Posterior contraction rates in a sparse non-linear mixed-effects model

Marion Naveau, Maud Delattre, Laure Sansonnet

Posterior contraction rates in a sparse non-linear mixed-effects model

Marion Naveau
Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris-Saclay, 75005, Paris, France.
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.

Maud Delattre
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.

Laure Sansonnet
Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris-Saclay, 75005, Paris, France.

Abstract

Recent works have shown an interest in investigating the frequentist asymptotic properties of Bayesian procedures for high-dimensional linear models under sparsity constraints. However, there exists a gap in the literature regarding analogous theoretical findings for non-linear models within the high-dimensional setting. The current study provides a novel contribution, focusing specifically on a non-linear mixed-effects model. In this model, the residual variance is assumed to be known, while the covariance matrix of the random effects and the regression vector are unknown and must be estimated. The prior distribution for the sparse regression coefficients consists of a mixture of a point mass at zero and a Laplace distribution, while an Inverse-Wishart prior is employed for the covariance parameter of the random effects. First, the effective dimension of this model is bounded with high posterior probabilities. Subsequently, we derive posterior contraction rates for both the covariance parameter and the prediction term of the response vector. Finally, under additional assumptions, the posterior distribution is shown to contract for recovery of the unknown sparse regression vector at the same rate as observed in the linear case.

Keywords Posterior contraction rate · Sparse priors · Non-linear mixed-effects models · High-dimensional regression

1 Introduction

Recent statistical literature has shown a keen interest in estimating high-dimensional models under sparsity assumptions, with different approaches proposed over the past few decades in both Bayesian and frequentist frameworks. The developed methodologies are numerous and use a large variety of techniques such as convex and non-convex penalization techniques, shrinkage methods and sparsity-inducing priors. In Bayesian analysis, a category of proposed priors includes those defined as mixtures of two distributions, commonly referred to as spike-and-slab priors. These priors have proven to be useful and relevant in many high-dimensional applications as demonstrated in (George and McCulloch, 1993, 1997; Tadesse and Vannucci, 2021).

The frequentist asymptotic properties of Bayesian sparse linear regression models with various types of mixture priors have been widely investigated, particularly in Narisetty and He (2014), Castillo et al. (2015), and Ročková and George (2018) with the spike-and-slab Gaussian prior, the discrete spike-and-slab prior, and the spike-and-slab lasso prior respectively. Subsequently, these investigations were extended to multivariate linear regression with an unknown residual covariance matrix, as discussed by Ning et al. (2020) with a discrete spike-and-slab prior and Shen and Deshpande (2022) with a multivariate spike-and-slab lasso prior.

The classical techniques for determining posterior contraction rates (see *e.g.* (Castillo et al., 2015)) face limitations when the residual covariance matrix is unknown. In such scenarios, the general theory based on the average squared Hellinger distance proves inadequate for obtaining rates in terms of the Euclidean norm for parameters. To overcome this difficulty, an alternative approach has been introduced, leveraging the average Rényi divergence of order 1/2. As underscored by Ning et al. (2020), this method enables the construction of exponentially powerful tests that are required by the general theory (Ghosal and Van der Vaart, 2017), facilitating a more effective determination of posterior contraction rates in Bayesian analysis. Another theoretical aspect requires adaptation to the general theory when the residual covariance matrix is unknown. Indeed, classical proofs require lower bounds for prior mass around true parameter values, but when the residual covariance matrix is unknown, this condition can only be fulfilled if the true parameter set is bounded, as discussed in works of Ning et al. (2020) and Jeong and Ghosal (2021a,b).

Recent advancements have expanded the study of estimation and selection properties to more complex models than sparse linear regression models, such as sparse generalized linear models (Jiang, 2007; Jeong and Ghosal, 2021a) or sparse linear regression models with nuisance parameters (Jeong and Ghosal, 2021b). To our knowledge, there are no similar theoretical results for non-linear models in high-dimensional contexts. The absence of theoretical results in this domain may reflect the inherent challenges and complexities associated with extending such analyses to non-linear models. The present paper provides a contribution in this direction, focusing on a specific non-linear model which also contains random effects. Mixed-effects models have been introduced to analyze observations collected repeatedly on several individuals in a population of interest, commonly encountered in fields such as pharmacokinetics or biological growth modeling for example (Pinheiro and Bates, 2000; Lavielle, 2014). These models, which are generally non-linear, may use high-dimensional covariates to describe inter-individual variability. Our paper deals with a generalization of the linear mixed-effects model to a non-linear marginal version where the fixed effects are non-linear functions of the regression parameter, while the random effects are incorporated into the model in a linear manner (see *e.g.* Demidenko (2013)). Such non-linear marginal mixed-effects models are easier to handle than more general non-linear mixed-effects models because the mean and the covariance matrix of the response variable are explicit. However, despite their practical appeal, there has been a lack of theoretical exploration concerning non-linear marginal mixed models in high-dimensional context. In this paper, posterior contraction rates are obtained for both the covariance matrix and the prediction term in a high-dimensional setting by using a mixture of a point mass at zero and a Laplace distribution prior on the regression coefficients and an inverse Wishart prior on the covariance matrix. These results are extended to the regression coefficients themselves under additional assumptions.

This paper is organized as follows. Section 2 describes the non-linear marginal mixed model to introduce the notation, defines the prior distributions, along with the necessary assumptions. Section 3 provides the main results on the posterior contraction. Finally, the proofs of the theorems are given in Section 4. Proofs of technical lemmas are postponed in Appendix.

Notation This paragraph describes the notations used in this paper for a generic matrix A and a generic vector $\theta \in \mathbb{R}^k$. We note $S_\theta = \{j|\theta_j \neq 0\}$ the support of θ and $s_\theta = |S_\theta|$ its cardinal. The euclidean norm, the ℓ_1-norm and the infinity norm are respectively noted $\|\theta\|_2 = \sqrt{\sum_{i=1}^k \theta_i^2}$, $\|\theta\|_1 = \sum_{i=1}^k |\theta_i|$, and $\|\theta\|_\infty = \max_i \|\theta_i\|$. The transpose of A is denoted by $A^\top$. For a square matrix A, we note $\rho_{min}(A)$ and $\rho_{max}(A)$ the minimum and maximum eigenvalues of A, respectively. The spectral norm of a matrix A is denoted $\|A\|_{sp} = \rho_{max}^{1/2}(A^\top A)$, and the Frobenius norm is noted $\|A\|_F = \mathrm{Tr}(A^\top A)^{1/2} = (\sum_{i,j} x_{ij}^2)^{1/2}$. The matrix norm $\|A\|_*$ is defined as $\|A\|_* = \max_j \|A_{\cdot j}\|_2$ for $A_{\cdot j}$ the j-th column of A. The identity matrix of size m is denoted I_m.

For sequences a_n and b_n, the notation $a_n \lesssim b_n$ means that for n large enough a_n is bounded above by a constant multiple of b_n, *i.e.* $a_n \leq Cb_n$ for n large enough, where $C > 0$ is independent of n. We denote $a_n = o(b_n)$ if $\dfrac{a_n}{b_n} \underset{n\to\infty}{\longrightarrow} 0$.

2 Model description

2.1 Non-linear marginal mixed-effects model

Mixed-effects models are sophisticated multivariate statistical models employed to analyze repeated observations, usually collected over time, on multiple statistical subjects, incorporating both fixed and random effects into the model for accurate description (Lavielle, 2014; Demidenko, 2013). We consider the following mixed-effects model:

$$Y_i = f_i(X_i\beta) + Z_i\xi_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{ind.}}{\sim} \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i}), \quad \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, \Gamma), \quad i = 1, \dots, n. \tag{1}$$

In the above equation, n is the number of individuals, $Y_i \in \mathbb{R}^{n_i}$ and n_i are the vector and the number of observations for subject i respectively, $\xi_i \in \mathbb{R}^q$ is a vector composed of q random effects, $\varepsilon_i \in \mathbb{R}^{n_i}$ is the error term. $X_i \in \mathbb{R}^{q\times p}$ and $Z_i \in \mathbb{R}^{n_i\times q}$ are design matrices composed of explanatory variables that relate the observations to fixed-effects $\beta \in \mathbb{R}^p$ and to the random effects respectively. In the development of a mixed-effects model, a significant focus lies on the covariate selection process in X_i, which amounts to the identification of non-null components in β, as it allows establishing connections between inter-individual variability and measured individual characteristics. In the above model, the function f_i defines the relationship between observations of subject i and the explanatory variables in X_i which are pivotal in this stage of model construction and are commonly denoted as covariates in subsequent discussions. To draw parallels with the classical literature on mixed-effects models for longitudinally repeated data, we define

$$f_i(x) = (f(x, t_{i1}), f(x, t_{i2}), \dots, f(x, t_{in_i}))^\top,$$

where t_{ij} represents the j-th observation time for individual i, and $f : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$ denotes a regression function chosen to effectively capture the longitudinally observed phenomenon. While opting for $f_i(X_i\beta) = X_i\beta$ yields the standard linear mixed-effects model, it's common in many applications to select a non-linear function f. Moreover, when f is non-linear, Model (1) is alternatively referred to as the non-linear marginal mixed-effects model (as discussed in (Demidenko, 2013)). The term "marginal mixed-effects model" is derived from the fact that, unlike numerous other non-linear mixed-effects models, both the expectation and variance of the observations possess an explicit expression. The distribution of Y_i defined through (1) is thus fully characterized:

$$Y_i \sim \mathcal{N}(f_i(X_i\beta), \Delta_{\Gamma,i}), \text{ where } \Delta_{\Gamma,i} = Z_i\Gamma Z_i^\top + \sigma^2 I_{n_i}. \tag{2}$$

In the following, the residual variance σ^2 and the number q of true random effects are assumed to be known. The aim is to estimate $(\beta, \Gamma) \in \mathcal{B} \times \mathcal{H}$ in an high-dimensional setting where $p >> n$ and obtain posterior contraction results. We establish below appropriate priors to achieve these goals.

