{(n)}) d\Pi_{n,\theta}(\eta), \tag{55}$$

by the substitution in the integral. Writing $\Delta_0^*$ the block diagonal matrix formed by stacking $\Delta_{\eta_0,i}^{1/2}$, $i = 1, \dots, n$, it can be seen that

$$\tilde\eta_n(\theta, \widetilde{\mathcal{H}}_n') = \left\{ \eta \in \mathcal{H} : \sqrt{\|\Delta_0^*(\tilde\xi_\eta - \tilde\xi_0 - H\tilde X(\theta-\theta_0))\|_2^2 + d_{B,n}^2(\eta,\eta_0)} \leq \tilde M_2' \bar\epsilon_n \right\}.$$

Hence, we see that $\tilde M_2$ can be chosen sufficiently larger than $\tilde M_2'$ such that $\tilde\eta_n(\theta, \widetilde{\mathcal{H}}_n') \subset \widetilde{\mathcal{H}}_n$ for every $\theta \in \widetilde\Theta_n$ as we have $\sqrt{n} d_{A,n}(\eta,\eta_0) \lesssim \|\tilde\xi_\eta - \tilde\xi_{\eta_0} - H\tilde X(\theta-\theta_0)\|_2 + \|X(\theta-\theta_0)\|_2$. Therefore, (55) is bounded by

$$\int_{\widetilde{\mathcal{H}}_n} p_{\theta_0,\eta}(Y^{(n)}) \exp\left( \left| \log \frac{d\Pi_{n,\theta}(\eta)}{d\Pi(\eta)} \right| \right) d\Pi(\eta) \lesssim \int_{\widetilde{\mathcal{H}}_n} p_{\theta_0,\eta}(Y^{(n)}) d\Pi(\eta),$$

by (C9), since $d\Pi(\eta) = d\Pi_{n,\theta_0}(\eta)$. This verifies (53) and thus the proof is complete. □

### *A.4. Proof of Theorems 5–6*

To prove the shape approximation in Theorem 5 and the selection results in Theorem 6, we first obtain two lemmas. The first shows that the remainder of the approximation goes to zero in $\mathbb{P}_0$- probability, which is a stronger version of Lemma 3. The second implies that with a point mass prior for $\theta$ at $\theta_0$, we also obtain a rate which is not worse than that in Theorem 3.

**Lemma 7.** *Suppose that* (C1), (C4), (C8*), and (C10*) are satisfied for some orthogonal projection $H$. Then, for $\Lambda_n^*(\theta,\eta) = (p_{\theta,\eta}/p_{\theta_0,\tilde{\eta}_n(\theta,\eta)})(Y^{(n)})$ and $\Lambda_n^\star(\theta)$ in (14) with the corresponding $H$, we have that

$$\mathbb{E}_0 \sup_{\theta\in\widehat{\Theta}_n} \sup_{\eta\in\widehat{\mathcal{H}}_n} |\log \Lambda_n^*(\theta,\eta) - \log \Lambda_n^\star(\theta)| \to 0.$$

*Proof.* Similar to the proof of Lemma 3, it suffices to show the following three assertions:

$$\sup_{\theta\in\widehat{\Theta}_n} \sup_{\eta\in\widehat{\mathcal{H}}_n} \left|(\theta-\theta_0)^T\tilde{X}^T(I-H)(I-\Delta_\eta^\star)(I-H)\tilde{X}(\theta-\theta_0)\right| \to 0, \tag{56}$$

$$\sup_{\theta\in\widehat{\Theta}_n} \sup_{\eta\in\widehat{\mathcal{H}}_n} \left|(\theta-\theta_0)^T\tilde{X}^T(I-H)\Delta_\eta^\star(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0}+H\tilde{X}(\theta-\theta_0))\right| \to 0, \tag{57}$$

$$\mathbb{E}_0 \sup_{\theta\in\widehat{\Theta}_n} \sup_{\eta\in\widehat{\mathcal{H}}_n} \left|(\theta-\theta_0)^T\tilde{X}^T(I-H)(I-\Delta_\eta^\star)U\right| \to 0. \tag{58}$$

First, note that the left side of (56) is bounded above by a constant multiple of

$$\begin{aligned}&\sup_{\theta\in\widehat{\Theta}_n} \sup_{\eta\in\widehat{\mathcal{H}}_n} \|I-\Delta_\eta^\star\|_{\text{sp}}\|\tilde{X}(\theta-\theta_0)\|_2^2 \\ &\quad\lesssim \sup_{\theta\in\widehat{\Theta}_n}\|X(\theta-\theta_0)\|_2^2 \sup_{\eta\in\widehat{\mathcal{H}}_n} \max_{1\le i\le n}\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\text{F}},\end{aligned} \tag{59}$$

where the inequality holds by (42) and the fact that $\max_i\|\Delta_{\eta,i}-\Delta_{\eta_0,i}\|_{\text{F}}^2 \le e_n + a_n d_{B,n}^2(\eta,\eta_0) \lesssim e_n + a_n(s_\star \log p)/n = o(1)$ on $\widehat{\mathcal{H}}_n$. We see that (59) is bounded above by a constant multiple of

$$\sup_{\theta\in\widehat{\Theta}_n}\|X\|_*\|\theta-\theta_0\|_1^2 \sup_{\eta\in\widehat{\mathcal{H}}_n}\sqrt{e_n+a_n d_{B,n}^2(\eta,\eta_0)} \lesssim s_\star^2\log p\sqrt{e_n+\frac{a_n s_\star\log p}{n}},$$

which goes to zero by (C10*).

Next, similar to (43), the left side of (57) is bounded by

$$\begin{aligned}&\sup_{\theta\in\widehat{\Theta}_n}\|X(\theta-\theta_0)\|_2 \sup_{\eta\in\widehat{\mathcal{H}}_n}\|(I-H)(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0})\|_2 \\ &+ \sup_{\theta\in\widehat{\Theta}_n} \sup_{\eta\in\widehat{\mathcal{H}}_n}\Big\{\Big(\|X(\theta-\theta_0)\|_2^2+\|X(\theta-\theta_0)\|_2\sqrt{n}d_{A,n}(\eta,\eta_0)\Big) \\ &\qquad\qquad \times \max_{1\le i\le n}\|\Delta_{\eta,i}^{-1}-\Delta_{\eta_0,i}^{-1}\|_{\text{sp}}\Big\}.\end{aligned}$$

Using the same approach used in (42), the display is further bounded above by a constant multiple of

$$s_\star\sqrt{\log p}\sup_{\eta\in\widehat{\mathcal{H}}_n}\|(I-H)(\tilde{\xi}_\eta-\tilde{\xi}_{\eta_0})\|_2+s_\star^2\log p\sqrt{e_n+\frac{a_n s_\star\log p}{n}},$$

which goes to zero by (C8*) and (C10*).

Now, using Lemma 4, note that (58) is bounded above by

$$\begin{aligned}&\sup_{\theta\in\widehat{\Theta}_n}\|\theta-\theta_0\|_1\,\mathbb{E}_0\sup_{\eta\in\widehat{\mathcal{H}}_n}\|\tilde{X}^T(I-H)(I-\Delta_\eta^\star)U\|_\infty\\ &\quad\lesssim s_\star\log p\left\{\sqrt{e_n+\frac{a_n s_\star\log p}{n}}\right.\\ &\qquad\qquad\left.+\sqrt{a_n}\int_0^{C_1\sqrt{(s_\star\log p)/n}}\sqrt{\log N(\delta,\widehat{\mathcal{H}}_n,d_{B,n})}d\delta\right\},\end{aligned}$$

for some $C_1>0$. This tends to zero by (C10*). □

**Lemma 8.** *Suppose that* (C1)–(C4), (C5*), *and* (C6) *are satisfied. Then there exists a constant $K_6>0$ such that*

$$\mathbb{E}_0\Pi^{\theta_0}\left(d_n(\eta,\eta_0)>K_6\bar{\epsilon}_n\,\middle|\,Y^{(n)}\right)\to 0,$$

*where $\Pi^{\theta_0}(\cdot\,|\,Y^{(n)})$ is the posterior distribution induced by the point mass prior for $\theta$ at $\theta_0$, i.e., $\delta_{\theta_0}(\theta)$, in place of the prior in* (4).

*Proof.* Since the prior for $\theta$ is the point mass at $\theta_0$, we can reduce to a low dimensional model $Y_i^*:=Y_i-X_i\theta_0=\xi_{\eta,i}+\varepsilon_i$, $i=1,\dots,n$. Then the lemma can be easily verified using the main results on posterior contraction in Section 3. The denominator of the posterior distribution with the Dirac prior at $\theta_0$ is bounded as in Lemma 1, which can be shown using (20) for the prior concentration condition (C2) and the expressions for the Kullback-Leibler divergence $K(p_{0,i},p_{\theta_0,\eta,i})$ and variation $V(p_{0,i},p_{\theta_0,\eta,i})$ with the true value $\theta_0$. For a local test relative to the average Rényi divergence, Lemma 2 applied with $\mathcal{F}_{1,n}$, modified so that it can be involved only with a given $\eta_1$ such that $R_n(p_0,p_{\theta_0,\eta_1})\ge\bar{\epsilon}_n^2$, implies that a small piece of the alternative is tested with exponentially small errors. Hence, by (C5*), we obtain the contraction rate $\bar{\epsilon}_n^2$ relative to $R_n(p_0,p_{\theta_0,\eta})$ for $\Pi^{\theta_0}(\cdot\,|\,Y^{(n)})$, as in the proof of Theorem 2. The lemma is then obtained by recovering the contraction rate of $\eta$ with respect to $d_n$ using the approach in the proof of Theorem 3. □

*Proof of Theorem 5.* Our proof is based on the proof of Theorem 6 in Castillo et al. [8], but is more involved due to $\eta$. We use the fact that for any probability measure $Q$ and its renormalized restriction $Q_{\mathcal{A}}(\cdot)=Q(\cdot\cap\mathcal{A})/Q(\mathcal{A})$ to a set $\mathcal{A}$, we have $\|Q-Q_{\mathcal{A}}\|_{\mathrm{TV}}\le 2Q(\mathcal{A}^c)$. First, using a sufficiently large constant $\hat{M}_2'$ that is smaller than $\hat{M}_2$, define $\widehat{\mathcal{H}}_n'$ as $\widehat{\mathcal{H}}_n$ in (12) such that $\widehat{\mathcal{H}}_n'\subset\widehat{\mathcal{H}}_n$. Let $\widetilde{\Pi}((\theta,\eta)\in\cdot)$

be the prior distribution restricted and renormalized on $\widehat{\Theta}_n \times \widehat{\mathcal{H}}'_n$ and $\widetilde{\Pi}((\theta,\eta) \in \cdot \,|\, Y^{(n)})$ be the corresponding posterior distribution. Also, $\widetilde{\Pi}^\infty(\theta \in \cdot \,|\, Y^{(n)})$ is the restricted and renormalized version of $\Pi^\infty(\theta \in \cdot \,|\, Y^{(n)})$ to the set $\widehat{\Theta}_n$. Then the left hand side of the theorem is bounded above by

$$\begin{aligned}
&\left\| \Pi(\theta \in \cdot \,|\, Y^{(n)}) - \widetilde{\Pi}(\theta \in \cdot \,|\, Y^{(n)}) \right\|_{\mathrm{TV}} + \left\| \widetilde{\Pi}(\theta \in \cdot \,|\, Y^{(n)}) - \widetilde{\Pi}^\infty(\theta \in \cdot \,|\, Y^{(n)}) \right\|_{\mathrm{TV}} \\
&+ \left\| \Pi^\infty(\theta \in \cdot \,|\, Y^{(n)}) - \widetilde{\Pi}^\infty(\theta \in \cdot \,|\, Y^{(n)}) \right\|_{\mathrm{TV}},
\end{aligned} \tag{60}$$

where the first summand goes to zero in $\mathbb{P}_0$-probability since $\Pi((\theta,\eta) \in \widehat{\Theta}_n \times \widehat{\mathcal{H}}'_n \,|\, Y^{(n)}) \to 1$ in $\mathbb{P}_0$-probability by Theorem 1 and Theorem 3.

To show that the second summand goes to zero in $\mathbb{P}_0$-probability, note that for every measurable $\mathcal{B} \subset \mathbb{R}^p$, we obtain

$$\begin{aligned}
\widetilde{\Pi}(\theta \in \mathcal{B} \,|\, Y^{(n)}) &\propto \int_{\mathcal{B}\cap\widehat{\Theta}_n} \int_{\widehat{\mathcal{H}}'_n} p_{\theta,\eta}(Y^{(n)})\, e^{-\lambda\|\theta\|_1} d\Pi(\eta) dV(\theta) \\
&= \int_{\mathcal{B}\cap\widehat{\Theta}_n} \int_{\widehat{\mathcal{H}}'_n} \Lambda_n^*(\theta,\eta)\, e^{-\lambda\|\theta\|_1} p_{\theta_0,\tilde{\eta}_n(\theta,\eta)}(Y^{(n)})\, d\Pi(\eta) dV(\theta), \\
\widetilde{\Pi}^\infty(\theta \in \mathcal{B} \,|\, Y^{(n)}) &\propto \int_{\mathcal{B}\cap\widehat{\Theta}_n} \Lambda_n^\star(\theta) dV(\theta) \\
&\propto \int_{\mathcal{B}\cap\widehat{\Theta}_n} \Lambda_n^\star(\theta)\, e^{-\lambda\|\theta_0\|_1} \int_{\mathcal{H}} p_{\theta_0,\eta}(Y^{(n)}) d\Pi(\eta) dV(\theta),
\end{aligned}$$

where $dV(\theta) = \sum_{S: s \le K_1 s_\star} \pi_p(s) \binom{p}{s}^{-1} (\lambda/2)^s d\{\mathcal{L}(\theta_S) \otimes \delta_0(\theta_{S^c})\}$. In the last line, the factor $e^{-\lambda\|\theta_0\|_1} \int_{\mathcal{H}} p_{\theta_0,\eta}(Y^{(n)}) d\Pi(\eta)$ cancels out in the normalizing constant, but is inserted for the sake of comparison. For any sequences of measures $\{\mu_S\}$ and $\{\nu_S\}$, if $\nu_S$ is absolutely continuous with respect to $\mu_S$ with the Radon-Nikodym derivative $d\nu_S/d\mu_S$, then it can be easily verified that

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

Hence, for $C_n = \int_{\mathcal{H}} p_{\theta_0,\eta}(Y^{(n)}) d\Pi(\eta)$, we see that the second summand of (60) is bounded by

$$2 \sup_{\theta\in\widehat{\Theta}_n} \left| 1 - \frac{1}{C_n} \int_{\widehat{\mathcal{H}}'_n} \frac{\Lambda_n^*(\theta,\eta) e^{-\lambda\|\theta\|_1}}{\Lambda_n^\star(\theta) e^{-\lambda\|\theta_0\|_1}} p_{\theta_0,\tilde{\eta}_n(\theta,\eta)}(Y^{(n)}) d\Pi(\eta) \right| .$$

Using the fact that $|\lambda(\|\theta\|_1 - \|\theta_0\|_1)| \le \lambda\|\theta - \theta_0\|_1 \lesssim \lambda s_\star \sqrt{\log p}/\|X\|_* \to 0$ on $\widehat{\Theta}_n$ and that $\sup\{|1 - \Lambda_n^*(\theta,\eta)/\Lambda_n^\star(\theta)| : \theta \in \widehat{\Theta}_n, \eta \in \widehat{\mathcal{H}}'_n\}$ goes to zero in $\mathbb{P}_0$-probability by Lemma 7, the last display is further bounded by

$$2 \sup_{\theta\in\widehat{\Theta}_n} \left| 1 - \{1 + o(1) + o_{\mathbb{P}_0}(1)\} \frac{1}{C_n} \int_{\widehat{\mathcal{H}}'_n} p_{\theta_0,\tilde{\eta}_n(\theta,\eta)}(Y^{(n)}) d\Pi(\eta) \right| . \tag{61}$$

Now, note that the map $\eta \mapsto \tilde{\eta}_n(\theta,\eta)$ is bijective for every fixed $\theta \in \widehat{\Theta}_n$. Thus for the set defined by $\tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n') = \{\tilde{\eta}_n(\theta,\eta) : \eta \in \widehat{\mathcal{H}}_n'\}$ with given $\theta \in \widehat{\Theta}_n$, we see that

$$\int_{\widehat{\mathcal{H}}_n'} p_{\theta_0,\tilde{\eta}_n(\theta,\eta)}(Y^{(n)})d\Pi(\eta) = \int_{\tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n')} p_{\theta_0,\eta}(Y^{(n)})d\Pi_{n,\theta}(\eta), \tag{62}$$

by the substitution in the integral. Similar to the proof of Theorem 4, observe that

$$\begin{aligned}\tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n') = \Big\{\eta \in \mathcal{H} : \|\Delta_0^*(\tilde{\xi}_\eta - \tilde{\xi}_0 - H\tilde{X}(\theta-\theta_0))\|_2 \le \hat{M}_2' s_\star\sqrt{(\log p)/n},\\ d_{B,n}(\eta,\eta_0) \le \hat{M}_2'\sqrt{(s_\star \log p)/n}\Big\}.\end{aligned}$$

Hence, we see that $\hat{M}_2$ can be chosen sufficiently large such that $\tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n') \subset \widehat{\mathcal{H}}_n$ for every $\theta \in \widehat{\Theta}_n$ as we have $\sqrt{n}d_{A,n}(\eta,\eta_0) \lesssim \|\tilde{\xi}_\eta - \tilde{\xi}_{\eta_0} - H\tilde{X}(\theta-\theta_0)\|_2 + \|X\|_*\|\theta-\theta_0\|_1$. Therefore, since $d\Pi(\eta) = d\Pi_{n,\theta_0}(\eta)$, one can see that (62) is written as

$$\{1+o(1)\}\int_{\tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n')} p_{\theta_0,\eta}(Y^{(n)})d\Pi(\eta),$$

by (C9*), and hence (61) is equal to

$$2\sup_{\theta\in\widehat{\Theta}_n}\left|1 - \{1+o_{\mathbb{P}_0}(1)\}\frac{\int_{\tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n')} p_{\theta_0,\eta}(Y^{(n)})d\Pi(\eta)}{\int_{\mathcal{H}} p_{\theta_0,\eta}(Y^{(n)})d\Pi(\eta)}\right|. \tag{63}$$

Now, observe that we also have the inequality of the other direction: $\|\tilde{\xi}_\eta - \tilde{\xi}_{\eta_0} - H\tilde{X}(\theta-\theta_0)\|_2 \lesssim \sqrt{n}d_{A,n}(\eta,\eta_0) + \|X\|_*\|\theta-\theta_0\|_1$. This means that $\hat{M}_2'$ can be chosen sufficiently large such that $\{\eta \in \mathcal{H} : d_n(\eta,\eta_0) \le K_6\bar{\epsilon}_n\} \subset \tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n')$ for every $\theta \in \widehat{\Theta}_n$. Hence, with appropriately chosen constants, we obtain

$$\begin{aligned}\inf_{\theta\in\widehat{\Theta}_n}\frac{\int_{\tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n')} p_{\theta_0,\eta}(Y^{(n)})d\Pi(\eta)}{\int_{\mathcal{H}} p_{\theta_0,\eta}(Y^{(n)})d\Pi(\eta)} &= \inf_{\theta\in\widehat{\Theta}_n}\Pi^{\theta_0}\left(\eta \in \tilde{\eta}_n(\theta,\widehat{\mathcal{H}}_n')\,|\,Y^{(n)}\right)\\ &\ge \Pi^{\theta_0}\left(d_n(\eta,\eta_0) \le K_6\bar{\epsilon}_n\,\middle|\,Y^{(n)}\right).\end{aligned}$$

The rightmost term goes to one with probability tending to one by Lemma 8. This implies that (63) goes to zero in $\mathbb{P}_0$-probability, completing the proof for the second part of (60).

Next, we show that $\Pi^\infty(\theta \in \widehat{\Theta}_n\,|\,Y^{(n)})$ goes to one in $\mathbb{P}_0$-probability to verify that the last summand in (60) goes to zero in $\mathbb{P}_0$-probability. Observe that $\Pi^\infty(\theta \in \widehat{\Theta}_n^c\,|\,Y^{(n)})$ is equal to

$$\frac{\int_{\widehat{\Theta}_n^c}\exp\left\{-\frac{1}{2}\|(I-H)\tilde{X}(\theta-\theta_0)\|_2^2 + U^T(I-H)\tilde{X}(\theta-\theta_0)\right\}dV(\theta)}{\int_{\mathbb{R}^p}\exp\left\{-\frac{1}{2}\|(I-H)\tilde{X}(\theta-\theta_0)\|_2^2 + U^T(I-H)\tilde{X}(\theta-\theta_0)\right\}dV(\theta)}. \tag{64}$$

Clearly, the denominator is bounded below by

$$\frac{\pi_p(s_0)}{\binom{p}{s_0}}\left(\frac{\lambda}{2}\right)^{s_0}\int_{\mathbb{R}^{s_0}}\exp\Big\{-\frac{1}{2}\|(I-H)\tilde{X}_{S_0}(\theta_{S_0}-\theta_{0,S_0})\|_2^2 \\ +U^T(I-H)\tilde{X}_{S_0}(\theta_{S_0}-\theta_{0,S_0})\Big\}d\theta_{S_0}. \tag{65}$$

Since the measure $Q$ defined by $Q(d\theta_{S_0}) = \exp\{-(1/2)\|(I-H)\tilde{X}_{S_0}(\theta_{S_0} - \theta_{0,S_0})\|_2^2\}$ is symmetric about $\theta_{0,S_0}$, the mean of $(\theta_{S_0} - \theta_{0,S_0})$ with respect to the normalized probability measure $\widetilde{Q} = Q/Q(\mathbb{R}^{s_0})$ is zero. Note also that $\Gamma_S = \tilde{X}_S^T(I-H)\tilde{X}_S$ is nonsingular for every $S$ such that $s \le K_1 s_\star$ by (C8*). Thus, by Jensen's inequality, (65) is bounded below by

$$\frac{\pi_p(s_0)}{\binom{p}{s_0}}\left(\frac{\lambda}{2}\right)^{s_0}\int_{\mathbb{R}^{s_0}}\exp\left\{-\frac{1}{2}\|(I-H)\tilde{X}_{S_0}(\theta_{S_0}-\theta_{0,S_0})\|_2^2\right\}d\theta_{S_0} \\ = \frac{\pi_p(s_0)}{\binom{p}{s_0}}\left(\frac{\lambda}{2}\right)^{s_0}\frac{(2\pi)^{s_0/2}}{\det(\Gamma_{S_0})^{1/2}}.$$

Applying the arithmetic-geometric mean inequality to the eigenvalues, we obtain $\det(\Gamma_{S_0}) \le (\operatorname{tr}(\Gamma_{S_0})/s_0)^{s_0} \le \|(I-H)\tilde{X}_{S_0}\|_*^{2s_0} \le \underline{\rho}_0^{-s_0}\|X\|_*^{2s_0}$, and hence $\det(\Gamma_{S_0})^{1/2}/\lambda^{s_0} \le \underline{\rho}_0^{-s_0/2}(L_1p^{L_2})^{s_0}$ by (4). Furthermore, we have $\pi_p(s_0) \gtrsim A_1^{s_0}p^{-A_3s_0}$ by (3) and $\binom{p}{s_0} \le p^{s_0}$. Hence, the preceding display is further bounded below by a constant multiple of

$$p^{-(1+L_2+A_3)s_0}\left(\frac{A_1\sqrt{\underline{\rho}_0\pi}}{L_1\sqrt{2}}\right)^{s_0}. \tag{66}$$

To bound the numerator of (64), let $D_n = 2\underline{\rho}_0^{-1/2}\sqrt{\log p}\|X\|_*$ and $\mathcal{U}_n = \{\|\tilde{X}^T(I-H)U\|_\infty \le D_n\}$. Then it suffices to show that (64) goes to zero in $\mathbb{P}_0$-probability on the set $\mathcal{U}_n$ as $\mathbb{P}_0(\mathcal{U}_n^c) \to 0$ by Lemma 5. Note that on the set $\mathcal{U}_n$ we have

$$U^T(I-H)\tilde{X}(\theta-\theta_0) \le D_n\|\theta-\theta_0\|_1 \\ \le D_n\frac{2\sqrt{\overline{\rho}_0}\|\tilde{X}(\theta-\theta_0)\|_2|S_{\theta-\theta_0}|^{1/2}}{\|X\|_*\phi_1(|S_{\theta-\theta_0}|)} - D_n\|\theta-\theta_0\|_1.$$

Using that $\|u\|_2 \lesssim \|(I-H)u\|_2$ for every $u \in \operatorname{span}(\tilde{X}_S)$ with $s \le K_1 s_\star$ by (C8*), the preceding display is, for some constant $C_1 > 0$, further bounded above by

$$D_n\frac{2\sqrt{\overline{\rho}_0}C_1\|(I-H)\tilde{X}(\theta-\theta_0)\|_2|S_{\theta-\theta_0}|^{1/2}}{\|X\|_*\phi_1(|S_{\theta-\theta_0}|)} - D_n\|\theta-\theta_0\|_1 \\ \le \frac{1}{2}\|(I-H)\tilde{X}(\theta-\theta_0)\|_2^2 + \frac{2\overline{\rho}_0C_1^2D_n^2|S_{\theta-\theta_0}|}{\|X\|_*^2\phi_1(|S_{\theta-\theta_0}|)^2} - D_n\|\theta-\theta_0\|_1,$$

by the Cauchy-Schwarz inequality. We have $s_{\theta-\theta_0} \le K_1 s_\star + s_0$ on the support of the measure $V$. Hence, on the event $\mathcal{U}_n$, the numerator of (64) is bounded above by

$$\begin{aligned}
&\exp\left\{\frac{2\overline{\rho}_0 C_1^2 D_n^2 (K_1 s_\star + s_0)}{\|X\|_*^2 \phi_1(K_1 s_\star + s_0)^2} - \frac{\hat{M}_1 D_n s_\star \sqrt{\log p}}{2\|X\|_*}\right\} \\
&\times \sum_{S: s \le K_1 s_\star} \frac{\pi_p(s)}{\binom{p}{s}} \int \left(\frac{\lambda}{2}\right)^s e^{-(D_n/2)\|\theta_S - \theta_{0,S}\|_1} d\theta_S \\
&\quad \le \exp\left\{\frac{8\overline{\rho}_0 C_1^2 (K_1+1) s_\star \log p}{\underline{\rho}_0 \phi_1(K_1 s_\star + s_0)^2} - \frac{\hat{M}_1 s_\star \log p}{\sqrt{\underline{\rho}_0}}\right\} \sum_{s=0}^{p} \pi_p(s) \left(L_3 \sqrt{\frac{\underline{\rho}_0}{n}}\right)^s,
\end{aligned}$$

since $D_n/2 \ge \lambda\sqrt{n}/(L_3\sqrt{\underline{\rho}_0})$. Note that we have

$$\sum_{s=0}^{p} \pi_p(s) \left(L_3 \sqrt{\frac{\underline{\rho}_0}{n}}\right)^s \lesssim \sum_{s=0}^{p} \left(\frac{A_2 L_3}{p^{A_4}} \sqrt{\frac{\underline{\rho}_0}{n}}\right)^s \lesssim 1,$$

by (3) and that $\phi_1(K_1 s_\star + s_0)$ in the denominators is bounded away from zero by the assumption. Thus, the last display combined with (66) shows that (64) goes to zero on the event $\mathcal{U}_n$, provided that $\hat{M}_1$ is chosen sufficiently large.