2.2 Prior specification

Drawing from classical literature in high-dimensional Bayesian analysis, this study adopts an approach employing priors that induce sparsity in β coefficients. For that purpose, we jointly consider a prior π_p on the number s of non-zero coefficients in β and a Laplace prior on the non-zero coefficients in β while setting the other components in β to zero:

$$(S, \beta) \mapsto \frac{\pi_p(s)}{\binom{p}{s}} g_S(\beta_S)\delta_0(\beta_{S^c}), \tag{3}$$

where S is a subset of s elements in $\{1, \dots, p\}$ that represents the support of β, *i.e.* the positions of its non-zero elements, S^c is the complementary subset of zero coefficients in β, $\beta_S = \{\beta_\ell | \ell \in S\}$ and $\beta_{S^c} = \{\beta_\ell | \ell \notin S\}$ are the non-zero and the zero coefficients in β respectively, δ_0 is the Dirac measure at zero on $\mathbb{R}^{p-s}$ and

$$g_S(\beta_S) = \prod_{\ell\in S} \frac{\lambda}{2} \exp(-\lambda|\beta_\ell|). \tag{4}$$

Concerning the random effects covariance matrix Γ, a conjugate inverse-Wishart prior is used:

$$\pi(\Gamma) \propto |\Gamma|^{-\frac{d+q+1}{2}} \exp\left(-\frac{1}{2}\text{Tr}(\Sigma\Gamma^{-1})\right),$$

where Σ is a positive definite matrix of size $q \times q$, and $d > q - 1$ the degree of freedom. This prior is chosen for a practical matter. Note that, as discussed in Ning et al. (2020), the inverse-Wishart prior may induce sub-optimal posterior contraction rate due to its weaker tail property when q increases to infinity. However, here q is assumed to be fixed so the rate should not be impacted by this property.

2.3 Assumptions

The frequentist assumption that the data, n independent observations $Y^{(n)} = (Y_i)_{1\leq i\leq n} \in \mathbb{R}^N$, where $N = \sum_{i=1}^n n_i$, has been generated from model (1) for a given sparse regression parameter β_0 and a given random effects covariance matrix Γ_0 is made. The expectation under these true parameters is denoted $\mathbb{E}_0$. The support of the true parameter β_0 is denoted S_0 and its cardinal s_0. The maximum number of observations per individual is defined as $J_n = \max_{1\leq i\leq n} n_i$.

2.3.1 Assumptions on the non-linear model structure

Assumptions have to be made on the regression function f to obtain posterior contraction. A first natural condition is the Lipschitz assumption, allowing for easy control of the regression function from its inputs.

Assumption A1. *$f : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$ is K-Lipschitz with respect to its first component:*

$$\forall x, y \in \mathbb{R}^q, \forall t \in \mathbb{R}, |f(x,t) - f(y,t)| \leq K\|x-y\|_2$$

Remark 1. *Under assumption A1, notice we have that $f_i : \mathbb{R}^q \to \mathbb{R}^{n_i}$ is $\sqrt{K^2 n_i}$-Lipschitz for $\|\cdot\|_2$.*

As outlined in the introduction, to satisfy the condition of prior mass around the true parameters, they should be confined within a specific subset of the parameter space characterized by bounded norms.

Assumption A2. *The true value β_0 belongs to $\mathcal{B}_0 := \left\{\beta \in \mathbb{R}^p : \|\beta\|_\infty \lesssim \lambda^{-1}\log(p)\right\}$, where λ is the regularization parameter of the Laplace distribution defined in equation* (4)*.*

Assumption A3. *The true covariance matrix of the random effects Γ_0 belongs to $\mathcal{H}_0 := \{\Gamma : 1 \lesssim \rho_{min}(\Gamma) \leq \rho_{max}(\Gamma) \lesssim 1\}$, and we denote $\underline{\rho_\Gamma} > 0$ and $\overline{\rho_\Gamma} > 0$ the bounds such that: $\underline{\rho_\Gamma} \leq \rho_{min}(\Gamma_0) \leq \rho_{max}(\Gamma_0) \leq \overline{\rho_\Gamma}$.*

Assumption A2 allows that the prior assigns sufficient mass on a Kullback-Leibler neighborhood of β_0. In the same way, assumption A3 enables to put sufficient mass around the true parameter Γ_0 in terms of Frobenius norm. Similar conditions can be found in the work of Ning et al. (2020), Jeong and Ghosal (2021b), and Song and Liang (2023). This is in contrast to Castillo et al. (2015)'s work where they obtain a result uniformly over the entire parameter space because they have explicit expressions to satisfy this condition directly in their case of univariate regression with known variance. Also, it is assumed that β_0 is not the zero vector, and that p does not converge faster than exponential of n (see assumption A4).

Assumption A4. *The true support size satisfies $s_0 > 0$ and the following high-dimensional setting is consider $s_0 \log(p) = o(n)$.*

2.3.2 Assumptions on the prior distributions

The importance of the prior π_p lies in its essential role in representing the sparsity of the parameter. The crucial aspect of the prior π_p on model dimension is to appropriately reduce the influence of larger models while maintaining sufficient weight for the true one. It is revealed that an exponential decrease effectively fulfills this requirement (Castillo et al., 2015). The following assumption is made on π_p.

Assumption A5 (Prior dimension)**.** *For some constants A_1, A_2, A_3, $A_4 > 0$,*

$$A_1 p^{-A_3}\pi_p(s-1) \leq \pi_p(s) \leq A_2 p^{-A_4}\pi_p(s-1) \qquad , \textit{for } s = 1, \ldots, p$$

Examples of priors satisfying this assumption A5 are given in Castillo and van der Vaart (2012) and Castillo et al. (2015). In fact, this type of prior is more generic than the discrete spike-and-slab prior. Indeed, if each coordinate β_ℓ is modeled as a mixture $(1-r)\delta_0 + rG$, where G follows the Laplace distribution, it can be realized as a prior of the form (3) by selecting π_p as a binomial distribution with parameters p and r. Since r controls the level of sparsity, which is unknown, a classical Bayesian strategy is to put a hyper-prior $Beta(1, p^u)$ with $u > 1$. Then, the overall prior satisfies the exponential decay rate A5. Furthermore, the regularization parameter of the Laplace prior λ must be bounded from below and above, as specified in the following assumption. Indeed, an excessively large value of λ would shrink non-zero coordinates of β towards 0, which is undesirable. Conversely, a too small value of λ may introduce false signals in the support, thereby slowing down the posterior contraction rate.

Assumption A6. *The regularization parameter λ of the Laplace prior on the non-zero coordinates of β satisfies:*

$$\frac{\|X\|_* K_n}{L_1 p^{L_2}} \leq \lambda \leq \frac{L_3\|X\|_* K_n}{\sqrt{n}},$$

for some constants L_1, L_2, $L_3 > 0$, *where* $K_n = \sqrt{K^2 J_n}$ *and* $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^{nq \times p}$.

Similar condition can be found in Jeong and Ghosal (2021b) for example.

2.3.3 Assumptions about the experimental design

For $\Gamma_1, \Gamma_2 \in \mathcal{H}$, we define the pseudo distance $d_n^2(\Gamma_1, \Gamma_2) = \dfrac{1}{n} \sum_{i=1}^n \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2$, where $\Delta_{\Gamma,i}$ is defined in equation (2). The following three assumptions on the model A7, A8, A9 enable to control the maximum Frobenius norm of the difference between covariance matrices from the average Frobenius norm:

$$\max_i \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2 \lesssim \|\Gamma_1 - \Gamma_2\|_F^2 \lesssim d_n^2(\Gamma_1, \Gamma_2) = \frac{1}{n} \sum_{i=1}^n \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2.$$

This point is demonstrated in Appendix B, Lemma B3.

Assumption A7. *The quantity* $\dfrac{1}{n} \sum_{i=1}^n \mathbb{1}_{n_i \geq q}$ *is bounded.*

Assumption A7 means that the number of individuals i such as the number of observations n_i is greater than the number of random effects q is of the order of n, that is n_i is probably greater than q, which seems statistically reasonable to be able to estimate q random effects. This is a necessary assumption for the identifiability of the model.

Assumption A8. *For each* $1 \leq i \leq n$ *such that* $n_i \geq q$, Z_i *is of full rank, i.e.* $\min_i \left\{ \rho_{min}^{1/2}(Z_i^\top Z_i) : n_i \geq q \right\} \gtrsim 1$. *We denote by* $\underline{\rho_Z}$ *the bound:* $\min_i \left\{ \rho_{min}^{1/2}(Z_i^\top Z_i) : n_i \geq q \right\} \geq \underline{\rho_Z}$.

Assumption A9. *For each* $1 \leq i \leq n$, *the maximum of* $\|Z_i\|_{sp}$ *is bounded, i.e.* $\max_i \|Z_i\|_{sp} \lesssim 1$. *We denote by* $\overline{\rho_Z}$ *the bound:* $\max_i \|Z_i\|_{sp} \leq \overline{\rho_Z}$.

Similar assumptions can be found in Theorem 10 of Jeong and Ghosal (2021b) for the linear mixed-effects model.

3 Posterior contraction results

In this section, we provide results on posterior contraction rates in sparse non-linear marginal mixed-effects model under suitable assumptions presented in Section 2.3. To achieve this, we first analyze a dimensionality property of the support of β. Then, we determine how quickly the posterior contracts based on the average Rényi divergence. Finally, we use this information about Rényi contraction to establish the rates for the parameters relative to more practical metrics.

3.1 Support size control

First, it is essential to examine the support size of β in order to then focus on models of relatively small sizes. The following theorem shows that the posterior distribution tends to concentrate on models of relatively small sizes, not much larger than the true one.

Theorem 1 (Effective dimension)**.** *In model* (1)*, with prior specifications outlined in Section 2.2, and assuming the validity of previous assumptions A1-A9, there exists a constant* $C_1 > 0$ *such that the following convergence holds:*

$$\sup_{\beta_0 \in \mathcal{B}_0, \Gamma_0 \in \mathcal{H}_0} \mathbb{E}_0 \left[\Pi \left(\beta : |S_\beta| > C_1 s_0 \middle| Y^{(n)} \right) \right] \underset{n \to \infty}{\longrightarrow} 0.$$

Proof of this theorem is provided in Section 4.1. The derivation of the posterior contraction rate heavily relies on a technical lemma which provides a lower bound for the denominator of the posterior distribution with probability tending to 1, see Lemma 1 in Section 4.1. More precisely, this lemma is employed in deriving our main results on effective dimension and posterior contraction rates, as outlined in Theorems 1 and 2.