Finally we conclude that (60) goes to zero in $\mathbb{P}_0$-probability. Since the total variation metric is bounded by 2, the convergence in mean holds as in the assertion. □

*Proof of Theorem 6.* Our proof follows the proof of Theorem 4 in Castillo et al. [8]. Since $\mathbb{E}_0\|\Pi(\theta \in \cdot \,|\, Y^{(n)}) - \Pi^\infty(\theta \in \cdot \,|\, Y^{(n)})\|_{\mathrm{TV}}$ tends to zero by Theorem 5, it suffices to show that $\mathbb{E}_0\Pi^\infty(\theta : S_\theta \in \mathcal{S}_n \,|\, Y^{(n)}) \to 0$ for $\mathcal{S}_n = \{S : s \le K_1 s_\star, S \supset S_0, S \ne S_0\}$. For the orthogonal projection defined by $\tilde{H}_S = (I - H)\tilde{X}_S \Gamma_S^{-1} \tilde{X}_S^T (I - H)$ with $\Gamma_S = \tilde{X}_S^T (I - H)\tilde{X}_S$, we see that $\Pi^\infty(\theta : S_\theta \in \mathcal{S}_n \,|\, Y^{(n)})$ is bounded by

$$\sum_{s=s_0+1}^{K_1 s_\star} \frac{\pi_p(s)\binom{p}{s_0}\binom{p-s_0}{s-s_0}}{\pi_p(s_0)\binom{p}{s}} \left(\frac{\lambda\sqrt{\pi}}{\sqrt{2}}\right)^{s-s_0} \max_{S \in \mathcal{S}_n : |S| = s} \left[\frac{\det(\Gamma_{S_0})^{1/2}}{\det(\Gamma_S)^{1/2}} e^{\|(\tilde{H}_S - \tilde{H}_{S_0})U\|_2^2/2}\right],$$

by (13), since $(\tilde{H}_S - \tilde{H}_{S_0})\tilde{X}\theta_0 = (\tilde{H}_S - \tilde{H}_{S_0})(I - H)\tilde{X}_{S_0}\theta_{0,S_0} = 0$ for every $S \in \mathcal{S}_n$ due to $S_0 \subset S$ on $\mathcal{S}_n$. Note that $\rho_k(\Gamma_{S_0}) \le \rho_k(\Gamma_S)$ for $k = 1, \dots, s_0$, because $\Gamma_{S_0}$ is a principal submatrix of $\Gamma_S$. Hence, $\det(\Gamma_{S_0})$ is equal to

$$\prod_{k=1}^{s_0} \rho_k(\Gamma_{S_0}) \le \prod_{k=1}^{s_0} \rho_k(\Gamma_S) \le \frac{\det(\Gamma_S)}{\rho_{\min}(\Gamma_S)^{s-s_0}} \le \frac{\det(\Gamma_S)}{(C_1 \overline{\rho}_0^{-1/2} \phi_2(s) \|X\|_*)^{2(s-s_0)}}, \tag{67}$$

for some $C_1 > 0$. The last inequality holds since by (C8*), there exists a constant $C_1 > 0$ such that $C_1^2\|v\|_2^2 \le \|(I - H)v\|_2^2$ for every $v \in \mathrm{span}(\tilde{X}_S)$ with $s \le K_1 s_\star$, and hence we have that by the definition of $\phi_2$,

$$\rho_{\min}(\Gamma_S) = \inf_{u\in\mathbb{R}^s, u\neq 0} \frac{\|(I-H)\tilde{X}_S u\|_2^2}{\|u\|_2^2} \geq \frac{C_1^2\phi_2(s)^2\|X\|_*^2}{\overline{\rho}_0}.$$

Now, we shall show that for any fixed $b>2$,

$$\mathbb{P}_0\left(\|(\tilde{H}_S - \tilde{H}_{S_0})U\|_2^2 \leq b(s-s_0)\log p, \text{ for every } S\in\mathcal{S}_n\right) \to 1. \tag{68}$$

Note that $\|(\tilde{H}_S - \tilde{H}_{S_0})U\|_2^2$ has a chi-squared distribution with degree of freedom $s-s_0$. Therefore, by Lemma 5 of Castillo et al. [8], there exists a constant $C_2$ such that for every $b>2$ and given $s\geq s_0+1$,

$$\mathbb{P}_0\left(\max_{S\in\mathcal{S}_n:|S|=s}\|(\tilde{H}_S - \tilde{H}_{S_0})U\|_2^2 > b\log N_s\right) \leq \left(\frac{1}{N_s}\right)^{(b-2)/4} e^{C_2(s-s_0)},$$

where $N_s = \binom{p-s_0}{s-s_0}$ is the cardinality of the set $\{S\in\mathcal{S}_n : |S| = s\}$. Since $N_s \leq (p-s_0)^{s-s_0} \leq p^{s-s_0}$, for $\mathcal{T}_n$ the event in the relation (68), it follows that

$$\mathbb{P}_0(\mathcal{T}_n^c) \leq \sum_{s=s_0+1}^{K_1 s_\star} \left(\frac{1}{N_s}\right)^{(b-2)/4} e^{C_2(s-s_0)}.$$

This goes to zero as $p\to\infty$, since for $s\leq K_1 s_\star$,

$$N_s \geq \frac{(p-s)^{s-s_0}}{(s-s_0)!} \geq \frac{(p-K_1 s_\star)^{s-s_0}}{(s-s_0)^{s-s_0}} \geq \left(\frac{p-K_1 s_\star}{K_1 s_\star}\right)^{s-s_0},$$

and $s_\star/p = o(1)$. To complete the proof, it remains to show that $\Pi^\infty(\theta : S_\theta \in \mathcal{S}_n \,|\, Y^{(n)})$ goes to zero on the set $\mathcal{T}_n$. Combining (67) and (68), we see that $\Pi^\infty(\theta : S_\theta \in \mathcal{S}_n \,|\, Y^{(n)})\mathbb{1}_{\mathcal{T}_n}$ is bounded by

$$\begin{aligned}
&\sum_{s=s_0+1}^{K_1 s_\star} \frac{\pi_p(s)\binom{p}{s_0}\binom{p-s_0}{s-s_0}}{\pi_p(s_0)\binom{p}{s}} \left(\frac{\lambda\sqrt{\pi}}{\sqrt{2}}\right)^{s-s_0} \left(\frac{\sqrt{\overline{\rho}_0 p^b}}{C_1\phi_2(s)\|X\|_*}\right)^{s-s_0} \\
&\leq \sum_{s=s_0+1}^{K_1 s_\star} \left(\frac{A_2}{p^{A_4}}\right)^{s-s_0} \binom{s}{s_0} \left(\frac{L_3}{C_1\phi_1(K_1 s_\star)}\sqrt{\frac{K_1 s_\star \pi\overline{\rho}_0 p^b}{2n}}\right)^{s-s_0},
\end{aligned}$$

which holds by the inequalities $\pi_p(s)/\pi_p(s_0) \leq (A_2 p^{-A_4})^{s-s_0}$ and $\binom{p}{s_0}\binom{p-s_0}{s-s_0}/\binom{p}{s} = \binom{s}{s_0}$. Note that for $s\leq K_1 s_\star$, we have that $\binom{s}{s_0} = \binom{s}{s-s_0} \leq (K_1 s_\star)^{s-s_0} \leq (K_1 C_2 p^a)^{s-s_0}$ for some $C_2>0$. Hence, the preceding display goes to zero provided that $a - A_4 + b/2 < 0$ since $s_\star = o(n)$. This condition can be translated to $a < A_4 - 1$ by choosing $b$ arbitrarily close to 2. □

## Appendix B: Proofs for the applications

### *B.1. Proof of Theorem 7*

We first verify the conditions for Theorem 3 to prove assertions (a) and (b).

- *Verification of* (C1): Let $\bar{\sigma}_{jk}$ be the $(j,k)$th element of $\Sigma - \Sigma_0$. Observe that $d_n^2(\Sigma, \Sigma_0)$ is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\|E_i^T(\Sigma - \Sigma_0)E_i\|_{\mathrm{F}}^2 = \frac{1}{n}\sum_{j=1}^{\overline{m}}\sum_{k=1}^{\overline{m}}\left[\bar{\sigma}_{jk}^2\sum_{i=1}^{n}e_{ij}e_{ik}\right] \gtrsim \frac{1}{c_n}\|\Sigma - \Sigma_0\|_{\mathrm{F}}^2. \quad (69)$$

  Hence, we see that $c_n$ has the same role as $a_n$. We also have $e_n = 0$ as the true $\Sigma_0$ belongs to the support of the prior.
- *Verification of* (C2): Note that

$$d_n^2(\Sigma_1, \Sigma_2) = \frac{1}{n}\sum_{i=1}^{n}\|E_i^T(\Sigma_1 - \Sigma_2)E_i\|_{\mathrm{F}}^2 \le \|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}^2, \quad (70)$$

  for every $\Sigma_1, \Sigma_2 \in \mathcal{H}$. Hence we obtain that for every $\bar{\epsilon}_n > n^{-1/2}$,

$$\log \Pi(d_n(\Sigma, \Sigma_0) \le \bar{\epsilon}_n) \ge \log \Pi(\|\Sigma - \Sigma_0\|_{\mathrm{F}} \le \bar{\epsilon}_n) \gtrsim \log \bar{\epsilon}_n \gtrsim -\log n,$$

  since $1 \lesssim \rho_{\min}(\Sigma_0) \le \rho_{\max}(\Sigma_0) \lesssim 1$. This leads us to choose $\bar{\epsilon}_n = \sqrt{(\log n)/n}$ for (C2) to be satisfied.
- *Verification of* (C3): The assumption $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$ given in the theorem directly satisfies (C3).
- *Verification of* (C4): We have the inequalities $\rho_{\min}(\Sigma_0) \le \rho_{\min}(E_i^T\Sigma_0 E_i) \le \rho_{\max}(E_i^T\Sigma_0 E_i) \le \rho_{\max}(\Sigma_0)$ for every $i \le n$ as $E_i^T\Sigma_0 E_i$ is a principal submatrix of $\Sigma_0$. Hence (C4) is directly satisfied by the assumption on $\Sigma_0$.
- *Verification of* (C5*): For a sufficiently large $M > 0$ and $s_\star = s_0 \vee (\log n/\log p)$, choose $\mathcal{H}_n = \{\Sigma : n^{-M} \le \rho_{\min}(\Sigma) \le \rho_{\max}(\Sigma) \le e^{Ms_\star \log p}\}$. Since $E_i^T\Sigma E_i$ is a principal submatrix of $\Sigma$, we have $\rho_{\min}(E_i^T\Sigma E_i) \ge \rho_{\min}(\Sigma) \ge n^{-M}$ for every $i \le n$ and $\Sigma \in \mathcal{H}_n$. Hence the minimum eigenvalue condition (6) is satisfied with $\log \gamma_n \asymp \log n$. Also, the entropy relative to $d_n$ is given by

$$\begin{aligned}&\log N\left(\frac{1}{6\overline{m}n^{M+3/2}}, \mathcal{H}_n, d_n\right)\\ &\quad \le \log N\left(\frac{1}{6\overline{m}n^{M+3/2}}, \left\{\Sigma : \|\Sigma\|_{\mathrm{F}} \le \sqrt{\overline{m}}e^{Ms_\star \log p}\right\}, \|\cdot\|_{\mathrm{F}}\right)\\ &\quad \lesssim \log n + s_\star \log p.\end{aligned}$$

  The entropy condition in (7) is thus satisfied if we choose $\epsilon_n = \sqrt{(s_\star \log p)/n}$. To verify the sieve condition (8), note that for some positive constants $b_1$, $b_2$, $b_3$, $b_4$ and $b_5$, an inverse Wishart distribution satisfies

$$\begin{aligned}\Pi(\Sigma : \rho_{\min}(\Sigma) < n^{-M}) &\le b_1 e^{-b_2 n^{b_3 M}},\\ \Pi(\Sigma : \rho_{\max}(\Sigma) > e^{Ms_\star \log p}) &\le b_4 e^{-b_5 M s_\star \log p};\end{aligned} \quad (71)$$

  see, for example, Lemma 9.16 of Ghosal and van der Vaart [17]. The sieve condition (8) is met provided that $M$ is chosen sufficiently large. Note that the condition $a_n\epsilon_n^2 \to 0$ is satisfied by the assumption $c_n s_\star \log p = o(n)$.

- *Verification of* (C6): The separability condition is trivially satisfied in this example as there is no nuisance mean part.

Therefore, the contraction properties in Theorem 3 are obtained with $s_\star = s_0 \vee (\log n/\log p)$, but $s_\star$ is replaced by $s_0$ since $s_0 > 0$ and $\log n \lesssim \log p$. The contraction rate for $\Sigma$ with respect to the Frobenius norm follows from (69). The optimal posterior contraction directly follows from Corollary 1. Assertions (a) and (b) are thus proved.

Next, we verify conditions (C8*)–(C10*) and (C11) to apply Theorems 5–6 and Corollaries 2–3.

- *Verification of* (C8*)–(C9*): These conditions are trivially satisfied with the zero matrix $H$ as there is no nuisance mean part.
- *Verification of* (C10*): Since the entropy in (C10*) is bounded above by a constant multiple of $\log N(\delta, \{\Sigma : \|\Sigma - \Sigma_0\|_{\mathrm{F}} \le \hat{M}_2\sqrt{c_n}\epsilon_n\}, \|\cdot\|_{\mathrm{F}}) \lesssim 0 \vee \log(3\hat{M}_2\sqrt{c_n}\epsilon_n/\delta)$ using (69) and (70), the term in (C10*) is bounded by a multiple of $(s_\star \vee \sqrt{\log c_n})\sqrt{c_n(s_\star \log p)^3/n}$ by Remark 6. This term tends to zero as $s_\star$ can be replaced by $s_0$.
- *Verification of* (C11): Note that $d_{B,n}(\Sigma_1, \Sigma_2) \le \|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}$ for every $\Sigma_1, \Sigma_2$ by (70), and hence it suffices to show that $\mathcal{H}$ is a separable metric space with the Frobenius norm. Since the support of the prior for $\Sigma$ is Euclidean, separability with the Frobenius norm is trivial.

Hence, under (C7*), Theorem 5 can be applied to obtain the distributional approximation in (15) with the zero matrix $H$. Under (C7*) and (C12), Theorem 6 implies the no-superset result in (16). If the beta-min condition (C13) is also met, the strong results in Corollary 2 and Corollary 3 hold. These establish (c)–(e).

### *B.2. Proof of Theorem 8*

We first verify the conditions for Theorem 3 for (a) and (b).

- *Verification of* (C1): Since $\Delta_{\eta,i}$ is the same for every $i \le n$ and the true parameters belong to the support of the prior, we see that $a_n = 1$ and $e_n = 0$ satisfy (C1).
- *Verification of* (C2): Observe that for every $\eta_1, \eta_2 \in \mathcal{H}$,

$$\begin{aligned}
\|\xi_{\eta_1} - \xi_{\eta_2}\|_2^2 &= |(\alpha_1 - \alpha_2) + (\mu_1^T\beta_1 - \mu_2^T\beta_2)|^2 + \|\mu_1 - \mu_2\|_2^2 \\
&\lesssim |\alpha_1 - \alpha_2|^2 + \|\mu_1\|_2^2\|\beta_1 - \beta_2\|_2^2 + (\|\beta_2\|_2^2 + 1)\|\mu_1 - \mu_2\|_2^2, \\
\|\Delta_{\eta_1} - \Delta_{\eta_2}\|_{\mathrm{F}}^2 &= |(\beta_1^T\Sigma_1\beta_1 - \beta_2^T\Sigma_2\beta_2) + (\sigma_1^2 - \sigma_2^2)|^2 \\
&\quad + 2\|\Sigma_1\beta_1 - \Sigma_2\beta_2\|_2^2 + \|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}^2 \\
&\lesssim (\|\beta_1\|_2^2 + 1)^2\|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}^2 + |\sigma_1^2 - \sigma_2^2|^2 \\
&\quad + (\|\beta_1\|_2^2 + \|\beta_2\|_2^2 + 1)\|\Sigma_2\|_{\mathrm{F}}^2\|\beta_1 - \beta_2\|_2^2.
\end{aligned} \tag{72}$$

Since $\|\beta_0\|_2$, $|\sigma_0^2|$, and $\|\Sigma_0\|_{\mathrm{F}}$ are bounded, it follows from the last display that there exists a constant $C_1$ such that $|\alpha-\alpha_0|+\|\beta-\beta_0\|_2+\|\mu-\mu_0\|_2+|\sigma^2-\sigma_0^2|+\|\Sigma-\Sigma_0\|_{\mathrm{F}}\le C_1\bar\epsilon_n$ implies $d_n(\eta,\eta_0)\le\bar\epsilon_n$ for any small $\bar\epsilon_n$. This shows that (C2) is satisfied as long as we choose $\bar\epsilon_n=\sqrt{\log n/n}$, as we have $|\alpha_0|\vee\|\beta_0\|_\infty\vee\|\mu_0\|_\infty\lesssim 1$, $\sigma_0^2\asymp 1$, and $1\lesssim\rho_{\min}(\Sigma_0)\le\rho_{\max}(\Sigma_0)\lesssim 1$.

- *Verification of* (C3): The assumption $\|\theta_0\|_\infty\lesssim\lambda^{-1}\log p$ given in the theorem directly satisfies (C3).
- *Verification of* (C4): Since $\Delta_\eta$ can be written as the sum of two positive definite matrices as

$$\Delta_\eta=\begin{pmatrix}\beta^T\Sigma\beta & \beta^T\Sigma\\ \Sigma\beta & \Sigma\end{pmatrix}+\begin{pmatrix}\sigma^2 & 0\\ 0 & \Psi\end{pmatrix},$$

condition (C4) is satisfied as we obtain $\sigma_0^2\wedge\rho_{\min}(\Psi)\le\rho_{\min}(\Delta_{\eta_0})\le\rho_{\max}(\Delta_{\eta_0})\le\|\Delta_{\eta_0}\|_{\mathrm{F}}$ by Weyl's inequality.

- *Verification of* (C5*): For a sufficiently large $M$ and $s_\star=s_0\vee(\log n/\log p)$, choose a sieve as

$$\begin{aligned}\mathcal{H}_n=\{(\alpha,\beta,\mu):|\alpha|^2+\|\beta\|_2^2+\|\mu\|_2^2\le n^{2M}\}\times\{\sigma:n^{-M}\le\sigma^2\le e^{Ms_\star\log p}\}\\ \times\{\Sigma:n^{-M}\le\rho_{\min}(\Sigma)\le\rho_{\max}(\Sigma)\le e^{Ms_\star\log p}\}.\end{aligned}$$

Then we have $\rho_{\min}(\Delta_\eta)\ge\sigma^2\wedge\rho_{\min}(\Psi)\ge n^{-M}$ for large $n$, and hence the minimum eigenvalue condition (6) is directly met with $\log\gamma_n\asymp\log n$ by the definition of the sieve. To see the entropy condition, observe from (72) that for every $\eta_1,\eta_2\in\mathcal{H}_n$,

$$\begin{aligned}d_n^2(\eta_1,\eta_2)\lesssim n^{4M}e^{2Ms_\star\log p}\big(|\alpha-\alpha_0|^2+\|\beta_1-\beta_2\|_2^2+\|\mu_1-\mu_2\|_2^2\\ +\|\Sigma_1-\Sigma_2\|_{\mathrm{F}}^2+|\sigma_1^2-\sigma_2^2|^2\big).\end{aligned}$$

Therefore, for $\delta_n=1/(6\overline{m}n^{3M+3/2}e^{Ms_\star\log p})$, the entropy relative to $d_n$ is bounded above by

$$\begin{aligned}&\log N\left(\delta_n,\{(\alpha,\beta,\mu):|\alpha|^2+\|\beta\|_2^2+\|\mu\|_2^2\le n^{2M}\},\|\cdot\|_2\right)\\ &+\log N\left(\delta_n,\{\sigma:\sigma^2\le e^{Ms_\star\log p}\},\|\cdot\|_2\right)\\ &+\log N\left(\delta_n,\{\Sigma:\|\Sigma\|_{\mathrm{F}}\le\sqrt{q}e^{Ms_\star\log p}\},\|\cdot\|_{\mathrm{F}}\right),\end{aligned}$$

each summand of which is bounded by a multiple of $\log n+s_\star\log p$. This shows that the choice $\epsilon_n=\sqrt{(s_\star\log p)/n}$ satisfies the entropy condition in (7). Further, it is easy to see that condition (8) holds using the tail bounds for normal and inverse Wishart distributions as in (71).

- *Verification of* (C6): Note that the mean of $Y$ is expressed as $X\theta+Z\xi_\eta$ for $Z=1_n\otimes I_{q+1}$. Since the condition $\varsigma_{\min}([X_S^*,1_n])\gtrsim 1$ implies $\varsigma_{\min}([X_S,Z])\gtrsim 1$, condition (C6) is satisfied by Remark 3.

Therefore we obtain the contraction properties of the posterior distribution as in (9) with $s_\star$ replaced by $s_0$ as $s_0>0$ and $\log n\lesssim\log p$. The rates for $\eta$ with respect to more concrete metrics than $d_n$ can now be obtained. Note that for small $\delta > 0$, $d_n(\eta, \eta_0) \le \delta$ directly implies $\|\mu - \mu_0\|_2 \le \delta$ and $\|\Sigma - \Sigma_0\|_{\mathrm{F}} \le \delta$ by the definition of $d_n$. For $\beta$, observe that

$$\begin{aligned}\|\beta - \beta_0\|_2 &\le \|\Sigma^{-1}\|_{\mathrm{sp}} \|\Sigma(\beta - \beta_0)\|_2 \\ &\le \|\Sigma^{-1}\|_{\mathrm{sp}} (\|\Sigma\beta - \Sigma_0\beta_0\|_2 + \|\Sigma - \Sigma_0\|_{\mathrm{F}} \|\beta_0\|_2) \\ &\lesssim \|\Sigma^{-1}\|_{\mathrm{sp}} \delta.\end{aligned}$$

Since $\|\Sigma^{-1}\|_{\mathrm{sp}}$ is bounded as $\|\Sigma - \Sigma_0\|_{\mathrm{F}} \le \delta$, the preceding display implies $\|\beta - \beta_0\|_2 \lesssim \delta$. Moreover, we have

$$\begin{aligned}|\alpha - \alpha_0| &\le |\mu^T\beta - \mu_0^T\beta_0| + \delta \\ &\lesssim \|\mu\|_2 \|\beta - \beta_0\|_2 + \|\beta_0\|_2 \|\mu - \mu_0\|_2 + \delta \\ &\lesssim (\|\mu\|_2 + 1)\delta,\end{aligned}$$

and

$$\begin{aligned}|\sigma^2 - \sigma_0^2| &\le |\beta^T\Sigma\beta - \beta_0^T\Sigma_0\beta_0| + |(\beta^T\Sigma\beta + \sigma^2) - (\beta_0^T\Sigma_0\beta_0 + \sigma_0^2)| \\ &\le \|\beta\|_2 \|\Sigma\beta - \Sigma_0\beta_0\|_2 + \|\beta_0\|_2 \|\Sigma_0\|_{\mathrm{sp}} \|\beta - \beta_0\|_2 + \delta \\ &\lesssim (\|\beta\|_2 + 1)\delta.\end{aligned}$$

These show that $|\alpha - \alpha_0| + |\sigma^2 - \sigma_0^2| \lesssim \delta$ as $\|\mu\|_2$ and $\|\beta\|_2$ are bounded. We finally conclude that $|\alpha - \alpha_0| + \|\beta - \beta_0\|_2 + \|\mu - \mu_0\|_2 + |\sigma^2 - \sigma_0^2| + \|\Sigma - \Sigma_0\|_{\mathrm{F}}$ contracts at the same rate of $d_n$. The optimal posterior contraction is directly obtained by Corollary 1. Thus assertions (a) and (b) hold.

Next, we verify conditions (C8*)–(C10*) and (C11) to apply Theorems 5–6 and Corollaries 2–3. The orthogonal projection defined by $H = \tilde{Z}(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T$ with $\tilde{Z} = 1_n \otimes \Delta_{\eta_0}^{-1/2}$ is used to check the conditions.

- *Verification of* (C8*): For $H$ defined above, it is easy to see that the first condition of (C8*) is satisfied. The second condition is directly satisfied by Remark 5.
- *Verification of* (C9*): Choose a map $(\alpha, \beta, \mu, \sigma^2, \Sigma) \mapsto (\alpha + n^{-1}1_n^T X^*(\theta - \theta_0), \beta, \mu, \sigma^2, \Sigma)$ for $\eta \mapsto \tilde{\eta}_n(\theta, \eta)$. To check (C9*), we shall verify that this map induces $\Phi(\tilde{\eta}_n(\theta, \eta)) = (\tilde{\xi}_\eta + H\tilde{X}(\theta - \theta_0), \tilde{\Delta}_\eta)$ as follows. Note that for matrices $R_k$, $k = 1, \dots, 6$, we have the properties of the Kronecker product that $(R_1 \otimes R_2)(R_3 \otimes R_4) = (R_1R_2 \otimes R_3R_4)$ and $(R_5 \otimes R_6)^{-1} = R_5^{-1} \otimes R_6^{-1}$ if the matrices allow such operations. Using these properties, we see that $H$ satisfies

$$\begin{aligned}H &= (1_n \otimes \Delta_{\eta_0}^{-1/2})(1_n^T 1_n \otimes \Delta_{\eta_0}^{-1})^{-1}(1_n \otimes \Delta_{\eta_0}^{-1/2})^T \\ &= \frac{1}{n}(1_n \otimes \Delta_{\eta_0}^{-1/2})\Delta_{\eta_0}(1_n \otimes \Delta_{\eta_0}^{-1/2})^T \\ &= \frac{1}{n}(1_n \otimes I_{q+1})(1_n^T \otimes I_{q+1}) \\ &= \frac{1}{n}(1_n 1_n^T \otimes I_{q+1}).\end{aligned}$$

Hence,

$$\begin{aligned} Z(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{X}(\theta-\theta_0) &= (I_n\otimes\Delta_{\eta_0}^{1/2})H(I_n\otimes\Delta_{\eta_0}^{-1/2})X(\theta-\theta_0) \\ &= HX(\theta-\theta_0) = 1_n\otimes\begin{pmatrix} n^{-1}1_n^TX^*(\theta-\theta_0) \\ 0_{q\times 1}\end{pmatrix}, \end{aligned}$$

which implies that the shift only for $\alpha$ as in the given map provides $\Phi(\tilde{\eta}_n(\theta,\eta)) = (\tilde{\xi}_\eta + H\tilde{X}(\theta-\theta_0), \tilde{\Delta}_\eta)$. Without loss of generality, we assume that the standard normal prior is used for $\alpha$. Now, observe that

$$\begin{aligned} \left|\log\frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}}(\eta)\right| &\lesssim \left|\alpha^2 - (\alpha + n^{-1}1_n^TX^*(\theta-\theta_0))^2\right| \\ &\le 2|\alpha||n^{-1}1_n^TX^*(\theta-\theta_0)| + (n^{-1}1_n^TX^*(\theta-\theta_0))^2, \end{aligned}$$

since the priors for the other parameters cancel out due to invariance. One can note that

$$\sup_{\eta\in\widehat{\mathcal{H}}_n}|\alpha| \lesssim s_\star\sqrt{(\log p)/n} + |\alpha_0| \lesssim 1,$$

and

$$\frac{1}{\sqrt{n}}\sup_{\theta\in\widehat{\Theta}_n}\|X(\theta-\theta_0)\|_2 \lesssim s_\star\sqrt{(\log p)/n}.$$

Thus, condition (C9*) is satisfied.

- *Verification of* (C10*): Note again that $d_{B,n}(\eta,\eta_0) \lesssim \|\Sigma-\Sigma_0\|_{\mathrm{F}} + |\sigma^2-\sigma_0^2| + \|\beta-\beta_0\|_2$ for every $\eta\in\widehat{\mathcal{H}}_n$. The inequality also holds for the other direction for every $\eta\in\widehat{\mathcal{H}}_n$, by the same argument used for the recovery in the proof of Theorem 8, (a)–(b). Hence, for some constants $C_1, C_2 > 0$, the entropy in (C10*) is bounded above by

$$\begin{aligned} &\log N\left(C_1\delta, \left\{\beta : \|\beta-\beta_0\|_2 \le C_2\hat{M}_2\epsilon_n\right\}, \|\cdot\|_2\right) \\ &+ \log N\left(C_1\delta, \left\{\sigma^2 : |\sigma^2-\sigma_0^2| \le C_2\hat{M}_2\epsilon_n\right\}, |\cdot|\right) \\ &+ \log N\left(C_1\delta, \left\{\Sigma : \|\Sigma-\Sigma_0\|_{\mathrm{F}} \le C_2\hat{M}_2\epsilon_n\right\}, \|\cdot\|_{\mathrm{F}}\right). \end{aligned}$$

  Since all nuisance parameters are of fixed dimensions, the last display is bounded by a multiple of $0\vee\log(3C_2\hat{M}_2\epsilon_n/C_1\delta)$ for every $\delta > 0$, so that (C10*) is bounded by $(s_\star^5\log^3 p/n)^{1/2}$ by Remark 6. Since $s_\star \lesssim s_0$ in this case, the condition is verified.
- *Verification of* (C11): Note that by (72), $d_{B,n}(\eta_1,\eta_2) \lesssim \|\Sigma_1-\Sigma_2\|_{\mathrm{F}} + |\sigma_1^2 - \sigma_2^2| + \|\beta_1-\beta_2\|_2$ for every $\eta_1,\eta_2\in\widehat{\mathcal{H}}_n$. Since each of the parameter spaces of $\Sigma$, $\sigma^2$, and $\beta$ is a separable metric space with each of these norms, (C11) is satisfied.

Therefore, under (C7*), Theorem 5 implies that the distributional approximation in (15) holds. Under (C7*) and (C12), we obtain the no-superset result in (16). The remaining assertions in the theorem are direct consequences of Corollary 2 and Corollary 3 if the beta-min condition (C13) is also satisfied. These prove (c)–(e).