3.2 Posterior contraction rates

As discussed in the introduction, the classical approach for determining posterior contraction rates encounters limitations when dealing with the unknown nature of the random effects covariance matrix. Indeed, this approach based on the average squared Hellinger distance faces inadequacies in obtaining rates in terms of the Euclidean norm for the parameters in this context. Specifically, the issue arises from the fact that establishing proximity using the average squared Hellinger distance between multivariate normal densities with individual-specific mean and an unknown covariance does not guarantee average proximity in terms of the Euclidean distance for the mean parameters in these densities. To overcome this challenge, the proposed solution is a direct utilization of the average Rényi divergence of order 1/2 (see Definition 1). This approach is highlighted for its high manageability in the context of multivariate normal distributions and its ability to ensure closeness in terms of the desired Euclidean distance. Examples of the application of this theory can be found in the works of Ning et al. (2020) and Jeong and Ghosal (2021b), further supporting the efficacy of the average Rényi divergence in overcoming the limitations associated with the unknown covariance matrix for random effects.

Definition 1. *For two n-variates densities $f = \prod_{i=1}^n f_i$ and $g = \prod_{i=1}^n g_i$ of independent variables, the average Rényi divergence (of order 1/2) is defined by:*

$$R_n(f,g) = -\frac{1}{n}\sum_{i=1}^n \log\left(\int \sqrt{f_i g_i}\right)$$

Based on the result of Theorem 1, the following theorem establishes the rate of contraction of the posterior distribution towards the truth with respect to the average Rényi divergence.

Theorem 2 (Contraction rate, Rényi)**.** *In model* (1)*, with prior specifications outlined in Section 2.2, we denote by $p_{\beta,\Gamma} = \prod_{i=1}^n p_{\beta,\Gamma,i}$ the joint density, with $p_{\beta,\Gamma,i}$ representing the density of the ith observation vector y_i, and p_0 representing the true joint density. Assuming the previous assumptions A1-A9 hold, as well as $\log(J_n) \lesssim \log(p)$, then there exists a constant $C_2 > 0$ such that:*

$$\sup_{\beta_0\in\mathcal{B}_0,\Gamma_0\in\mathcal{H}_0} \mathbb{E}_0\left[\Pi\left((\beta,\Gamma)\in\mathcal{B}\times\mathcal{H} : R_n(p_{\beta,\Gamma},p_0) > C_2\frac{s_0\log(p)}{n}\middle|Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0.$$

The proof can be found in Section 4.2. This proof is based on the general theory of posterior contraction rate of Ghosal et al. (2000); Ghosal and Van Der Vaart (2007); Ghosal and Van der Vaart (2017), which relies on the construction and existence of exponentially powerful tests (see also Castillo (2024) for more details).

While Theorem 2 provides a fundamental result on posterior contraction, it does not offer precise interpretations for the parameters β and Γ. The following theorem relies on the form of the average Rényi divergence to obtain more concrete contraction rates. Specifically, it demonstrates that the posterior distribution of the prediction term and Γ contracts towards their true respective values at certain rates, relative to metrics more easily understandable than the average Rényi divergence.

Theorem 3 (Recovery)**.** *In model* (1)*, with prior specifications outlined in Section 2.2, and assuming Assumptions A1-A9, as well as $\log(J_n) \lesssim \log(p)$, then there exist constants $C_3, C_4, C_5 > 0$ such that:*

$$\sup_{\beta_0\in\mathcal{B}_0,\Gamma_0\in\mathcal{H}_0} \mathbb{E}_0\left[\Pi\left(\Gamma : d_n(\Gamma,\Gamma_0) > C_3\sqrt{\frac{s_0\log(p)}{n}}\middle|Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0,$$

$$\sup_{\beta_0\in\mathcal{B}_0,\Gamma_0\in\mathcal{H}_0} \mathbb{E}_0\left[\Pi\left(\Gamma : \|\Gamma-\Gamma_0\|_F > C_4\sqrt{\frac{s_0\log(p)}{n}}\middle|Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0,$$

$$\sup_{\beta_0\in\mathcal{B}_0,\Gamma_0\in\mathcal{H}_0} \mathbb{E}_0\left[\Pi\left(\beta : \sqrt{\frac{1}{n}\sum_{i=1}^n \|f_i(X_i\beta)-f_i(X_i\beta_0)\|_2^2} > C_5\sqrt{\frac{s_0\log(p)}{n}}\middle|Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0.$$

The proof can be found in Section 4.3. By comparing our theorem to Castillo et al. (2015)'s results in Bayesian, or Bühlmann and Van De Geer (2011)'s results in frequentist framework, in simple linear regression, it can be observed that the same rates are achieved for the prediction term. For the covariance term, the rate obtained in Theorem 3 coincides with that obtained for linear regression with nuisance parameters by Jeong and Ghosal (2021b).

The last theorem gives precise interpretations of the posterior contraction result for the parameter β. The posterior contraction rates with respect to more concrete metrics are derived based on additional conditions, summarized by Assumptions A10 and A11.

Assumption A10. *For all $i \in \{1, \dots, n\}$, $\delta > 0$ and $t \in \mathbb{R}$,*

$$\Big\{\beta \in \mathbb{R}^p : \big|f(X_i\beta, t) - f(X_i\beta_0, t)\big| \leq \delta\Big\} \subset \Big\{\beta \in \mathbb{R}^p : \big|f(X_i\beta, t) - f(X_i\beta_0, t)\big| \gtrsim \|X_i(\beta - \beta_0)\|_2\Big\}.$$

This assumption, as well as Assumption A7, is a necessary condition for the identifiability of the model and allows to derive posterior contraction rate for β from the third assertion of Theorem 3. In particular, this assumption implies that for all $i \in \{1, \dots, n\}$ and $t \in \mathbb{R}$, the true $X_i\beta_0$ does not lie in a neighborhood of critical points of $f(\cdot, t)$, which seems a reasonable assumption. To ensure the identifiability of the parameter β, a kind of assumption of "local invertibility" for the Gram matrix $X^\top X$ is also required. For this purpose, we define the following compatibility numbers drawing from the literature (Castillo et al., 2015).

Definition 2. *For all $s > 0$, the smallest scaled singular value of dimension s is defined as:*

$$\phi_2(s) = \inf_{\beta : 1 \leq s_\beta \leq s} \frac{\|X\beta\|_2}{\|X\|_* \|\beta\|_2}.$$

Definition 3. *For all $s > 0$, the uniform compatibility number in dimension s is defined as:*

$$\phi_1(s) = \inf_{\beta : 1 \leq s_\beta \leq s} \frac{\|X\beta\|_2 \sqrt{s_\beta}}{\|X\|_* \|\beta\|_1}.$$

Assumption A11. *For each $1 \leq i \leq n$, the maximum of $\|X_i\|_*$ is bounded, i.e. $\max_i \|X_i\|_* \lesssim 1$, and $\beta_0 \in \overline{\mathcal{B}}_0 := \left\{\beta \in \mathcal{B}_0 : \dfrac{s_0^2 \log(p)}{\|X\|_*^2 \phi_1^2((C_1 + 1)s_0)} = o(1)\right\}$.*

Typically, the first assertion in this assumption is commonly satisfied in practical scenarios. The second assertion concerns the true parameter β_0 and will be used for recovery of $X\beta$ in ℓ_2-norm and β with respect to the ℓ_1-norm and the ℓ_2-norm. Since $\phi_2(s) \leq \phi_1(s)$ for all $s > 0$ by the Cauchy-Schwarz inequality, ϕ_1 can be removed if the smallest scaled singular value ϕ_2 is bounded away from zero. Note that under specific conditions on the design matrix, the compatibility numbers can be bounded away from zero (see Example 7 of Castillo et al. (2015) for further discussion). Thus, since by the first assertion of Assumption A11, $\|X\|_*^2 \lesssim n$, the second assertion implies that $s_0^2 \log(p) = o(n)$. This condition is similar to that obtained in Jeong and Ghosal (2021a).

Theorem 4 (Posterior contraction rate for β)**.** *In model* (1)*, with prior specifications outlined in Section 2.2, and assuming Assumptions A1-A11, as well as $\log(J_n) \lesssim \log(p)$, then there exist constants $C_6, C_7, C_8 > 0$ such that:*

$$\sup_{\beta_0 \in \overline{\mathcal{B}}_0, \Gamma_0 \in \mathcal{H}_0} \mathbb{E}_0 \left[\Pi\left(\beta : \|X(\beta - \beta_0)\|_2 > C_6\sqrt{s_0 \log(p)} \middle| Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0,$$

$$\sup_{\beta_0 \in \overline{\mathcal{B}}_0, \Gamma_0 \in \mathcal{H}_0} \mathbb{E}_0 \left[\Pi\left(\beta : \|\beta - \beta_0\|_2 > C_7 \frac{\sqrt{s_0 \log(p)}}{\|X\|_* \phi_2((C_1 + 1)s_0)} \middle| Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0,$$

$$\sup_{\beta_0 \in \overline{\mathcal{B}}_0, \Gamma_0 \in \mathcal{H}_0} \mathbb{E}_0 \left[\Pi\left(\beta : \|\beta - \beta_0\|_1 > C_8 \frac{s_0\sqrt{\log(p)}}{\|X\|_* \phi_1((C_1 + 1)s_0)} \middle| Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0.$$

The proof can be found in Section 4.4. These rates coincide with those obtained by Jeong and Ghosal (2021a) or Jeong and Ghosal (2021b), respectively for generalized linear model and linear regression with nuisance parameters. Since the compatibility numbers can be bounded away from zero under some conditions (see above), they can be removed from the rates.

4 Proofs of main theorems

In this section, proofs of the main theorems are provided. First, additional notations used for the proofs are introduced. Let $\Lambda_n(\beta, \Gamma) = \prod_{i=1}^n p_{\beta,\Gamma,i}/p_{0,i}$ be the likelihood ratio of $p_{\beta,\Gamma} = \prod_{i=1}^n p_{\beta,\Gamma,i}$, where $p_{\beta,\Gamma,i}$ is the density of the i-th observation vector y_i, and $p_0 = \prod_{i=1}^n p_{0,i} = \prod_{i=1}^n p_{\beta_0,\Gamma_0,i}$ the density with the true parameters β_0 and Γ_0. For two densities f and g, let $K(f, g) = \int f(x) \log\left(f(x)/g(x)\right) dx$ the Kullback-Leibler divergence, and $V(f, g) = \int f(x) \left|\log\left(f(x)/g(x)\right) - K(f, g)\right|^2 dx$ the Kullback-Leibler variation.

4.1 Proof of Theorem 1

The proof of Theorem 1 is based on the following technical lemma which provides a lower bound for the denominator of the posterior distribution, with the probability tending to 1.

Lemma 1. *Suppose that Assumptions A1-A9 are satisfied. Then, there exists a positive constant M such that*

$$\mathbb{P}_0\left(\int \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma) \geq \pi_p(s_0)e^{-M(s_0\log(p)+\log(n))}\right) \longrightarrow 1, \tag{5}$$

when n tends to infinity.