We complete the proof by showing that the covariance matrix of the nonzero part can be written as in the theorem. For given $S$, we obtain

$$\begin{aligned}&\tilde X_S^T(I_{n(q+1)}-H)\tilde X_S\\&\quad=X_S^{*T}\left(I_n\otimes\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}^T\right)(I_n\otimes I_{q+1}-H)\left(I_n\otimes\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}\right)X_S^*\\&\quad=\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}^T\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}X_S^{*T}H^*X_S^*,\end{aligned}$$

where $\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}$ is the first column of $\Delta_{\eta_0}^{-1/2}$. Note that $\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}^T\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}=\{\Delta_{\eta_0}^{-1}\}_{1,1}$, where $\{\Delta_{\eta_0}^{-1}\}_{1,1}$ is the top-left element of $\Delta_{\eta_0}^{-1}$, which is equal to $(\beta_0^T\Sigma_0\beta_0+\sigma_0^2-\beta_0^T\Sigma_0(\Sigma_0+\Psi)^{-1}\Sigma_0\beta_0)^{-1}=(\sigma_0^2+\beta_0^T\Sigma_0(\Sigma_0+\Psi)^{-1}\Psi\beta_0)^{-1}$ by direct calculations. For the mean $\hat\theta_S$, observe that

$$\begin{aligned}&\tilde X_S^T(I_{n(q+1)}-H)(U+\tilde X\theta_0)\\&\quad=X_S^{*T}\left(I_n\otimes\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}^T\right)\left(I_n\otimes I_{q+1}-\frac{1}{n}1_n1_n^T\otimes I_{q+1}\right)\\&\qquad\times\Big\{\left(I_n\otimes\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}\right)\left(Y^*-(\alpha_0+\mu_0^T\beta_0)1_n\right)\\&\qquad\quad+\left(I_n\otimes\{\Delta_{\eta_0}^{-1/2}\}_{\cdot(-1)}\right)(W-1_n\otimes\mu_0)\Big\},\end{aligned}$$

where $\{\Delta_{\eta_0}^{-1/2}\}_{\cdot(-1)}$ is the submatrix of $\Delta_{\eta_0}^{-1/2}$ consisting of columns except for $\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}$ the first column. Since $\{\Delta_{\eta_0}^{-1/2}\}_{\cdot 1}^T\{\Delta_{\eta_0}^{-1/2}\}_{\cdot(-1)}=\{\Delta_{\eta_0}^{-1}\}_{1,(-1)}$, where $\{\Delta_{\eta_0}^{-1}\}_{1,(-1)}$ is the first row of $\Delta_{\eta_0}^{-1}$ with the top-left element excluded, the last display is equal to

$$\begin{aligned}X_S^{*T}\Big\{H^*\Big[&\{\Delta_{\eta_0}^{-1}\}_{1,1}\left(Y^*-(\alpha_0+\mu_0^T\beta_0)1_n\right)\\&+\left(I_n\otimes\{\Delta_{\eta_0}^{-1}\}_{1,(-1)}\right)(W-1_n\otimes\mu_0)\Big]\Big\}.\end{aligned}$$

As we have $\{\Delta_{\eta_0}^{-1}\}_{1,(-1)}=-\{\Delta_{\eta_0}^{-1}\}_{1,1}^{-1}\beta_0^T\Sigma_0(\Sigma_0+\Psi)^{-1}$ by direct calculations, it follows that

$$\begin{aligned}\hat\theta_S=\left(X_S^{*T}H^*X_S^*\right)^{-1}X_S^{*T}\Big\{H^*\Big[&\left(Y^*-(\alpha_0+\mu_0^T\beta_0)1_n\right)\\&-\left(I_n\otimes(\beta_0^T\Sigma_0(\Sigma_0+\Psi)^{-1})\right)(W-1_n\otimes\mu_0)\Big]\Big\}.\end{aligned}$$

This completes the proof.

### B.3. *Proof of Theorem 9*

We shall verify the conditions for the posterior contraction in Theorem 3 to prove (a)–(b). First we give the bounds for the eigenvalues of each correlation matrix. It can be shown that

$$1-\alpha = \rho_{\min}\left(G_i^{\mathrm{CS}}(\alpha)\right) \le \rho_{\max}\left(G_i^{\mathrm{CS}}(\alpha)\right) = 1+(m_i-1)\alpha, \tag{73}$$

$$\frac{1-\alpha^2}{(1+|\alpha|)^2} \le \rho_{\min}\left(G_i^{\mathrm{AR}}(\alpha)\right) \le \rho_{\max}\left(G_i^{\mathrm{AR}}(\alpha)\right) \le \frac{1-\alpha^2}{(1-|\alpha|)^2}, \tag{74}$$

$$1-2|\alpha| \le \rho_{\min}\left(G_i^{\mathrm{MA}}(\alpha)\right) \le \rho_{\max}\left(G_i^{\mathrm{MA}}(\alpha)\right) \le 1+2|\alpha|. \tag{75}$$

The first assertion in (73) follows directly from the identity $\rho_k(G_i^{\mathrm{CS}}(\alpha)) = \rho_k(\alpha 1_{m_i} 1_{m_i}^T) + 1 - \alpha$ for every $k \le m_i$. For (74), see Theorem 2.1 and Theorem 3.5 of Fikioris [12]. The assertion in (75) is due to Theorem 2.2 of Kulkarni et al. [21].

- *Verification of* (C1): For the autoregressive correlation matrix, note that

$$\begin{aligned}&\max_{1\le i\le n}\left\|\sigma^2 G_i^{\mathrm{AR}}(\alpha) - \sigma_0^2 G_i^{\mathrm{AR}}(\alpha_0)\right\|_{\mathrm{F}}^2 \\ &\quad = \overline{m}(\sigma^2-\sigma_0^2)^2 + 2\sum_{k=1}^{\overline{m}-1}(\overline{m}-k)(\sigma^2\alpha^k - \sigma_0^2\alpha_0^k)^2.\end{aligned}$$

  Using $\overline{m}n \asymp n_*$, we have that

$$\begin{aligned}\sum_{k=1}^{\overline{m}-1}(\overline{m}-k)(\sigma^2\alpha^k-\sigma_0^2\alpha_0^k)^2 &\lesssim \frac{1}{n}\sum_{k=1}^{\overline{m}-1}(\sigma^2\alpha^k-\sigma_0^2\alpha_0^k)^2\sum_{i=1}^{n}\{(m_i-k)\vee 0\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m_i-1}(m_i-k)(\sigma^2\alpha^k-\sigma_0^2\alpha_0^k)^2,\end{aligned}$$

  and hence

$$\max_{1\le i\le n}\|\sigma^2 G_i^{\mathrm{AR}}(\alpha)-\sigma_0^2 G_i^{\mathrm{AR}}(\alpha_0)\|_{\mathrm{F}}^2 \lesssim \frac{1}{n}\sum_{i=1}^{n}\|\sigma^2 G_i^{\mathrm{AR}}(\alpha)-\sigma_0^2 G_i^{\mathrm{AR}}(\alpha_0)\|_{\mathrm{F}}^2.$$

  This gives us $a_n \asymp 1$ for the autoregressive matrices. Similarly, we can also show that $a_n \asymp 1$ satisfies (C1) for the compound-symmetric and the moving average correlation matrices. Also, we have $e_n = 0$ for (C1) as the true parameter values $\alpha_0$ and $\sigma_0^2$ are in the support of the prior.
- *Verification of* (C2): Since the nuisance parameters are of fixed dimensions, condition (C2) is satisfied with $\bar{\epsilon}_n = \sqrt{(\log n)/n}$ due to the restricted range of the true parameters, $\sigma_0^2 \asymp 1$ and $\alpha_0 \in [b_1+\epsilon, b_2-\epsilon]$ for some fixed $\epsilon > 0$.
- *Verification of* (C3): The assumption $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$ given in the theorem directly satisfies (C3).
- *Verification of* (C4): Using (73)–(75), we see that for the compound-symmetric correlation matrix, condition (C4) is satisfied with the bounded range of the true parameters provided that $\overline{m}$ is bounded. For the other correlation matrices, condition (C4) is satisfied even with increasing $\overline{m}$.

- *Verification of* (C5*): For a sufficiently large $M > 0$ and $s_\star = s_0 \vee (\log n / \log p)$, choose a sieve $\mathcal{H}_n = \{\sigma^2 : n^{-M} \le \sigma^2 \le e^{M s_\star \log p}\} \times \{\alpha : b_1 + n^{-M} \le \alpha \le b_2 - n^{-M}\}$. Then using (73)–(75), it is easy to see that the minimum eigenvalue of each correlation matrix is bounded below by a polynomial in $n$, which implies that condition (6) is satisfied with $\log \gamma_n \asymp \log n$. For the entropy calculation, note that for every type of correlation matrix,

$$\begin{aligned} d_n^2(\eta_1, \eta_2) &= \frac{1}{n} \sum_{i=1}^n \|\sigma_1^2 G_i(\alpha_1) - \sigma_2^2 G_i(\alpha_2)\|_{\mathrm{F}}^2 \\ &\le \frac{1}{n} \sum_{i=1}^n \left\{ (\sigma_1^2 - \sigma_2^2)^2 \|G_i(\alpha_1)\|_{\mathrm{F}}^2 + \sigma_2^4 \|G_i(\alpha_1) - G_i(\alpha_2)\|_{\mathrm{F}}^2 \right\}. \end{aligned} \tag{76}$$

  From the identity $\alpha_1^k - \alpha_2^k = (\alpha_1 - \alpha_2) \sum_{j=0}^{k-1} \alpha_1^j \alpha_2^{k-1-j}$ for every integer $k \ge 1$, we have that $|\alpha_1^k - \alpha_2^k| \lesssim k|\alpha_1 - \alpha_2|$ for every $\alpha_1, \alpha_2 \in (b_1, b_2)$. By this inequality we obtain $\|G_i(\alpha_1) - G_i(\alpha_2)\|_{\mathrm{F}}^2 \lesssim \overline{m}^4 |\alpha_1 - \alpha_2|^2$ for every correlation matrix. Then, the last display is bounded by a multiple of $\overline{m}^2(\sigma_1^2 - \sigma_2^2)^2 + e^{2Ms_\star \log p} \overline{m}^4 (\alpha_1 - \alpha_2)^2$ for every $\eta_1, \eta_2 \in \mathcal{H}_n$. The entropy in (7) is thus bounded by

$$\log N\big(\delta_n, \{\sigma^2 : 0 < \sigma^2 \le e^{M s_\star \log p}\}, |\cdot|\big) + \log N\big(\delta_n, \{\alpha : 0 < \alpha < 1\}, |\cdot|\big),$$

  for $\delta_n = (6\overline{m}^3 n^{3/2 + C_1} e^{M s_\star \log p})^{-1}$ with some constant $C_1 > 0$. It can be easily checked that each term in the last display is bounded by a multiple of $s_\star \log p$, by which the entropy condition in (7) is satisfied with $\epsilon_n = \sqrt{(s_\star \log p)/n}$. Using the tail bounds of inverse gamma distributions and properties of the density $\Pi(d\alpha)$ near the boundaries, condition (8) is satisfied as long as $M$ is chosen sufficiently large.
- *Verification of* (C6): The separation condition is trivially satisfied as there is no nuisance mean part.

Therefore, we obtain the posterior contraction properties of $\theta$ with $s_\star = s_0 \vee (\log n / \log p)$ by Theorem 3. The term $s_\star$ can be replaced by $s_0$ since $s_0 > 0$ and $\log n \lesssim \log p$. Since we have $m_i(\sigma^2 - \sigma_0^2)^2 \le \|\sigma^2 G_i(\alpha) - \sigma_0^2 G_i(\alpha_0)\|_{\mathrm{F}}^2$ by the diagonal entries of each matrix, the contraction rate $\sqrt{(s_0 \log p)/(\overline{m} n)}$ is obtained for $\sigma^2$ with respect to the $\ell_2$-norm, for every correlation matrix, as $\overline{m} n \asymp n_*$. In particular, for the compound-symmetric correlation matrix, this rate is reduced to $\sqrt{(s_0 \log p)/n}$ since $\overline{m}$ is bounded in that case. We also have $m_i(\sigma^2\alpha - \sigma_0^2\alpha_0)^2 \le \|\sigma^2 G_i(\alpha) - \sigma_0^2 G_i(\alpha_0)\|_{\mathrm{F}}^2$ for every correlation matrix, as there are more than $m_i$ entries that is equal to $\sigma^2\alpha - \sigma_0^2\alpha_0$. Hence, by the relation $|\alpha - \alpha_0| \lesssim |\sigma^2\alpha - \sigma_0^2\alpha_0| + |\alpha||\sigma^2 - \sigma_0^2|$, the same rate is also obtained for $\alpha$ relative to the $\ell_2$-norm. The optimal posterior contraction directly follows from Corollary 1. Thus assertions (a)–(b) hold.

Next, we verify conditions (C8*)–(C10*) and (C11) to apply Theorems 5–6 and Corollaries 2–3.

- *Verification of* (C8*)–(C9*): These conditions are trivially satisfied with the zero matrix $H$ since there is no nuisance mean part.

- *Verification of* (C10*): Using the results of contraction rates of $\sigma^2$ and $\alpha$, note that there exists a constant $C_2 > 0$ such that $\{\eta \in \mathcal{H} : d_{B,n}(\eta, \eta_0) \leq \hat{M}_2\epsilon_n\} \subset \{\sigma^2 : |\sigma^2 - \sigma_0^2| \leq C_2\epsilon_n/\sqrt{\overline{m}}\} \times \{\alpha : |\alpha - \alpha_0| \leq C_2\epsilon_n/\sqrt{\overline{m}}\}$. Thus the entropy in (C10*) is bounded by $0 \vee 2\log(3C_2\epsilon_n/\sqrt{\overline{m}}\delta)$. By Remark 6, (C10*) is bounded by a multiple of $\{(s_\star^5 \log^3 p)/n\}^{1/2}$, which goes to zero by the assumption since $s_\star \lesssim s_0$.
- *Verification of* (C11): Using (76), we have $d_{B,n}(\eta_1, \eta_2) \lesssim \overline{m}|\sigma_1^2 - \sigma_2^2| + \overline{m}^2|\alpha_1 - \alpha_2|$ for every $\eta_1, \eta_2 \in \widehat{\mathcal{H}}_n$. Since the parameter spaces of $\alpha$ and $\sigma^2$ are Euclidean and hence separable under the $\ell_2$-metric, condition (C11) is satisfied.

Therefore, under (C7*), the distributional approximation in (15) holds with the zero matrix $H$ by Theorem 5. Under (C7*) and (C12), Theorem 6 implies that the no-superset result in (16) holds. The strong results in Corollary 2 and Corollary 3 follow explicitly from the beta-min condition (C13). These prove (c)–(e).

### *B.4. Proof of Theorem 10*

We verify the conditions for the posterior contraction in Theorem 3 to show (a)–(b).