This lemma is demonstrated in Appendix A.1.

Proof of Theorem 1. Let $B = \{(\beta,\Gamma) \in \mathcal{B}\times\mathcal{H} : |S_\beta| > \tilde{s}\}$, with an integer $\tilde{s} \geq s_0$. First, by Bayes formula:

$$\Pi(B|Y^{(n)}) = \frac{\int_B \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}{\int \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}. \tag{6}$$

Let us prove that $\mathbb{E}_0\left[\Pi(B|Y^{(n)})\right]$ tends to 0 as n tends to infinity uniformly for $\beta_0 \in \mathcal{B}_0$ and $\Gamma_0 \in \mathcal{H}_0$, and choose a suitable $\tilde{s}$. Let $\mathcal{A}_n$ be the event that appears in Equation (5). We can write

$$\mathbb{E}_0\left[\Pi\left(B|Y^{(n)}\right)\right] = \mathbb{E}_0\left[\Pi\left(B|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}\right] + \mathbb{E}_0\left[\Pi\left(B|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n^c}\right]. \tag{7}$$

where the second term tends to 0 by using Lemma 1.

Concerning the first term, by definition of $\mathcal{A}_n$, we have that

$$\begin{aligned}
\mathbb{E}_0\left[\Pi\left(B|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}\right] &= \mathbb{E}_0\left[\frac{\int_B \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}{\int \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}\mathbb{1}_{\mathcal{A}_n}\right] \\
&\leq \mathbb{E}_0\left[\int_B \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)\pi_p(s_0)^{-1}e^{M(s_0\log(p)+\log(n))}\mathbb{1}_{\mathcal{A}_n}\right] \\
&\leq \pi_p(s_0)^{-1}\exp\{M(s_0\log(p)+\log(n))\}\mathbb{E}_0\left[\int_B \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)\mathbb{1}_{\mathcal{A}_n}\right]
\end{aligned}$$

Now, we get that

$$\begin{aligned}
\mathbb{E}_0\left[\int_B \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)\mathbb{1}_{\mathcal{A}_n}\right] &\leq \mathbb{E}_0\left[\int_B \frac{p_{\beta,\Gamma}(y)}{p_0(y)}d\Pi(\beta,\Gamma)\right] \\
&= \int\int_B p_{\beta,\Gamma}(y)d\Pi(\beta,\Gamma)dy \\
&= \Pi(B)
\end{aligned}$$

using Fubini-Tonelli theorem and since $p_{\beta,\Gamma}$ is a density. Thus,

$$\mathbb{E}_0\left[\Pi\left(B|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}\right] \leq \pi_p(s_0)^{-1}\exp\{M(s_0\log(p)+\log(n))\}\Pi(B),$$

and by Assumption A5,

$$\begin{aligned}
\Pi(B) = \Pi(|S_\beta| > \tilde{s}) &= \sum_{s=\tilde{s}+1}^{p}\pi_p(s) \\
&\leq \pi_p(s_0)\sum_{s=\tilde{s}+1}^{p}\left(A_2p^{-A_4}\right)^{s-s_0} \\
&= \pi_p(s_0)\left(A_2p^{-A_4}\right)^{\tilde{s}+1-s_0}\sum_{k=0}^{p-\tilde{s}-1}\left(A_2p^{-A_4}\right)^k \\
&\leq \pi_p(s_0)\left(A_2p^{-A_4}\right)^{\tilde{s}+1-s_0}\frac{1}{1-A_2p^{-A_4}},
\end{aligned}$$

for p large enough to ensure that $A_2 p^{-A_4} < 1$. Thus finally we have

$$
\begin{aligned}
&\mathbb{E}_0\left[\Pi\left(B|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}\right] \\
&\leq \pi_p(s_0)^{-1}\exp\{M(s_0\log(p)+\log(n))\}\Pi(B) \\
&\leq \exp\left\{M(s_0\log(p)+\log(n))+(\tilde{s}+1-s_0)\log(A_2p^{-A_4})\right\}\frac{1}{1-A_2p^{-A_4}} \\
&= \exp\left\{\log(p)\left(Ms_0+M\frac{\log(n)}{\log(p)}-A_4(\tilde{s}+1-s_0)\right)+(\tilde{s}+1-s_0)\log(A_2)\right\}\frac{1}{1-A_2p^{-A_4}}
\end{aligned}
$$

where $\log(n)/\log(p) \leq 1$ as $p > n$. Thus, as $(1-A_2p^{-A_4})^{-1}$ tends to 1 when $n \to \infty$, we choose $\tilde{s}$ as the largest integer that is smaller than $C_1 s_0$ (such as $\tilde{s}+1 > C_1 s_0$), for some constant C_1 large enough to have $Ms_0 + M - A_4(C_1 s_0 - s_0) < 0$, and then we have that

$$
\begin{aligned}
&\mathbb{E}_0\left[\Pi\left(B|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}\right] \\
&\leq \exp\left\{\log(p)\left(Ms_0+M-A_4(C_1s_0-s_0)\right)+(\tilde{s}+1-s_0)\log(A_2)\right\}\frac{1}{1-A_2p^{-A_4}} \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}
$$

Finally, by Equation (7), we conclude that $\mathbb{E}_0\left[\Pi(B|Y^{(n)})\right] \longrightarrow 0$, for this well-chosen $\tilde{s}$. Thus, we have also that $\mathbb{E}_0\left[\Pi\left(\beta : |S_\beta| > C_1 s_0 \middle| Y^{(n)}\right)\right] \longrightarrow 0$, which concludes the proof of the theorem.

□

4.2 Proof of Theorem 2

Proof of Theorem 2. Let $\mathcal{B}_n = \{\beta \in \mathcal{B} | s_\beta \leq C_1 s_0\}$, $R_n^*(\beta,\Gamma) = R_n(p_{\beta,\Gamma}, p_0)$ and $\epsilon_n = \sqrt{\dfrac{s_0\log(p)}{n}}$.

$$
\begin{aligned}
&\mathbb{E}_0\left[\Pi\left((\beta,\Gamma)\in\mathcal{B}\times\mathcal{H} : R_n^*(\beta,\Gamma) > C_2\epsilon_n^2 \middle| Y^{(n)}\right)\right] \\
&\leq \mathbb{E}_0\left[\Pi\left((\beta,\Gamma)\in\mathcal{B}_n\times\mathcal{H} : R_n^*(\beta,\Gamma) > C_2\epsilon_n^2 \middle| Y^{(n)}\right)\right] + \mathbb{E}_0\left[\Pi\left(\mathcal{B}_n^c|Y^{(n)}\right)\right]
\end{aligned}
$$

where the second term tends to 0 when n goes to infinity by Theorem 1.

Therefore, given $D = \left\{(\beta,\Gamma)\in\mathcal{B}_n\times\mathcal{H} : R_n^*(\beta,\Gamma) > C_2\epsilon_n^2\right\}$, proving Theorem 2 consists in showing that $\mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\right]$ goes to 0 as n tends to infinity uniformly for $\beta_0 \in \mathcal{B}_0$ and $\Gamma_0 \in \mathcal{H}_0$.

This proof is based on the construction and existence of exponentially powerful tests to show contraction rates of posterior distributions (see Ghosal et al. (2000); Ghosal and Van der Vaart (2017) for more details). More precisely, we want to construct a test φ_n such that on an appropriate sieve $\mathcal{B}_n^* \times \mathcal{H}_n \subset \mathcal{B}_n \times \mathcal{H}$ we have, for some constants M_1, $M_2 > 0$:

$$
\mathbb{E}_0[\varphi_n] \lesssim e^{-M_1 n\epsilon_n^2} \quad , \quad \sup_{(\beta,\Gamma)\in\mathcal{B}_n^*\times\mathcal{H}_n : R_n^*(\beta,\Gamma) > C_2\epsilon_n^2} \mathbb{E}_{(\beta,\Gamma)}[1-\varphi_n] \leq e^{-M_2 n\epsilon_n^2} \tag{8}
$$

where the sieve $\mathcal{B}_n^* \times \mathcal{H}_n$ shall satisfy that the prior mass of $\mathcal{B}_n \backslash \mathcal{B}_n^*$ and $\mathcal{H}\backslash\mathcal{H}_n$ decreases rapidly enough to balance the denominator of the posterior. Indeed, assuming that we have constructed such a test, then, for $\mathcal{A}_n$ the event that appears in Equation (5):

$$
\begin{aligned}
\mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\right] &= \mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}\right] + \mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n^c}\right] \\
&= \mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}(1-\varphi_n) + \Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}\varphi_n\right] + \mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n^c}\right] \\
&\leq \mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}(1-\varphi_n)\right] + \mathbb{E}_0\left[\varphi_n\right] + \mathbb{P}_0\left(\mathcal{A}_n^c\right)
\end{aligned}
$$

where by construction of φ_n, $\mathbb{E}_0\left[\varphi_n\right] \underset{n\to\infty}{\longrightarrow} 0$, and $\mathbb{P}_0\left(\mathcal{A}_n^c\right) \underset{n\to\infty}{\longrightarrow} 0$ by Lemma 1.

Now for the first term, by the Bayes formula (6), we have that

$$\begin{aligned}
\mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}(1-\varphi_n)\right] &= \mathbb{E}_0\left[\frac{\int_D \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}{\int \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)}\mathbb{1}_{\mathcal{A}_n}(1-\varphi_n)\right] \\
&\le \mathbb{E}_0\left[\int_D \Lambda_n(\beta,\Gamma)d\Pi(\beta,\Gamma)\pi_p(s_0)^{-1}e^{M(s_0\log(p)+\log(n))}(1-\varphi_n)\right]
\end{aligned}$$

But, grant Assumption A5, we have that: $\pi_p(s_0)^{-1} \le A_1^{-1}p^{A_3}\pi_p(s_0-1)^{-1}$ and by iteration

$$-\log(\pi_p(s_0)) \lesssim s_0\log(p) - \log(\pi_p(0)) \lesssim s_0\log(p)$$

since $1 = \sum_{s=1}^p \pi_p(s) \le \sum_{s=1}^p (A_2p^{-A_4})^s\pi_p(0) \lesssim \pi_p(0)$ by assumption A5. Thus, for a constant C large enough, $\pi_p(s_0)^{-1}e^{M(s_0\log(p)+\log(n))} \le e^{Cs_0\log(p)} = e^{Cn\epsilon_n^2}$, since $\log(n) \lesssim s_0\log(p)$. So, by using the Fubini-Tonelli theorem,

$$\begin{aligned}
\mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}(1-\varphi_n)\right] &\le \int_D \mathbb{E}_{(\beta,\Gamma)}\left[1-\varphi_n\right]d\Pi(\beta,\Gamma)\times e^{Cn\epsilon_n^2} \\
&\le \left(\int_{D\cap(\mathcal{B}_n^*\times\mathcal{H}_n)} \mathbb{E}_{(\beta,\Gamma)}\left[1-\varphi_n\right]d\Pi(\beta,\Gamma) + \Pi(\mathcal{B}_n\backslash\mathcal{B}_n^*) + \Pi(\mathcal{H}\backslash\mathcal{H}_n)\right)\times e^{Cn\epsilon_n^2} \\
&\le \left(\sup_{(\beta,\Gamma)\in D\cap(\mathcal{B}_n^*\times\mathcal{H}_n)}\left\{\mathbb{E}_{(\beta,\Gamma)}\left[1-\varphi_n\right]\right\} + \Pi(\mathcal{B}_n\backslash\mathcal{B}_n^*) + \Pi(\mathcal{H}\backslash\mathcal{H}_n)\right)\times e^{Cn\epsilon_n^2} \\
&\le \left(e^{-M_2n\epsilon_n^2} + \Pi(\mathcal{B}_n\backslash\mathcal{B}_n^*) + \Pi(\mathcal{H}\backslash\mathcal{H}_n)\right)\times e^{Cn\epsilon_n^2}
\end{aligned}$$

by construction of φ_n, equation (8). Then for M_2 large enough and by the condition on the prior mass of $\mathcal{B}_n\backslash\mathcal{B}_n^*$ and $\mathcal{H}\backslash\mathcal{H}_n$, we have that $\mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\mathbb{1}_{\mathcal{A}_n}(1-\varphi_n)\right] \underset{n\to\infty}{\longrightarrow} 0$, and finally $\mathbb{E}_0\left[\Pi\left(D|Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 0$, what was wanted to be demonstrated.

Thus, to complete the proof, we need to demonstrate the existence of such a test φ_n satisfying (8) on an appropriate sieve $\mathcal{B}_n^*\times\mathcal{H}_n$ such that the prior mass of $\mathcal{B}_n\backslash\mathcal{B}_n^*$ and $\mathcal{H}\backslash\mathcal{H}_n$ have an exponential decrease.

Construction of the test φ_n: To this end, we want to apply Lemma D.3 of Ghosal and Van der Vaart (2017), which directly allows to construct the test φ_n with appropriate control of error probabilities as described in (8) to test the true value against the whole of the alternative intersected with the sieve. To apply this lemma, we need to construct local tests with exponentially small errors to compare the true value with a subset of the alternative, centered at any $(\beta_1,\Gamma_1)\in\mathcal{B}\times\mathcal{H}$ which is adequately distant from the true value with respect to the average Rényi divergence. The other condition to apply this Lemma is that the minimum number N_n^* of these small subsets of the alternative needed to cover a sieve $\mathcal{B}_n^*\times\mathcal{H}_n$ is appropriately controlled in terms of ϵ_n.

First, the following lemma constructs an appropriate local test by employing the likelihood ratio to compare the true value with a subset of the alternative and by controlling the second order moment of the likelihood ratios in these small pieces of the alternative. For $(\beta_1,\Gamma_1)\in\mathcal{B}\times\mathcal{H}$, we denote by p_1 the associated density, and $\mathbb{E}_1$ and $\mathbb{P}_1$ the expectation and probability under p_1.

Lemma 2. *For a given positive sequence (γ_n), $(\beta_1,\Gamma_1)\in\mathcal{B}\times\mathcal{H}$ such that $R_n(p_0,p_1)\ge\epsilon_n^2$, where $\epsilon_n = \sqrt{\dfrac{s_0\log(p)}{n}}$, define*

$$\mathcal{F}_{1,n} = \left\{(\beta,\Gamma)\in\mathcal{B}\times\mathcal{H} : \frac{1}{n}\sum_{i=1}^n \|f_i(X_i\beta)-f_i(X_i\beta_1)\|_2^2 \le \frac{\epsilon_n^2}{16\gamma_n},\ d_n(\Gamma,\Gamma_1)\le\frac{\epsilon_n^2}{2J_n\gamma_n},\ \max_{1\le i\le n}\|\Delta_{\Gamma,i}^{-1}\|_{sp}\le\gamma_n\right\}.$$

Grant Assumptions A7-A9 and A4, then there exists a test $\overline{\varphi}_n$ such that

$$\mathbb{E}_0[\overline{\varphi}_n]\le e^{-n\epsilon_n^2},\quad \textit{and} \quad \sup_{(\beta,\Gamma)\in\mathcal{F}_{1,n}}\mathbb{E}_{\beta,\Gamma}[1-\overline{\varphi}_n]\le e^{-n\epsilon_n^2/16}.$$

This lemma is demonstrated in Appendix A.2.

Now, we still have to construct an appropriate sieve $\mathcal{B}_n^* \times \mathcal{H}_n$ such that the prior mass of $\mathcal{B}_n \backslash \mathcal{B}_n^*$ and $\mathcal{H} \backslash \mathcal{H}_n$ have an exponential decrease, and the minimum number N_n^* of the small subsets of the alternative needed to cover the sieve satisfies $\log(N_n^*) \lesssim n\epsilon_n^2$.

Define the sieve as follows:

$$\mathcal{B}_n^* = \left\{\beta \in \mathcal{B} | s_\beta \leq C_1 s_0, \|\beta\|_\infty \leq \frac{p^{L_2+2}}{K_n \|X\|_*}\right\}, \quad \mathcal{H}_n = \left\{\Gamma \in \mathcal{H} | n^{-M} \leq \rho_{min}(\Gamma) \leq \rho_{max}(\Gamma) \leq e^{Mn\epsilon_n^2}\right\}$$

for a constant M, and define $\mathcal{F}_{1,n}$ as in Lemma 2 with $\gamma_n = n^M / \underline{\rho_Z}^2$. Remark that, with this choice of γ_n, the last condition $\max_{1\leq i \leq n} \|\Delta_{\Gamma,i}^{-1}\|_{sp} \leq \gamma_n$ is always satisfy in the sieve. Indeed, $\|\Delta_{\Gamma,i}^{-1}\|_{sp} = \rho_{max}(\Delta_{\Gamma,i}^{-1}) = \rho_{min}^{-1}(\Delta_{\Gamma,i})$. But $\rho_{min}(\Delta_{\Gamma,i}) \geq \sigma^2 + \rho_{min}(Z_i \Gamma Z_i^\top) \geq \rho_{min}(\Gamma)\underline{\rho_Z}^2 \geq n^{-M}\underline{\rho_Z}^2$ by Assumption A8 and since $\Gamma \in \mathcal{H}_n$. So finally, $\max_{1\leq i\leq n} \|\Delta_{\Gamma,i}^{-1}\|_{sp} \leq \gamma_n$ for $\Gamma \in \mathcal{H}_n$.

First, we show that $\mathcal{B}_n \backslash \mathcal{B}_n^*$ and $\mathcal{H} \backslash \mathcal{H}_n$ have an exponential decrease. Using Assumption A5, we obtain that

$$\begin{aligned}
\Pi(\mathcal{B}_n \backslash \mathcal{B}_n^*) &= \Pi\left(\left\{\beta \in \mathcal{B} | s_\beta \leq C_1 s_0, \|\beta\|_\infty > \frac{p^{L_2+2}}{K_n\|X\|_*}\right\}\right) \\
&= \sum_{S: s \leq C_1 s_0} \frac{\pi_p(s)}{\binom{p}{s}} \int_{\left\{\beta_S : \|\beta_S\|_\infty > \frac{p^{L_2+2}}{K_n\|X\|_*}\right\}} g_S(\beta_S) d\beta_S \\
&\leq \sum_{S: s \leq C_1 s_0} \frac{(A_2 p^{-A_4})^s}{\binom{p}{s}} \int_{\left\{\beta_S : \|\beta_S\|_\infty > \frac{p^{L_2+2}}{K_n\|X\|_*}\right\}} g_S(\beta_S) d\beta_S \\
&\leq \sum_{S: s \leq C_1 s_0} \frac{(A_2 p^{-A_4})^s}{\binom{p}{s}} \sum_{\ell \in S} \int_{\left\{|\beta_\ell| > \frac{p^{L_2+2}}{K_n\|X\|_*}\right\}} \frac{\lambda}{2} e^{-\lambda|\beta_\ell|} d\beta_\ell
\end{aligned}$$

Then, by using the tail probability of the Laplace distribution $\int_{|x|>t} \frac{\lambda}{2} e^{-\lambda|x|}dx = e^{-\lambda t}$ for every $t > 0$, and since there is $\binom{p}{s}$ support S of size s, we obtain:

$$\begin{aligned}
\Pi(\mathcal{B}_n \backslash \mathcal{B}_n^*) &\leq \sum_{S: s \leq C_1 s_0} \frac{(A_2 p^{-A_4})^s}{\binom{p}{s}} s e^{-\lambda \frac{p^{L_2+2}}{K_n\|X\|_*}} \\
&\leq \sum_{s=1}^{C_1 s_0} s (A_2 p^{-A_4})^s e^{-\lambda \frac{p^{L_2+2}}{K_n\|X\|_*}} \\
&\leq C_1 s_0 e^{-\lambda \frac{p^{L_2+2}}{K_n\|X\|_*}} \sum_{s=1}^{C_1 s_0} (A_2 p^{-A_4})^s \\
&\lesssim s_0 e^{-\lambda \frac{p^{L_2+2}}{K_n\|X\|_*}} \lesssim s_0 e^{-\frac{\lambda}{L_1} p^2}
\end{aligned}$$

Thus, $\Pi(\mathcal{B}_n \backslash \mathcal{B}_n^*) e^{Cn\epsilon_n^2} \underset{n\to\infty}{\longrightarrow} 0$ for every $C > 0$ since $n\epsilon_n^2 = s_0 \log(p) = o(p^2)$.