- *Verification of* (C1): Using the assumption $\max_i \|Z_i\|_{\text{sp}} \lesssim 1$, note that

$$\begin{aligned}
&\max_{1\leq i\leq n} \|Z_i(\Psi - \Psi_0)Z_i^T\|_{\text{F}}^2 \\
&\quad\leq \|\Psi - \Psi_0\|_{\text{F}}^2 \max_{1\leq i\leq n} \|Z_i\|_{\text{sp}}^4 \\
&\quad\lesssim \frac{1}{\sum_{i=1}^n \mathbb{1}(m_i \geq q)} \sum_{i:m_i\geq q} \|Z_i(\Psi - \Psi_0)Z_i^T\|_{\text{F}}^2 \|(Z_i^T Z_i)^{-1} Z_i^T\|_{\text{sp}}^4 \qquad (77)\\
&\quad\lesssim \frac{1}{n}\sum_{i=1}^n \|Z_i(\Psi - \Psi_0)Z_i^T\|_{\text{F}}^2,
\end{aligned}$$

  where the last inequality holds since $\min_i\{\varsigma_{\min}(Z_i) : m_i \geq q\} \gtrsim 1$ and $\sum_{i=1}^n \mathbb{1}(m_i \geq q) \asymp n$. Thus we have $a_n \asymp 1$ and $e_n = 0$.
- *Verification of* (C2): The condition is satisfied with $\bar{\epsilon}_n = \sqrt{(\log n)/n}$ as $\Psi$ is fixed dimensional and we have $1 \lesssim \rho_{\min}(\Psi_0) \leq \rho_{\max}(\Psi_0) \lesssim 1$.
- *Verification of* (C3): The assumption $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$ given in the theorem directly satisfies (C3).
- *Verification of* (C4): By Weyl's inequality, we obtain that

$$\min_{1\leq i\leq n} \rho_{\min}(\sigma^2 I_{m_i} + Z_i\Psi_0 Z_i^T) \geq \sigma^2 + \min_{1\leq i\leq n} \rho_{\min}(Z_i\Psi_0 Z_i^T), \qquad (78)$$

$$\max_{1\leq i\leq n} \rho_{\max}(\sigma^2 I_{m_i} + Z_i\Psi_0 Z_i^T) \leq \sigma^2 + \rho_{\max}(\Psi_0) \max_{1\leq i\leq n} \|Z_i\|_{\text{sp}}^2. \qquad (79)$$

  Since $Z_i\Psi_0 Z_i^T$ is nonnegative definite, the right hand side of (78) is further bounded below by $\sigma^2$, while the right hand side of (79) is bounded. The condition (C4) is thus satisfied.

- *Verification of* (C5*): For a sufficiently large $M$ and $s_\star = s_0 \vee (\log n/\log p)$, define a sieve as $\mathcal{H}_n = \{\Psi : n^{-M} \le \rho_{\min}(\Sigma) \le \rho_{\max}(\Sigma) \le e^{Ms_\star \log p}\}$, so that the minimum eigenvalue condition (6) can be satisfied with $\log \gamma_n \asymp \log n$. Similar to the proof of Theorem 7, it can be easily shown that conditions (7) and (8) are satisfied with $\epsilon_n = \sqrt{(s_\star \log p)/n}$.
- *Verification of* (C6): The separation condition is trivially satisfied as there is no nuisance mean part.

Therefore, the posterior contraction rates for $\theta$ are given by Theorem 3 with $s_\star$ replaced by $s_0$ since $s_0 > 0$ and $\log n \lesssim \log p$. The contraction rate for $\Sigma$ relative to the Frobenius norm is a direct consequence of (77). The optimal posterior contraction easily follows from Corollary 1. Thus assertions (a)–(b) hold.

Now, we verify conditions (C8*)–(C10*) and (C11) to apply Theorems 5–6 and Corollaries 2–3.

- *Verification of* (C8*)–(C9*): These conditions are trivially satisfied with the zero matrix $H$ since there is no nuisance mean part.
- *Verification of* (C10*): For some $C_1 > 0$, the entropy in (C10*) is bounded above by a multiple of $\log N(\delta, \{\Sigma : \|\Sigma - \Sigma_0\|_{\mathrm{F}} \le \hat{M}_2 C_1 \epsilon_n\}, \|\cdot\|_{\mathrm{F}}) \lesssim 0 \vee \log(3\hat{M}_2 C_1 \epsilon_n/\delta)$ by (77). The expression in (C10*) is thus bounded by a constant multiple of $s_\star^5 \log^3 p$ by Remark 6. This tends to zero since $s_\star \lesssim s_0$.
- *Verification of* (C11): It is easy to see that $d_{B,n}(\eta, \eta_0) \lesssim \|\Psi - \Psi_0\|_{\mathrm{F}}$ since $\max_i \|Z_i\|_{\mathrm{sp}} \lesssim 1$. The separability of the space is thus trivial.

Hence, under (C7*), Theorem 5 can be applied to obtain the distributional approximation in (15) with the zero matrix $H$. Under (C7*) and (C12), we obtain the no-superset result in (16) by Theorem 6. The strong results in Corollary 2 and Corollary 3 follow explicitly from the beta-min condition (C13). These establish (c)–(e).

### *B.5. Proof of Theorem 11*

We verify the conditions for the posterior contraction in Theorem 3.

- *Verification of* (C1): Since $\Delta_{\eta,i} = \Omega^{-1}$ for every $i \le n$ and $\Omega_0 \in \mathcal{M}_0^+(cL)$ for some $0 < c < 1$, $a_n = 1$ and $e_n = 0$ satisfy (C1).
- *Verification of* (C2): Using (i) of Lemma 10 and the relation $1 - x \asymp 1 - x^{-1}$ as $x \to 1$, observe that $\|\Omega^{-1} - \Omega_0^{-1}\|_{\mathrm{F}} \lesssim \|\Omega - \Omega_0\|_{\mathrm{F}} \lesssim \bar{\epsilon}_n$ if the right hand side is small enough. Thus, there exists a constant $C_1 > 0$ such that $\{\Omega : \|\Omega^{-1} - \Omega_0^{-1}\|_{\mathrm{F}} \le \bar{\epsilon}_n\} \supset \{\Omega : \|\Omega - \Omega_0\|_{\mathrm{F}} \le C_1 \bar{\epsilon}_n\}$. Furthermore, although the components of $\Omega$ are not a priori independent as the prior is truncated to $\mathcal{M}_0^+(L)$, the truncation can only increase prior concentration since $\Omega_0 \in \mathcal{M}_0^+(cL)$ for some $0 < c < 1$. Hence, for some $C_2 > 0$,

$$\Pi\left(\|\Omega^{-1} - \Omega_0^{-1}\|_{\mathrm{F}} \le \bar{\epsilon}_n\right) \ge \Pi\left(\|\Omega - \Omega_0\|_\infty \le C_2 \bar{\epsilon}_n/\overline{m}\right) \gtrsim \left(\frac{C_2 \bar{\epsilon}_n}{\overline{m}}\right)^{\overline{m}+d},$$

  which justifies the choice $\bar{\epsilon}_n \asymp \sqrt{(\overline{m} + d)(\log n)/n}$ for (C2).

- *Verification of* (C3): The assumption $\|\theta_0\|_\infty \lesssim \lambda^{-1}\log p$ given in the theorem directly satisfies (C3).
- *Verification of* (C4): This is trivially met as $\Omega_0 \in \mathcal{M}_0^+(cL)$ for some $0<c<1$.
- *Verification of* (C5*): Note that the minimum eigenvalue condition (6) is trivially satisfied with $\gamma_n = 1$ since the prior is put on $\mathcal{M}_0^+(L)$. Now, for $\bar{r}_n = Ms_\star \log p/\log n$ with $s_\star = s_0 \vee (n\bar{\epsilon}_n^2/\log p)$ and sufficiently large $M$, choose a sieve as $\mathcal{H}_n = \{\Omega \in \mathcal{M}_0^+(L) : \sum_{j,k}\mathbb{1}\{\omega_{jk}\neq 0\} \le \bar{r}_n\}$, that is, the maximum number of edges of $\Omega$ does not exceed $\bar{r}_n$. Then, for $\delta_n = 1/6\overline{m}n^{3/2}$, the entropy in (7) is bounded by

  $$\begin{aligned}\log N(\delta_n/\overline{m}, \mathcal{H}_n, \|\cdot\|_\infty) &\le \log\left\{\left(\frac{\overline{m}L}{\delta_n}\right)^{\overline{m}+\bar{r}_n}\binom{\binom{\overline{m}}{2}}{\bar{r}_n}\right\}\\ &\le (\overline{m}+\bar{r}_n)\log(\overline{m}L/\delta_n) + 2\bar{r}_n\log\overline{m},\end{aligned}$$

  where in the second term, the factor $(\overline{m}L/\delta_n)^{\overline{m}}$ comes from the diagonal elements of $\Omega$, while the rest is from the off-diagonal entries. It is easy to see that the last display is bounded by a multiple of $s_\star \log p$ with chosen $\bar{r}_n$, and hence the entropy condition in (7) is satisfied. Lastly, note that for some $C_3 > 0$,

  $$\log \Pi(\mathcal{H}\setminus\mathcal{H}_n) = \log\Pi(|\Upsilon| > \bar{r}_n) \lesssim -\bar{r}_n \log \bar{r}_n \le -C_3 M s_\star \log p.$$

  Therefore, condition (8) is satisfied with sufficiently large $M$.
- *Verification of* (C6): The separation condition is trivially met as there is no nuisance mean part.

Therefore, we obtain the posterior contraction properties for $\theta$ by Theorem 3. The theorem also implies that the posterior distribution of $\Omega^{-1}$ contracts to $\Omega_0^{-1}$ at the rate $\epsilon_n = \sqrt{(s_0\log p \vee (\overline{m}+d)\log n)/n}$ with respect to the Frobenius norm. This is also translated as convergence of $\Omega$ to $\Omega_0$ at the same rate, since we obtain

$$\|\Omega-\Omega_0\|_{\mathrm{F}}^2 \lesssim \|\Omega^{-1}-\Omega_0^{-1}\|_{\mathrm{F}}^2 \lesssim \epsilon_n^2, \tag{80}$$

by (i) of Lemma 10 and the inequality $1-x \asymp 1-x^{-1}$ as $x\to 1$. The assertion for the optimal posterior contraction is directly justified by Corollary 1. These prove (a)–(b).

Next, we verify conditions (C8)–(C11) to obtain the optimal posterior contraction by applying Theorem 4.

- *Verification of* (C8)–(C9): These conditions are trivially satisfied with the zero matrix $H$ since there is no nuisance mean part.
- *Verification of* (C10): Note that by (80), there exists a constant $C_4 > 0$ such that the entropy in (C10) is bounded by $\log N(\delta, \{\Omega : \|\Omega-\Omega_0\|_{\mathrm{F}} \le C_4\bar{\epsilon}_n\}, d_{B,n})$ for every $\delta > 0$. Using (81), the entropy is further bounded by $\log N(C_5\delta, \{\Omega : \|\Omega-\Omega_0\|_{\mathrm{F}} \le C_4\bar{\epsilon}_n\}, \|\cdot\|_{\mathrm{F}})$ for some $C_5 > 0$. This is clearly bounded by a multiple of $0 \vee \overline{m}^2\log(3C_4\bar{\epsilon}_n/C_5\delta)$, and hence using Remark 6 we bound (C10) by a multiple of $(\sqrt{\bar{s}_\star}\vee\overline{m})\sqrt{(\bar{s}_\star\log p)/n}$ which goes to zero by assumption.

- *Verification of* (C11): For every $\Omega_1, \Omega_2 \in \widehat{\mathcal{H}}_n$, note that
$$\|\Omega_1^{-1} - \Omega_2^{-1}\|_{\mathrm{F}} \lesssim \|\Omega_1 - \Omega_2\|_{\mathrm{F}} \lesssim \|\Omega_1^{-1} - \Omega_2^{-1}\|_{\mathrm{F}} \lesssim \epsilon_n, \tag{81}$$
using (i) of Lemma 10 and the inequality $1 - x \asymp 1 - x^{-1}$ as $x \to 1$ again. By the first inequality, it suffices to show that $\mathcal{H}$ is separable metric space with the Frobenius norm. This is trivial as the parameter space is Euclidean.

Hence, under condition (C7), Theorem 4 verifies (c).

Now, we verify conditions (C8*)–(C10*) to apply Theorems 5–6 and Corollaries 2–3.

- *Verification of* (C8*)–(C9*): These are trivially satisfied for the same reason as (C8)–(C9).
- *Verification of* (C10*): Similar to the verification of (C10), the entropy in (C10*) is bounded by a multiple of $0 \vee \overline{m}^2 \log(3C_6\epsilon_n/\delta)$ for some $C_6 > 0$. Hence using Remark 6 we bound (C10*) by a multiple of $(s_\star \vee \overline{m})\sqrt{(s_\star \log p)^3/n}$ which goes to zero by assumption.

Therefore, under (C7*), we obtain the distributional approximation in (15) with the zero matrix $H$ by Theorem 5. Under (C7*) and (C12), the no-superset result in (16) holds by Theorem 6. Lastly, we obtain the strong results in Corollary 2 and Corollary 3 if the beta-min condition (C13) is also met. These prove (d)–(f).

### *B.6. Proof of Theorem 12*

To verify the conditions for Theorem 3, we will use the following properties of $B$-splines.

For any $f \in \mathfrak{C}^\alpha[0,1]$, there exists $\beta_* \in \mathbb{R}^J$ with $\|\beta_*\|_\infty < \|f\|_{\mathfrak{C}^\alpha}$ such that
$$\|\beta_*^T B_J - f\|_\infty \lesssim J^{-\alpha}\|f\|_{\mathfrak{C}^\alpha}, \tag{82}$$
by the well-known approximation theory of B-splines [11, page 170]. Writing $f_\beta = \beta^T B_J$, this gives
$$\|f_\beta - f\|_{2,n} \le \|f_\beta - f\|_\infty \lesssim J^{-\alpha}\|f\|_{\mathfrak{C}^\alpha} + \|f_\beta - f_{\beta_*}\|_\infty. \tag{83}$$
We also use the following inequalities: for every $\beta \in \mathbb{R}^J$,
$$\|\beta\|_\infty \lesssim \|f_\beta\|_\infty \le \|\beta\|_\infty, \quad \|\beta\|_2 \lesssim \sqrt{J}\|f_\beta\|_{2,n} \lesssim \|\beta\|_2. \tag{84}$$
See Lemma E.6 of Ghosal and van der Vaart [17] for proofs with respect to $L_\infty$- and $L_2$-norms. Hence the first relation can be formally justified. For the second relation with respect to the empirical $L_2$-norm, we assume that $z_i$ are sufficiently regularly distributed as in (7.12) of Ghosal and van der Vaart [16].

- *Verification of* (C1): If $v_0$ is strictly positive on $[0,1]$, then $v_0$ satisfies the same approximation rule in (82) for some $\beta_* \in (0,\infty)^J$ with $\|\beta_*\|_\infty < \|v_0\|_{\mathfrak{C}^\alpha}$ (see Lemma E.5 of Ghosal and van der Vaart [17]). Therefore the approximation in (83) also holds for $v_0$ even if $\beta$ is restricted to have positive entries only, and thus by (82) and (84),

$$\|v_{\beta_*} - v_0\|_\infty \lesssim J^{-\alpha}, \quad \text{for some } \beta_* \in (0,\infty)^J,$$
$$\|v_{\beta_1} - v_{\beta_2}\|_\infty \lesssim \sqrt{J}\|v_{\beta_1} - v_{\beta_2}\|_{2,n}, \quad \beta_1, \beta_2 \in (0,\infty)^J,$$

which tells us that we have $a_n \asymp J$ and $e_n \asymp J^{1-2\alpha}$ for (C1).

- *Verification of* (C2): Note that if $J^{-\alpha} \lesssim \bar{\epsilon}_n$, it follows that for some $C_1 > 0$,

$$\log \Pi(\beta : \|v_\beta - v_0\|_{2,n} \le \bar{\epsilon}_n) \ge \log \Pi(\beta : \|\beta - \beta_*\|_\infty \le C_1 \bar{\epsilon}_n) \gtrsim J \log \bar{\epsilon}_n.$$

  This implies that condition (C2) is satisfied with $\bar{\epsilon}_n = \sqrt{(J \log n)/n}$.
- *Verification of* (C3): The assumption $\|\theta_0\|_\infty \lesssim \lambda^{-1} \log p$ given in the theorem directly satisfies (C3).
- *Verification of* (C4): Since $v_0$ is strictly positive on $[0,1]$ and belongs to a fixed multiple of the unit ball of $\mathfrak{C}^\alpha[0,1]$, we have that

$$1 \lesssim \inf_{z\in[0,1]} v_0(z) \le \sup_{z\in[0,1]} v_0(z) \lesssim 1.$$

  The condition (C4) is thus satisfied.
- *Verification of* (C5*): For a sufficiently large $M$, choose a sieve as $\mathcal{H}_n = \prod_{j=1}^J \{\beta_j : n^{-M} \le \beta_j \le n^M\}$. Then the minimum eigenvalue condition (6) is satisfied with $\log \gamma_n \asymp \log n$ because for every $i \le n$,

$$\inf_{\beta\in\mathcal{H}_n} v_\beta(z_i) = \inf_{\beta\in\mathcal{H}_n} \sum_{j=1}^J B_{J,j}(z_i)\beta_j \ge \inf_{\beta\in\mathcal{H}_n} \min_{1\le j\le J} \beta_j \sum_{j=1}^J B_{J,j}(z_i) \ge n^{-M},$$

  where $B_{J,j}$ and $\beta_j$ denote the $j$th components of $B_J$ and $\beta$, respectively. To check the entropy condition in (7), note that for every $\eta_1, \eta_2 \in \mathcal{H}_n$, we have $d_n(\eta_1,\eta_2) \lesssim \|\beta_1 - \beta_2\|_\infty$ by (84). Hence, for some $C_2 > 0$, the entropy in (7) is bounded above by a multiple of

$$\log N\left(\frac{1}{C_2 \overline{m} n^{M+3/2}}, \{\beta : \|\beta\|_\infty \le n^M\}, \|\cdot\|_\infty\right) \lesssim J \log n.$$

  The condition (8) holds since an inverse Gaussian prior on each $\beta_j$ produces $\Pi(\mathcal{H} \setminus \mathcal{H}_n) \lesssim J e^{-C_3 n^M}$ for some constant $C_3$, by its exponentially small bounds for tail probabilities on both sides. By matching $J^{-\alpha} \asymp \bar{\epsilon}_n$ and $n\bar{\epsilon}_n^2 \asymp J \log n$, we obtain $J \asymp (n/\log n)^{1/(2\alpha+1)}$ and $\bar{\epsilon}_n = (\log n/n)^{\alpha/(2\alpha+1)}$. Note that the conditions $a_n \epsilon_n^2 \to 0$ and $e_n \to 0$ hold only if $\alpha > 1/2$.
- *Verification of* (C6): The separation condition holds as there is no additional mean part.

Hence, we obtain the posterior contraction rates for $\theta$ by Theorem 3. The contraction rate for $v$ is also obtained by the same theorem. The assertion for the optimal posterior contraction is directly justified by Corollary 1. Hence we have verified (a)–(b).

Now, we verify (C8)–(C11) for the optimal posterior contraction in Theorem 4.

- *Verification of* (C8)–(C9): These conditions are trivially satisfied as there is no nuisance mean part.
- *Verification of* (C10): Note that by the inequality $\|v_\beta - v_0\|_{2,n} \lesssim \|v_\beta - v_{\beta_*}\|_{2,n} + \bar\epsilon_n$, the entropy in the integrand is bounded by

  $$\log N\left(\delta\sqrt{J}, \left\{\beta : \|\beta - \beta_*\|_2 \le C_4\sqrt{J}\bar\epsilon_n\right\}, \|\cdot\|_2\right) \lesssim 0 \vee J\log\left(\frac{3C_4\bar\epsilon_n}{\delta}\right),$$

  for some $C_4 > 0$. Thus, the second term of (C10) is bounded by $J\bar\epsilon_n$ by Remark 6, while the first term is bounded by $\sqrt{J\bar s_\star^2(\log p)/n}$. Since $\bar s_\star = (J\log n)/\log p \lesssim J$, (C10) is bounded by $J\bar\epsilon_n = (n/\log n)^{(1-\alpha)/(2\alpha+1)}$, which tends to zero as $\alpha > 1$.
- *Verification of* (C11): For every $v_{\beta_1}, v_{\beta_2} \in \widehat{\mathcal{H}}_n$, note that $d_{B,n}(\eta_1,\eta_2) = \|v_{\beta_1} - v_{\beta_2}\|_{2,n} \lesssim \|\beta_1 - \beta_2\|_2$ by (84). Since we put a prior for $v$ using the B-splines through a Euclidean parameter $\beta$, the separability is trivially satisfied.

Therefore, since (C7) is satisfied the assumption, assertion (c) holds by Theorem 4.

Next, we verify conditions (C8*)–(C10*) to apply Theorems 5–6 and Corollaries 2–3.

- *Verification of* (C8*)–(C9*): These are trivially satisfied for the same reason as before.
- *Verification of* (C10*): Similar to the verification of (C10), the entropy of interest is bounded by a constant multiple of $0 \vee J\log(3C_5\epsilon_n/\delta)$ for some $C_5 > 0$. Thus, (C10*) is bounded above by a multiple of $\{(s_\star^2 \vee J)J(s_\star\log p)^3/n\}^{1/2}$ by Remark 6, and hence goes to zero by the assumption. The condition $\alpha > 2$ is seen to be necessary by the inequality

  $$(s_\star^2 \vee J)J(s_\star\log p)^3/n \ge J^2n^2\bar\epsilon_n^6 = n^{2(-\alpha+2)/(2\alpha+1)}\log n^{2(3\alpha-1)/(2\alpha+1)}.$$

Under (C7*), the distributional approximation in (15) holds with the zero matrix $H$ by Theorem 5. Under (C7*) and (C12), the no-superset result in (16) holds by Theorem 6. We also obtain the strong results in Corollary 2 and Corollary 3 if the beta-min condition (C13) is also met. These prove (d)–(f).

### B.7. Proof of Theorem 13

We verify the conditions for the posterior contraction in Theorem 3.

- *Verification of* (C1): Since $\Delta_{\eta,i} = \sigma^2$ for every $i \le n$ and $\sigma_0^2$ belongs to the support of the prior, we have $a_n = 1$ and $e_n = 0$.
- *Verification of* (C2): Note that we write $d_n^2(\eta,\eta_0) = |\sigma^2 - \sigma_0^2|^2 + \|g_\beta - g_0\|_{2,n}^2$. To verify the prior concentration condition, observe that

  $$\begin{aligned}&\log\Pi\left(\eta \in \mathcal{H} : d_n(\eta,\eta_0) \le \bar\epsilon_n\right)\\ &\quad\ge \log\Pi\left(\beta : \|g_\beta - g_0\|_{2,n} \le \frac{\bar\epsilon_n}{\sqrt{2}}\right) + \log\Pi\left(\sigma : |\sigma^2 - \sigma_0^2| \le \frac{\bar\epsilon_n}{\sqrt{2}}\right),\end{aligned}$$

where the second term on the right hand side is trivially bounded below by a constant multiple of $\log \bar{\epsilon}_n$. Using (82)–(84), it is easy to see that if $J^{-\alpha} \lesssim \bar{\epsilon}_n$,

$$\log \Pi\left(\beta : \|g_\beta - g_0\|_{2,n} \le \frac{\bar{\epsilon}_n}{\sqrt{2}}\right) \ge \log \Pi(\beta : \|\beta - \beta_*\|_\infty \le C_1 \bar{\epsilon}_n) \gtrsim J \log \bar{\epsilon}_n,$$

for some $C_1 > 0$. Since $\bar{\alpha} \le \alpha$, this implies that (C2) is satisfied with $\bar{\epsilon}_n = \sqrt{(J \log n)/n}$.

- *Verification of* (C3): The assumption $\|\theta_0\|_\infty \lesssim \lambda^{-1} \log p$ given in the theorem directly satisfies the condition.
- *Verification of* (C4): This is directly satisfied by $\sigma_0^2 \asymp 1$.
- *Verification of* (C5*): For a sufficiently large constant $M$ and $s_\star = s_0 \vee (J \log n/\log p)$, choose $\mathcal{H}_n = \{g_\beta : \|\beta\|_\infty \le n^M\} \times \{\sigma : n^{-M} \le \sigma^2 \le e^{M s_\star \log p}\}$, from which the minimum eigenvalue condition (6) is directly satisfied with $\log \gamma_n \asymp \log n$. To check the entropy condition in (7), note that for every $\eta_1, \eta_2 \in \mathcal{H}_n$, we have $d_n^2(\eta_1, \eta_2) \lesssim \|\beta_1 - \beta_2\|_\infty^2 + |\sigma_1^2 - \sigma_2^2|^2$ by (84). Hence, for some $C_3 > 0$, the entropy in (7) is bounded above by a multiple of

$$\log N\left(\frac{1}{C_3 \overline{m} n^{M+3/2}}, \{\beta : \|\beta\|_\infty \le n^M\}, \|\cdot\|_\infty\right) \\ + \log N\left(\frac{1}{C_3 \overline{m} n^{M+3/2}}, \{\sigma : \sigma^2 \le e^{M s_\star \log p}\}, |\cdot|\right).$$

  The display is further bounded by a multiple of $J \log n + s_\star \log p$, and hence (7) is satisfied with $\epsilon_n = \sqrt{(s_\star \log p)/n}$. Using the tail bounds of normal and inverse gamma distributions, condition (8) is also satisfied.
- *Verification of* (C6): The separation condition holds by Remark 3 as we have $d_{A,n}(\eta_*, \eta_0) = \|g_{\beta_*} - g_0\|_{2,n} \lesssim \bar{\epsilon}_n$ for $\eta_* = (g_{\beta_*}, \sigma_0^2)$ in view of (82).

Therefore, the contraction rates for $\theta$ are given by Theorem 3. The rate for $g$ is also obtained by the same theorem. The assertion for the optimal posterior contraction is directly justified by Corollary 1. We thus see (a)–(b) hold.

Now, we verify (C8)–(C11) for Theorem 4.

- *Verification of* (C8): Observe that the left hand side of the first line of (C8) is equal to

$$\frac{1}{(s_0 \vee 1) \log p} \|\tilde{\xi}_{\eta_0} - H \tilde{\xi}_{\eta_0}\|_2^2 = \frac{n}{\sigma_0^2 (s_0 \vee 1) \log p} \|g_0 - \hat{\beta}_J^T B_J\|_{2,n}^2,$$

  where $\hat{\beta}_J = (W_J^T W_J)^{-1} W_J^T (g_0(z_1), \ldots, g_0(z_n))^T$ is the the least squares solution. Since $\hat{\beta}_J$ is the solution minimizing $\|g_0 - \hat{\beta}_J^T B_J\|_{2,n}^2$, for some $\beta_* \in \mathbb{R}^J$, the last display is bounded above by

$$\frac{n}{\sigma_0^2 \log p} \|g_0 - \beta_*^T B_J\|_\infty^2 \lesssim \frac{n}{J^{2\alpha} \log p}, \qquad (85)$$

  by (82), where $s_0 \vee 1$ is replaced by 1 as $s_0$ is unknown. Plugging in $J \asymp (n/\log n)^{1/(2\bar{\alpha}+1)}$, it is easy to see that the right hand side of (85) is the same order of $(\log n)^{2\alpha/(2\bar{\alpha}+1)} n^{(-2\alpha+2\bar{\alpha}+1)/(2\bar{\alpha}+1)}/\log p$. This tends to zero by the given boundedness assumption. The necessary condition $\bar{\alpha} < \alpha$ is implied by this, because $\log p = o(n)$. The second condition of (C8) is satisfied by Remark 5.