Now,

$$
\begin{aligned}
\Pi(\mathcal{H}\backslash\mathcal{H}_n) &= \Pi\left(\left\{\Gamma \in \mathcal{H} | \rho_{min}(\Gamma) < n^{-M} \text{ or } \rho_{max}(\Gamma) > e^{Mn\epsilon_n^2}\right\}\right) \\
&\leq \Pi\left(\left\{\Gamma \in \mathcal{H} | \rho_{min}(\Gamma) < n^{-M}\right\}\right) + \Pi\left(\left\{\Gamma \in \mathcal{H} | \rho_{max}(\Gamma) > e^{Mn\epsilon_n^2}\right\}\right) \\
&= \Pi\left(\left\{\Gamma \in \mathcal{H} | \rho_{max}(\Gamma^{-1}) \geq n^{M}\right\}\right) + \Pi\left(\left\{\Gamma \in \mathcal{H} | \rho_{min}(\Gamma^{-1}) \leq e^{-Mn\epsilon_n^2}\right\}\right) \\
&\leq b_1 e^{-b_2 n^{b_3 M}} \times b_4 e^{-b_5 Mn\epsilon_n^2}
\end{aligned}
$$

for some constants $b_1, b_2, b_3, b_4, b_5 > 0$ by Lemma 9.16 of Ghosal and Van der Vaart (2017) since $\Gamma^{-1} \sim \mathcal{W}_q(d, \Sigma^{-1})$. So, $\Pi(\mathcal{H}\backslash\mathcal{H}_n)e^{Cn\epsilon_n^2} \underset{n\to\infty}{\longrightarrow} 0$ for every $C > 0$, for M large enough.

Finally, we have to prove that the minimum number N_n^* of the small subsets of the alternative of the form $\mathcal{F}_{1,n}$ needed to cover the sieve satisfies $\log(N_n^*) \lesssim n\epsilon_n^2$. First, note that for every $\beta, \beta' \in \mathcal{B}$, by Assumption A1, the inequality $\|X\theta\|_2 \leq \|X\|_* \|\theta\|_1$, we have:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \|f_i(X_i\beta) - f_i(X_i\beta')\|_2^2 &\leq \frac{1}{n}\sum_{i=1}^{n} K_n^2 \|X_i(\beta - \beta')\|_2^2 = \frac{K_n^2}{n}\|X(\beta - \beta')\|_2^2 \\
&\leq \frac{K_n^2}{n}\|X\|_*^2 \|\beta - \beta'\|_1^2 \\
&\leq \frac{p^2 K_n^2}{n}\|X\|_*^2 \|\beta - \beta'\|_\infty^2
\end{aligned}
$$

Thus, we define

$$
\mathcal{F}'_{1,n} = \left\{(\beta, \Gamma) \in \mathcal{B} \times \mathcal{H} : \frac{p^2 K_n^2}{n}\|X\|_*^2 \|\beta - \beta'\|_\infty^2 + d_n^2(\Gamma, \Gamma_1) \leq \frac{1}{16 J_n^2 \gamma_n^2 n^3}, \max_{1\leq i\leq n} \|\Delta_{\Gamma,i}^{-1}\|_{sp} \leq \gamma_n\right\},
$$

with the same (β_1, Γ_1) used in $\mathcal{F}_{1,n}$, and therefore we have $\mathcal{F}'_{1,n} \subset \mathcal{F}_{1,n}$. Thus, since N_n^* is the minimum number of the small subsets of the alternative of the form $\mathcal{F}_{1,n}$ needed to cover the sieve $\mathcal{B}_n^* \times \mathcal{H}_n$, with $\mathcal{F}_{1,n} \supset \mathcal{F}'_{1,n}$, we have that N_n^* is bounded above by the minimum number of the small subsets of the alternative of the form $\mathcal{F}'_{1,n}$ needed to cover the sieve $\mathcal{B}_n^* \times \mathcal{H}_n$. This last minimum number is denoted by N'_n. In the following, for a pseudo-metric space $(\mathcal{F}, d)$, let $N(\epsilon, \mathcal{F}, d)$ denote the minimal number of ϵ-balls that cover $\mathcal{F}$.

Now, note that if $(\beta, \Gamma) \in \mathcal{B}_n^* \times \mathcal{H}_n$ with $\|\beta - \beta_1\|_\infty \leq \frac{1}{6npK_n J_n \gamma_n \|X\|_*}$ and $d_n(\Gamma, \Gamma_1) \leq \frac{1}{6n^{3/2} J_n \gamma_n}$, then $(\beta, \Gamma) \in \mathcal{F}'_{1,n}$ (by using that the last condition is satisfy for all $\Gamma \in \mathcal{H}_n$). Thus,

$$
N'_n \leq N\left(\frac{1}{6npK_n J_n \gamma_n \|X\|_*}, \mathcal{B}_n^*, \|\cdot\|_\infty\right) \times N\left(\frac{1}{6n^{3/2} J_n \gamma_n}, \mathcal{H}_n, d_n\right).
$$

Then,

$$
\log(N_n^*) \leq \log N\left(\frac{1}{6npK_n J_n \gamma_n \|X\|_*}, \mathcal{B}_n^*, \|\cdot\|_\infty\right) + \log N\left(\frac{1}{6n^{3/2} J_n \gamma_n}, \mathcal{H}_n, d_n\right). \tag{9}
$$

Recall that $\mathcal{B}_n^* = \left\{\beta \in \mathcal{B} | s_\beta \leq C_1 s_0, \|\beta\|_\infty \leq \frac{p^{L_2+2}}{K_n \|X\|_*}\right\}$, to cover $\mathcal{B}_n^*$, we have to choose at most $\lfloor C_1 s_0 \rfloor$ non-zero β coordinates and we need to recover a ball in $\mathbb{R}^{\lfloor C_1 s_0 \rfloor}$ with radius $\frac{p^{L_2+2}}{K_n \|X\|_*}$, with balls of radius $\frac{1}{6npK_n J_n \gamma_n \|X\|_*}$. Therefore,

$$
\begin{aligned}
N\left(\frac{1}{6npK_n J_n \gamma_n \|X\|_*}, \mathcal{B}_n^*, \|\cdot\|_\infty\right) &\leq \binom{p}{\lfloor C_1 s_0 \rfloor}\left(6p^{L_2+3} n J_n \gamma_n\right)^{\lfloor C_1 s_0 \rfloor} \\
&\lesssim \left(6p^{L_2+4} n J_n \gamma_n\right)^{\lfloor C_1 s_0 \rfloor} \text{ as } \binom{p}{\lfloor C_1 s_0 \rfloor}(\lfloor C_1 s_0 \rfloor)! \leq p^{\lfloor C_1 s_0 \rfloor}
\end{aligned}
$$

So, the first term in the right side of equation (9) is bounded by:

$$\log N\left(\frac{1}{6npK_nJ_n\gamma_n\|X\|_*}, \mathcal{B}_n^*, \|\cdot\|_\infty\right) \lesssim s_0\log(p) = n\epsilon_n^2$$

since, by the assumption of Theorem 2, we have $\log(J_n) \lesssim \log(p)$ and $\log(n) \lesssim \log(p)$.

Similarly, for the second term in the right side of equation (9), note that $\mathcal{H}_n \subset \left\{\Gamma \in \mathcal{H} : \|\Gamma\|_F \leq \sqrt{q}e^{Mn\epsilon_n^2}\right\}$, and therefore by assumption A9:

$$\begin{aligned}
\log N\left(\frac{1}{6n^{3/2}J_n\gamma_n}, \mathcal{H}_n, d_n\right) &\leq \log N\left(\frac{1}{6n^{3/2}J_n\gamma_n}, \left\{\Gamma \in \mathcal{H} : \|\Gamma\|_F \leq \sqrt{q}e^{Mn\epsilon_n^2}\right\}, d_n\right)\\
&\leq \log N\left(\frac{1}{6n^{3/2}J_n\gamma_n\overline{\rho_Z}^2}, \left\{\Gamma \in \mathcal{H} : \|\Gamma\|_F \leq \sqrt{q}e^{Mn\epsilon_n^2}\right\}, \|\cdot\|_F\right)\\
&\lesssim q^2\log(\sqrt{q}e^{Mn\epsilon_n^2}n^{3/2}J_n\gamma_n) \lesssim n\epsilon_n^2
\end{aligned}$$

as $\log(J_n) \lesssim \log(p)$ and $\log(n) \lesssim \log(p)$.

Finally, $\log(N_n^*) \lesssim n\epsilon_n^2$. Thus, Lemma D.3 of Ghosal and Van der Vaart (2017) can be applied and gives that for every $\epsilon > \epsilon_n$, there exists a test φ_n satisfying

$$\mathbb{E}_0[\varphi_n] \leq 2e^{B_1n\epsilon_n^2 - n\epsilon^2} \text{ and } \sup_{(\beta,\Gamma)\in\mathcal{B}_n^*\times\mathcal{H}_n : R_n^*(\beta,\Gamma)>\epsilon^2} \mathbb{E}_{(\beta,\Gamma)}[1-\varphi_n] \leq e^{-n\epsilon^2/16}$$

for some constant $B_1 > 0$. Then, choosing $\epsilon = C_2\epsilon_n$ for C_2 large enough, we obtain that the test φ_n satisfies (8), which concludes the proof, as demonstrated above.

□

4.3 Proof of Theorem 3

Proof of Theorem 3. The contraction rate of the posterior distribution with respect to the average Rényi divergence $R_n^*(\beta,\Gamma) = R_n(p_{\beta,\Gamma}, p_0)$ is provided by Theorem 2. Denote $\epsilon_n = \sqrt{\frac{s_0\log(p)}{n}}$ this rate. We have that, for all $\beta_0 \in \mathcal{B}_0, \Gamma_0 \in \mathcal{H}_0$,

$$\mathbb{E}_0\left[\Pi\left(\mathcal{R}|Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 1.$$

where $\mathcal{R} = \left\{(\beta,\Gamma) \in \mathcal{B}\times\mathcal{H} : R_n^*(\beta,\Gamma) \leq C_2\epsilon_n^2\right\}$. However, since for all $(\beta,\Gamma) \in \mathcal{B}\times\mathcal{H}$, $p_{\beta,\Gamma} = \prod_{i=1}^n p_{\beta,\Gamma,i}$, with $p_{\beta,\Gamma,i} = \mathcal{N}_{n_i}(f_i(X_i\beta), \Delta_{\Gamma,i})$ in Model (1), the average Rényi divergence is equal to:

$$\begin{aligned}
R_n^*(\beta,\Gamma) = R_n(p_{\beta,\Gamma}, p_0) &= -\frac{1}{n}\sum_{i=1}^n \log\left(\int\sqrt{p_{\beta,\Gamma,i}(y_i)p_{0,i}(y_i)}dy_i\right)\\
&= -\frac{1}{n}\sum_{i=1}^n \log\left(1 - g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i})\right) + \frac{1}{4n}\sum_{i=1}^n \|(\Delta_{\Gamma,i}+\Delta_{\Gamma_0,i})^{-1/2}(f_i(X_i\beta) - f_i(X_i\beta_0))\|_2^2
\end{aligned}$$

where we used the Sherman-Morrison-Woodbury formula, with

$$g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i}) = 1 - \frac{\det(\Delta_{\Gamma,i})^{1/4}\det(\Delta_{\Gamma_0,i})^{1/4}}{\det((\Delta_{\Gamma,i}+\Delta_{\Gamma_0,i})/2)^{1/2}}.$$

Remark that for all $1 \leq i \leq n$, $g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i}) \geq 0$ since, with $\Delta_{\Gamma,i}^* = \Delta_{\Gamma_0,i}^{-1/2}\Delta_{\Gamma,i}\Delta_{\Gamma_0,i}^{-1/2}$:

$$\begin{aligned}
\frac{\det(\Delta_{\Gamma,i})^{1/4}\det(\Delta_{\Gamma_0,i})^{1/4}}{\det((\Delta_{\Gamma,i}+\Delta_{\Gamma_0,i})/2)^{1/2}} &= \left(\frac{1}{2^{n_i}}\det({\Delta_{\Gamma,i}^*}^{1/2} + {\Delta_{\Gamma,i}^*}^{-1/2})\right)^{-1/2}\\
&= \left(\prod_{k=1}^{n_i}\frac{1}{2}(d_k^{1/2} + d_k^{-1/2})\right)^{-1/2} \leq 1
\end{aligned}$$

where d_k are the eigenvalues of $\Delta^*_{\Gamma,i}$, and using that $\forall x \leq 0, x + x^{-1} \geq 2$.