- *Verification of* (C9): Let $\tilde{\eta}_n(\theta,\eta) = (g_\beta(\cdot) + B_J^T(\cdot)(W_J^T W_J)^{-1} W_J^T X(\theta - \theta_0), \sigma^2)$ for a given $\theta$, where $\eta = (g_\beta(\cdot), \sigma^2)$. This setting satisfies $\Phi(\tilde{\eta}_n(\theta,\eta)) = (\tilde{\xi}_\eta + H\tilde{X}(\theta - \theta_0), \tilde{\Delta}_\eta)$. Since each entry of $\beta$ has the standard normal prior, $g_\beta(\cdot)$ is a zero mean Gaussian process with the covariance kernel $K(t_1, t_2) = B_J(t_1)^T B_J(t_2)$, and thus its reproducing kernel Hilbert space (RKHS) $\mathbb{K}$ is the set of all functions of the form $\sum_k \zeta_k B_J(t_k)^T B_J(\cdot)$ with coefficients $\zeta_k$, $k \in \{1, 2, \dots\}$. It is easy to see that the shift $(\theta-\theta_0)^T X^T W_J (W_J^T W_J)^{-1} B_J(\cdot)$ is in the RKHS $\mathbb{K}$ since it is expressed as $(\theta-\theta_0)^T X^T W_J (W_J^T W_J)^{-1} \tilde{W}_J^{-1} \tilde{W}_J B_J(\cdot)$ using an invertible matrix $\tilde{W}_J \in \mathbb{R}^{J\times J}$ with rows $B_J(t_k)$ evaluated by some $t_k$, $k = 1, \dots, J$. Hence, by the Cameron-Martin theorem, for $\nu = (\nu_1, \dots, \nu_J)^T = (\tilde{W}_J^T)^{-1}(W_J^T W_J)^{-1} W_J^T X(\theta - \theta_0)$ and $\|\cdot\|_{\mathbb{K}}$ the RKHS norm, we see that
$$\log \frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}}(\eta) = \sum_{k=1}^J \nu_k g_\beta(t_k) - \frac{1}{2}\|\nu^T \tilde{W}_J B_J\|_{\mathbb{K}}^2 = \nu^T \tilde{W}_J \beta - \frac{1}{2}\|\tilde{W}_J^T \nu\|_2^2,$$
almost surely. This gives that
$$\begin{aligned}\left|\log \frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}}(\eta)\right| &\lesssim \|\beta\|_2 \|(W_J^T W_J)^{-1} W_J^T X(\theta - \theta_0)\|_2 \\ &\quad + \|(W_J^T W_J)^{-1} W_J^T X(\theta - \theta_0)\|_2^2.\end{aligned} \tag{86}$$
Note that we have
$$\begin{aligned}\sup_{\eta\in\tilde{\mathcal{H}}_n} \|\beta\|_2 &\le \sup_{\eta\in\tilde{\mathcal{H}}_n} \|\beta - \beta_*\|_2 + \|\beta_*\|_2 \\ &\lesssim \sqrt{J} \sup_{\eta\in\tilde{\mathcal{H}}_n} \|g_\beta - g_{\beta_*}\|_{2,n} + 1 \lesssim \sqrt{J}\bar{\epsilon}_n + 1,\end{aligned}$$
and
$$\sup_{\theta\in\tilde{\Theta}_n} \|(W_J^T W_J)^{-1} W_J^T X(\theta - \theta_0)\|_2 \lesssim \frac{\|W_J\|_{\mathrm{sp}} \sup_{\theta\in\tilde{\Theta}_n} \|X(\theta - \theta_0)\|_2}{\rho_{\min}(W_J^T W_J)} \lesssim \sqrt{J}\bar{\epsilon}_n,$$
using (84). Since $\sqrt{J}\bar{\epsilon}_n$ is bounded due to $\bar{\alpha} \ge 1/2$, (86) is bounded.
- *Verification of* (C10): Since the entropy in the integral in (C10) is bounded above by a multiple of $0 \vee \log(3\tilde{M}_2\bar{\epsilon}_n/\delta)$ for every $\delta > 0$, the second term of (C10) is bounded by a constant multiple of $\bar{\epsilon}_n$ due to Remark 6. The first term is $\bar{\epsilon}_n^2 \sqrt{n/\log p} = (\log n)^{2\bar{\alpha}/(2\bar{\alpha}+1)} n^{(-\bar{\alpha}+1/2)/(2\bar{\alpha}+1)}/\sqrt{\log p}$ that tends to zero by the boundedness assumption.
- *Verification of* (C11): Since we have $d_{B,n}(\eta_1, \eta_2) = |\sigma_1^2 - \sigma_2^2|$ for every $\sigma_1^2, \sigma_2^2 \in (0, \infty)$ and the parameter space of $\sigma^2$ is Euclidean, the condition is trivially satisfied.

Therefore, assertion (c) holds by Theorem 4 since (C7) is also satisfied by the given assumption.

Lastly, we verify conditions (C8*)–(C10*) to apply Theorems 5–6 and Corollaries 2–3.

- *Verification of* (C8*): Similar to the verification of (C8), the first line of (C8*) is equal to
$$s_\star^2 \log p \|\tilde{\xi}_{\eta_0} - H\tilde{\xi}_{\eta_0}\|_2^2 \lesssim \frac{n s_\star^2 \log p}{J^{2\alpha}}.$$
Plugging in $J \asymp (n/\log n)^{1/(2\bar{\alpha}+1)}$, it is easy to see that this tends to zero by the given boundedness condition, which requires that $\bar{\alpha} < \alpha - 1/2$.
- *Verification of* (C9*): Similar to the verification of (C9), we now have
$$\begin{aligned}\sup_{\eta \in \widehat{\mathcal{H}}_n} \|\beta\|_2 &\lesssim s_\star \sqrt{(J \log p)/n} + 1, \\ \sup_{\theta \in \widehat{\Theta}_n} \|(W_J^T W_J)^{-1} W_J^T X(\theta - \theta_0)\|_2 &\lesssim \frac{\|W_J\|_{\mathrm{sp}} \sup_{\theta \in \widehat{\Theta}_n} \|X(\theta - \theta_0)\|_2}{\rho_{\min}(W_J^T W_J)} \\ &\lesssim s_\star \sqrt{(J \log p)/n}.\end{aligned}$$
Since $J \log n = n\bar{\epsilon}_n^2 \le s_\star \log p$, (86) tends to zero since $s_\star^5 \log^3 p = o(n)$.
- *Verification of* (C10*): By the similar calculations as before, we see that (C10*) is bounded by $(s_\star^5 \log^3 p/n)^{1/2}$ which tends to zero. The condition $\bar{\alpha} > 1$ is necessary since $(s_\star^5 \log^3 p)/n \ge n^2 \bar{\epsilon}_n^6 = (\log n)^{6\bar{\alpha}/(2\bar{\alpha}+1)} n^{-2(\bar{\alpha}-1)/(2\bar{\alpha}+1)}$.

Therefore, under (C7*), we have the distributional approximation in (15) by Theorem 5. Under (C7*) and (C12), Theorem 6 implies that the no-superset result in (16) holds. The stronger assertions in (17) and (18) are explicitly derived from Corollary 2 and Corollary 3 if the beta-min condition (C13) is also met.

## Appendix C: Auxiliary results

Here we provide some auxiliary results used to prove the main results.

**Lemma 9.** *Let $p_k$ be the density of $\mathrm{N}_r(\mu_k, \Sigma_k)$ for $k = 1, 2$. Then,*
$$\begin{aligned}K(p_1, p_2) =& \frac{1}{2}\left\{\log \frac{\det \Sigma_2}{\det \Sigma_1} + \mathrm{tr}(\Sigma_1 \Sigma_2^{-1}) - r + \|\Sigma_2^{-1/2}(\mu_1 - \mu_2)\|_2^2\right\}, \\ V(p_1, p_2) =& \frac{1}{2}\Big\{\mathrm{tr}(\Sigma_1 \Sigma_2^{-1} \Sigma_1 \Sigma_2^{-1}) - 2\mathrm{tr}(\Sigma_1 \Sigma_2^{-1}) + r\Big\} + \|\Sigma_1^{1/2} \Sigma_2^{-1}(\mu_1 - \mu_2)\|_2^2.\end{aligned}$$

*Proof.* Let $Z = \Sigma_1^{-1/2}(X - \mu_1) \sim \mathrm{N}_r(0, I)$ for $X \sim p_1$ and $A = \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2}$. Then by direct calculations, we have
$$\begin{aligned}K(p_1, p_2) &= \mathbb{E}_{p_1}\left\{\log \frac{p_1}{p_2}(X)\right\} \\ &= \frac{1}{2}\left\{\log \frac{\det \Sigma_2}{\det \Sigma_1} + \mathbb{E}_{p_1} Z^T A Z - r + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2)\right\},\end{aligned}$$

which verifies the first assertion because $\mathbb{E}_{p_1} Z^T A Z = \mathrm{tr} A$. After some algebra, we also obtain

$$\begin{aligned} V(p_1, p_2) &= \mathbb{E}_{p_1} \left\{ \log \frac{p_1}{p_2}(X) - K(p_1, p_2) \right\}^2 \\ &= \frac{1}{4} \mathbb{E}_{p_1} \left\{ -Z^T Z + Z^T A Z + 2(\mu_1 - \mu_2)^T \Sigma_2^{-1} \Sigma_1^{1/2} Z - \mathrm{tr}(A) + r \right\}^2 . \end{aligned}$$

The rightmost side involves forms of $\mathbb{E}_{p_1}(Z Z^T Q_1 Z)$ and $\mathbb{E}_{p_1}(Z^T Q_1 Z Z^T Q_2 Z)$ for two positive definite matrices $Q_1$ and $Q_2$. It is easy to see that the former is zero, while it can be shown the latter equals $2\mathrm{tr}(Q_1 Q_2) + \mathrm{tr}(Q_1)\mathrm{tr}(Q_2)$; for example, see Lemma 6.2 of Magnus [22]. Plugging in this for the expected values of the products of quadratic forms, it is easy (but tedious) to verify the second assertion. □

**Lemma 10.** *For $r \times r$ positive definite matrices $\Sigma_1$ and $\Sigma_2$, let $d_1, \ldots, d_r$ be the eigenvalues of $\Sigma_2^{1/2} \Sigma_1^{-1} \Sigma_2^{1/2}$. Then the following assertions hold:*

(i) $\rho_{\max}^{-2}(\Sigma_2) \|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}^2 \le \sum_{k=1}^r (d_k^{-1} - 1)^2 \le \rho_{\min}^{-2}(\Sigma_2) \|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}^2$,

(ii) $\max_k |d_k - 1|$ *can be made arbitrarily small if $g^2(\Sigma_1, \Sigma_2)$ is chosen sufficiently small, where $g$ is defined in* (33).

*Proof.* Let $A = \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2}$. Since the eigenvalues of $A - I_r$ are $d_1^{-1} - 1, \ldots, d_r^{-1} - 1$, we can see that $\|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}^2$ is equal to

$$\|\Sigma_2^{1/2} (A - I_r) \Sigma_2^{1/2}\|_{\mathrm{F}}^2 \le \rho_{\max}^2(\Sigma_2) \|A - I_r\|_{\mathrm{F}}^2 = \rho_{\max}^2(\Sigma_2) \sum_{k=1}^r (d_k^{-1} - 1)^2 .$$

Conversely, using the sub-multiplicative property of the Frobenius norm, $\|BC\|_{\mathrm{F}} \le \|B\|_{\mathrm{sp}} \|C\|_{\mathrm{F}}$, it can be seen that $\sum_{k=1}^r (d_k^{-1} - 1)^2$ is equal to

$$\|A - I_r\|_{\mathrm{F}}^2 = \|\Sigma_2^{-1/2} (\Sigma_1 - \Sigma_2) \Sigma_2^{-1/2}\|_{\mathrm{F}}^2 \le \rho_{\max}^2(\Sigma_2^{-1}) \|\Sigma_1 - \Sigma_2\|_{\mathrm{F}}^2 .$$

These verify (i). Now, note that by direct calculations,

$$\begin{aligned} \frac{(\det \Sigma_1)^{1/4} (\det \Sigma_2)^{1/4}}{\det((\Sigma_1 + \Sigma_2)/2)^{1/2}} &= \left\{ \frac{1}{2^r} \det(A^{1/2} + A^{-1/2}) \right\}^{-1/2} \\ &= \left\{ \prod_{k=1}^r \frac{1}{2} (d_k^{1/2} + d_k^{-1/2}) \right\}^{-1/2} . \end{aligned}$$

Hence, $g^2(\Sigma_1, \Sigma_2) < \delta$ for a sufficiently small $\delta > 0$ implies that

$$\prod_{k=1}^r \frac{1}{2} (d_k^{1/2} + d_k^{-1/2}) < (1 - \delta^2/2)^{-2} .$$

Since every term in the product of the last display is greater than or equal to 1, we have $(d_k^{1/2} + d_k^{-1/2})/2 < (1 - \delta^2/2)^{-2}$ for every $k$. As a function of $d_k$, $(d_k^{1/2} + d_k^{-1/2})/2$ has the global minimum at $d_k = 1$, and hence $\delta$ can be chosen sufficiently small to make $|d_k - 1|$ small for every $k = 1, \ldots, r$, which establishes (ii). □

## References

[1] Atchadé, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *The Annals of Statistics*, 45(5):2248–2273. MR3718168

[2] Bai, R., Moran, G. E., Antonelli, J., Chen, Y., and Boland, M. R. (2020). Spike-and-slab group lassos for grouped regression and sparse generalized additive models. *Journal of the American Statistical Association*, to appear.

[3] Belitser, E. and Ghosal, S. (2020). Empirical Bayes oracle uncertainty quantification for regression. *The Annals of Statistics*, 48(6):3113–3137. MR4185802

[4] Bickel, P. J. and Kleijn, B. J. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237. MR3013185

[5] Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624. MR3036420

[6] Carroll, R. J., Ruppert, D., Crainiceanu, C. M., and Stefanski, L. A. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC. MR2243417

[7] Castillo, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152(1-2):53–99. MR2875753

[8] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018. MR3375874

[9] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101. MR3059077

[10] Chae, M., Lin, L., and Dunson, D. B. (2019). Bayesian sparse linear regression with unknown symmetric error. *Information and Inference: A Journal of the IMA*, 8(3):621–653. MR3994400

[11] De Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer. MR0507062

[12] Fikioris, G. (2018). Spectral properties of Kac–Murdock–Szegö matrices with a complex parameter. *Linear Algebra and its Applications*, 553:182–210. MR3809375

[13] Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons. MR0898653

[14] Gao, C., van der Vaart, A. W., and Zhou, H. H. (2020). A general framework for Bayes structured linear models. *The Annals of Statistics*, 48(5):2848–2878. MR4152123

[15] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531. MR1790007

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

223. MR2332274

[17] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. MR3587782

[18] Jeong, S. (2020). Posterior contraction in group sparse logit models for categorical responses. *arXiv preprint* arXiv:2010.03513.

[19] Jeong, S. and Ghosal, S. (2021). Posterior contraction in sparse generalized linear models. *Biometrika*, 108(2):367–379. MR4259137

[20] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660. MR2980074

[21] Kulkarni, D., Schmidt, D., and Tsui, S.-K. (1999). Eigenvalues of tridiagonal pseudo-Toeplitz matrices. *Linear Algebra and its Applications*, 297:63–80. MR1723838

[22] Magnus, J. R. (1978). The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32(4):201–210. MR0528403

[23] Martin, R., Mess, R., and Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847. MR3624879

[24] Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817. MR3210987

[25] Ning, B., Jeong, S., and Ghosal, S. (2020). Bayesian linear regression for multivariate responses under group sparsity. *Bernoulli*, 26(3):2353–2382. MR4091112

[26] Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437. MR3766957

[27] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515. MR2417391

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

[29] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer. MR1385671