Thus, by Theorem 2, this implies that, for $(\beta, \Gamma) \in \mathcal{R}$:

$$\begin{aligned}
\epsilon_n^2 &\gtrsim -\frac{1}{n}\sum_{i=1}^n \log\left(1 - g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i})\right) + \frac{1}{4n}\sum_{i=1}^n \|(\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i})^{-1/2}(f_i(X_i\beta) - f_i(X_i\beta_0))\|_2^2 \\
&\gtrsim -\frac{1}{n}\sum_{i=1}^n \log\left(1 - g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i})\right) \geq \frac{1}{n}\sum_{i=1}^n g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i})
\end{aligned}$$

since $\log(1-x) \leq -x$ for all $x \geq 0$. Now, by Lemma 10 of Jeong and Ghosal (2021b), we obtain that $g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i}) \gtrsim \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2$ if $g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i})$ is small enough for each $i \in \{1, \dots, n\}$. Thus, by defining $I_{n,\delta} = \{1 \leq i \leq n : g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i}) \geq \delta\}$, for some $\delta > 0$ small, we have:

$$\begin{aligned}
\epsilon_n^2 &\gtrsim \frac{1}{n}\sum_{i \notin I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 + \frac{1}{n}\sum_{i \in I_{n,\delta}} g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i}) \\
&\gtrsim \frac{1}{n}\sum_{i \notin I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 = \frac{1}{n}\sum_{i=1}^n \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 - \frac{1}{n}\sum_{i \in I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 \\
&\geq M_1 d_n^2(\Gamma, \Gamma_0) - M_1 \frac{|I_{n,\delta}|}{n} \max_{1 \leq i \leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 \\
&\geq M_1 d_n^2(\Gamma, \Gamma_0) - M_2 \epsilon_n^2 \max_{1 \leq i \leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 \\
&\geq (M_1 - M_3\epsilon_n^2) d_n^2(\Gamma, \Gamma_0)
\end{aligned}$$

where we used that $\frac{|I_{n,\delta}|}{n} \lesssim \epsilon_n^2$ since

$$\epsilon_n^2 \;\gtrsim\; \frac{1}{n}\sum_{i \notin I_{n,\delta}} g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i}) \;\gtrsim\; \frac{|I_{n,\delta}|}{n} \times \delta,$$

and thanks to Lemma B3 for the last inequality. Then, since $\epsilon_n^2 \underset{n\to\infty}{\longrightarrow} 0$, $M_1 - M_3\epsilon_n^2$ is bounded away from 0, the last inequation implies that $\epsilon_n \gtrsim d_n(\Gamma, \Gamma_0)$, which proves the first assertion of Theorem 3. Now, thanks to Lemma B3, we have also that $\epsilon_n \gtrsim \|\Gamma - \Gamma_0\|_F$, which proves the second assertion of Theorem 3.

Also, by Theorem 2, for $(\beta, \Gamma) \in \mathcal{R}$, since $g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i}) \geq 0$, we have that:

$$\begin{aligned}
\epsilon_n^2 &\gtrsim -\frac{1}{n}\sum_{i=1}^n \log\left(1 - g^2(\Delta_{\Gamma,i}, \Delta_{\Gamma_0,i})\right) + \frac{1}{4n}\sum_{i=1}^n \|(\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i})^{-1/2}(f_i(X_i\beta) - f_i(X_i\beta_0))\|_2^2 \\
&\geq \frac{M_4}{4n}\sum_{i=1}^n \|(\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i})^{-1/2}(f_i(X_i\beta) - f_i(X_i\beta_0))\|_2^2 \\
&\geq \frac{M_4}{4n}\sum_{i=1}^n \rho_{min}((\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i})^{-1})\|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2
\end{aligned}$$

using that for A symmetric matrix, $\|Ax\|_2^2 \geq \rho_{min}(A^2)\|x\|_2^2$ for all vector x.

Now, since $\rho_{min}((\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i})^{-1}) = \rho_{max}^{-1}(\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i})$, we want to upper bound $\rho_{max}(\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i})$ uniformly across $i \in \{1, \dots, n\}$. Thus, by using Weyl's inequality:

$$\rho_{max}(\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i}) \leq \rho_{max}(\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}) + 2\rho_{max}(\Delta_{\Gamma_0,i}) \leq \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F + 2\overline{\rho_\Delta}$$

by Lemma B2 where $\overline{\rho_\Delta}$ denotes the uniform upper bound of the eigenvalues of $(\Delta_{\Gamma_0,i})_i$. Thus, by Lemma B3,

$$\max_{1 \leq i \leq n} \rho_{max}(\Delta_{\Gamma,i} + \Delta_{\Gamma_0,i}) \leq \max_{1 \leq i \leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F + 2\overline{\rho_\Delta} \leq d_n(\Gamma, \Gamma_0) + 2\overline{\rho_\Delta} \leq C_3\epsilon_n + 2\overline{\rho_\Delta},$$

by the first assertion of Theorem 3.

Finally, we obtain that:

$$\epsilon_n^2 \geq \frac{M_4}{4n(C_3\epsilon_n + 2\overline{\rho_\Delta})} \sum_{i=1}^{n} \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2,$$

and since $C_3\epsilon_n + 2\overline{\rho_\Delta} \underset{n\to\infty}{\longrightarrow} 2\overline{\rho_\Delta}$, we finally obtain that for n large enough $\epsilon_n \gtrsim \sqrt{\frac{1}{n}\sum_{i=1}^n \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2}$, which gives the last assertion of Theorem 3.

□

4.4 Proof of Theorem 4

Proof of Theorem 4. Let us consider $\beta_0 \in \overline{\mathcal{B}}_0 \subset \mathcal{B}_0$ and $\Gamma_0 \in \mathcal{H}_0$. The contraction rate of the posterior distribution for the prediction term is provided by Theorem 3 and we have in particular that: for all $\beta_0 \in \overline{\mathcal{B}}_0, \Gamma_0 \in \mathcal{H}_0$,

$$\mathbb{E}_0\left[\Pi\left(\mathcal{P}_n \middle| Y^{(n)}\right)\right] \underset{n\to\infty}{\longrightarrow} 1,$$

where $\mathcal{P}_n = \left\{\beta : \frac{1}{n}\sum_{i=1}^n \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2 \lesssim \epsilon_n^2\right\}$ and $\epsilon_n = \sqrt{\frac{s_0\log(p)}{n}}$.

Now, note that for $i \in \{1,\ldots,n\}$, and $\delta > 0$, if we assume that $\|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2 \leq \delta$, then for all $j \in \{1,\ldots,n_i\}$, $|f(X_i\beta, t_{ij}) - f(X_i\beta_0, t_{ij})| \leq \delta$ and so by Assumption A10

$$|f(X_i\beta, t_{ij}) - f(X_i\beta_0, t_{ij})| \gtrsim \|X_i(\beta-\beta_0)\|_2.$$

Therefore, $\|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2 \gtrsim n_i\|X_i(\beta-\beta_0)\|_2^2 \gtrsim \|X_i(\beta-\beta_0)\|_2^2$, as $n_i \geq 1$. Thus, using the same idea used in the proof of Theorem 3, we define $I_{n,\delta} = \{1 \leq i \leq n : \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2 > \delta\}$, for some $\delta > 0$ small. Thus, for $\beta \in \mathcal{P}_n$,

$$\begin{aligned}
\epsilon_n^2 &\gtrsim \frac{1}{n}\sum_{i=1}^n \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2 = \frac{1}{n}\sum_{i\in I_{n,\delta}} \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2 + \frac{1}{n}\sum_{i\notin I_{n,\delta}} \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2 \\
&\gtrsim \frac{1}{n}\sum_{i\notin I_{n,\delta}} \|X_i(\beta-\beta_0)\|_2^2 = \frac{1}{n}\sum_{i=1}^n \|X_i(\beta-\beta_0)\|_2^2 - \frac{1}{n}\sum_{i\in I_{n,\delta}} \|X_i(\beta-\beta_0)\|_2^2 \\
&\gtrsim \frac{1}{n}\|X(\beta-\beta_0)\|_2^2 - \frac{|I_{n,\delta}|}{n}\max_{1\leq i\leq n} \|X_i(\beta-\beta_0)\|_2^2
\end{aligned}$$

Then, by the first assertion of Assumption A11, we have that

$$\max_{1\leq i\leq n} \|X_i(\beta-\beta_0)\|_2^2 \leq \max_{1\leq i\leq n} \|X_i\|_*^2\|\beta-\beta_0\|_1^2 \lesssim \|\beta-\beta_0\|_1^2.$$

Remark that for β such as $s_\beta \leq C_1 s_0$, we have $s_{\beta-\beta_0} \leq s_\beta + s_0 \leq (C_1+1)s_0$, thus by Theorem 1 and by Definition 3, we obtain that: $\|\beta-\beta_0\|_1^2 \leq \dfrac{s_0\|X(\beta-\beta_0)\|_2^2}{\|X\|_*^2\phi_1^2((C_1+1)s_0)}$, so

$$\max_{1\leq i\leq n} \|X_i(\beta-\beta_0)\|_2^2 \lesssim \frac{s_0\|X(\beta-\beta_0)\|_2^2}{\|X\|_*^2\phi_1^2((C_1+1)s_0)}.$$

Then, by using $|I_{n,\delta}| \lesssim n\epsilon_n^2$, since $\epsilon_n^2 \gtrsim \frac{1}{n}\sum_{i\in I_{n,\delta}} \|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2 \gtrsim \frac{|I_{n,\delta}|}{n} \times \delta$, we have that:

$$\begin{aligned}
\epsilon_n^2 &\gtrsim \left(1 - \frac{s_0|I_{n,\delta}|}{\|X\|_*^2\phi_1^2((C_1+1)s_0)}\right)\frac{1}{n}\|X(\beta-\beta_0)\|_2^2 \\
&\gtrsim \left(1 - \frac{s_0 n\epsilon_n^2}{\|X\|_*^2\phi_1^2((C_1+1)s_0)}\right)\frac{1}{n}\|X(\beta-\beta_0)\|_2^2.
\end{aligned}$$

Moreover, since $\beta_0 \in \overline{\mathcal{B}}_0$, $1 - \dfrac{s_0 n \epsilon_n^2}{\|X\|_*^2 \phi_1^2((C_1+1)s_0)}$ is bounded away from 0. This implies that $\|X(\beta - \beta_0)\|_2^2 \lesssim n\epsilon_n^2$ which gives the first assertion of Theorem 4.

Finally, by definition of the uniform compatibility number ϕ_1 and the smallest scaled singular value ϕ_2, we obtain that:

$$\epsilon_n^2 \gtrsim \frac{\|X\|_*^2 \phi_1^2((C_1+1)s_0)}{s_0 n} \|\beta - \beta_0\|_1^2,$$

$$\text{and } \epsilon_n^2 \gtrsim \frac{\|X\|_*^2 \phi_2^2((C_1+1)s_0)}{n} \|\beta - \beta_0\|_2^2,$$

which proves the last two assertions of the theorem. □

Acknowledgements

This work was funded by the Stat4Plant project ANR-20-CE45-0012. The authors thank Professor Ismaël Castillo (Sorbonne Université) for his helpful comments and recommendations about this work.

References

Bühlmann, P. and Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

Castillo, I. (2024). Bayesian nonparametric statistics, St-Flour lecture notes. arXiv preprint arXiv:2402.16422.

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. Publisher: Institute of Mathematical Statistics.

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.

Demidenko, E. (2013). Mixed models: theory and applications with R. John Wiley & Sons.

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

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

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. Annals of Statistics, pages 500–531. Publisher: JSTOR.

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

Ghosal, S. and Van der Vaart, A. (2017). Fundamentals of nonparametric Bayesian inference, volume 44. Cambridge University Press.

Jeong, S. and Ghosal, S. (2021a). Posterior contraction in sparse generalized linear models. Biometrika, 108(2):367–379. Publisher: Oxford University Press.

Jeong, S. and Ghosal, S. (2021b). Unified Bayesian theory of sparse linear regression with nuisance parameters. Electronic Journal of Statistics, 15(1):3040–3111. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

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

Lavielle, M. (2014). Mixed effects models for the population approach: models, tasks, methods and tools. CRC press, New York.

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

Ning, B., Jeong, S., and Ghosal, S. (2020). Bayesian linear regression for multivariate responses under group sparsity. Bernoulli, 26(3):2353–2382. Publisher: International Statistical Institute.

Pinheiro, J. C. and Bates, D. M. (2000). Mixed-effects Models in S and S-PLUS. Springer, New York.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. Journal of the American Statistical Association, 113(521):431–444. Publisher: Taylor & Francis.

Shen, Y. and Deshpande, S. K. (2022). On the posterior contraction of the multivariate spike-and-slab LASSO. arXiv preprint arXiv:2209.04389.

Song, Q. and Liang, F. (2023). Nearly optimal Bayesian shrinkage for high dimensional regression. Science China Mathematics, 66:409–442.

Tadesse, M. G. and Vannucci, M. (2021). Handbook of Bayesian variable selection. CRC Press, Boca Raton.

Appendices

Appendix A Proofs of technical lemmas

A.1 Proof of Lemma 1

First, we define a Kullback-Leibler neighbourhood around $p_{0,i}$

$$\mathcal{D}_n = \left\{ (\beta, \Gamma) \in \mathcal{B} \times \mathcal{H} \,\middle|\, \sum_{i=1}^n K(p_{0,i}, p_{\beta,\Gamma,i}) \leq c_1 \log(n),\ \sum_{i=1}^n V(p_{0,i}, p_{\beta,\Gamma,i}) \leq c_1 \log(n) \right\}$$

for a constant c_1 large enough. Then, by Lemma 10 of Ghosal and Van Der Vaart (2007), we have that, for every $C > 0$,

$$\mathbb{P}_0 \left(\int_{\mathcal{D}_n} \Lambda_n(\beta, \Gamma) d\Pi(\beta, \Gamma) \leq e^{-(1+C)c_1 \log(n)} \Pi(\mathcal{D}_n) \right) \leq \frac{1}{C^2 c_1 \log(n)}. \tag{10}$$

The proof of the lemma consists in showing that there exists $M > 0$ such that $e^{-(1+C)c_1 \log(n)} \Pi(\mathcal{D}_n) \gtrsim \pi_p(s_0) e^{-M(s_0 \log(p) + \log(n))}$. Indeed by combining this result with Inequality (10), since $\int_{\mathcal{D}_n} \Lambda_n(\beta, \Gamma) d\Pi(\beta, \Gamma) \leq \int \Lambda_n(\beta, \Gamma) d\Pi(\beta, \Gamma)$, we have that

$$\begin{aligned}
&\mathbb{P}_0 \left(\int \Lambda_n(\beta, \Gamma) d\Pi(\beta, \Gamma) \geq \pi_p(s_0) e^{-M(s_0 \log(p) + \log(n))} \right) \\
&\geq \mathbb{P}_0 \left(\int_{\mathcal{D}_n} \Lambda_n(\beta, \Gamma) d\Pi(\beta, \Gamma) \geq \pi_p(s_0) e^{-M(s_0 \log(p) + \log(n))} \right) \\
&\geq \mathbb{P}_0 \left(\int_{\mathcal{D}_n} \Lambda_n(\beta, \Gamma) d\Pi(\beta, \Gamma) \leq e^{-(1+C)c_1 \log(n)} \Pi(\mathcal{D}_n) \right) \\
&\geq 1 - \frac{1}{C^2 c_1 \log(n)} \underset{n \to \infty}{\longrightarrow} 1,
\end{aligned}$$

that concludes the proof. Thus, it remains to show that $e^{-(1+C)c_1 \log(n)} \Pi(\mathcal{D}_n) \geq \pi_p(s_0) e^{-M(s_0 \log(p) + \log(n))}$, or, more precisely, we need to exhibit a lower bound of $\Pi(\mathcal{D}_n)$.

In Model (1), we have that $p_{\beta,\Gamma,i} = \mathcal{N}(f_i(X_i\beta), \Delta_{\Gamma,i})$, with $\Delta_{\Gamma,i} = Z_i \Gamma Z_i^\top + \sigma^2 I_{n_i}$. By Lemma 9 of Jeong and Ghosal (2021b), the Kullback-Leibler divergence and variation of the i-th individual are respectively expressed as:

$$\begin{aligned}
K(p_{0,i}, p_{\beta,\Gamma,i}) &= \frac{1}{2} \left[\log\left(\frac{|\Delta_{\Gamma,i}|}{|\Delta_{\Gamma_0,i}|} \right) + \mathrm{Tr}(\Delta_{\Gamma_0,i} \Delta_{\Gamma,i}^{-1}) - n_i + \left\| \Delta_{\Gamma,i}^{-1/2} \left(f_i(X_i\beta) - f_i(X_i\beta_0) \right) \right\|_2^2 \right], \\
V(p_{0,i}, p_{\beta,\Gamma,i}) &= \frac{1}{2} \left[\mathrm{Tr}\left(\Delta_{\Gamma_0,i} \Delta_{\Gamma,i}^{-1} \Delta_{\Gamma_0,i} \Delta_{\Gamma,i}^{-1} \right) - 2\mathrm{Tr}(\Delta_{\Gamma_0,i} \Delta_{\Gamma,i}^{-1}) + n_i \right] + \left\| \Delta_{\Gamma_0,i}^{1/2} \Delta_{\Gamma,i}^{-1} \left(f_i(X_i\beta) - f_i(X_i\beta_0) \right) \right\|_2^2.
\end{aligned}$$

Then, by denoting $\rho_{i,k}$, for $k = 1, \ldots, n_i$, the eigenvalues of $\Delta_{\Gamma_0,i}^{1/2} \Delta_{\Gamma,i}^{-1} \Delta_{\Gamma_0,i}^{1/2}$, we obtain that

$$\begin{aligned}
K(p_{0,i}, p_{\beta,\Gamma,i}) &= \frac{1}{2} \left[-\sum_{k=1}^{n_i} \log(\rho_{i,k}) - \sum_{k=1}^{n_i} (1 - \rho_{i,k}) + \| \Delta_{\Gamma,i}^{-1/2} \left(f_i(X_i\beta) - f_i(X_i\beta_0) \right) \|_2^2 \right], \\
V(p_{0,i}, p_{\beta,\Gamma,i}) &= \frac{1}{2} \sum_{k=1}^{n_i} (1 - \rho_{i,k})^2 + \| \Delta_{\Gamma_0,i}^{1/2} \Delta_{\Gamma,i}^{-1} \left(f_i(X_i\beta) - f_i(X_i\beta_0) \right) \|_2^2.
\end{aligned}$$

Our goal is to find a lower bound of $\Pi(\mathcal{D}_n)$, so we want to find an upper bound of $\sum_{i=1}^n K(p_{0,i}, p_{\beta,\Gamma,i})$ and $\sum_{i=1}^n V(p_{0,i}, p_{\beta,\Gamma,i})$.

Let us first focus on the term $V(p_{0,i}, p_{\beta,\Gamma,i})$. By Lemma 10 of Jeong and Ghosal (2021b), we obtain that:

$$\sum_{k=1}^{n_i}(1-\rho_{i,k}^{-1})^2 \leq \rho_{min}^{-2}(\Delta_{\Gamma_0,i})\|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2.$$

By using Weyl's inequality and Assumptions A9 and A3, Lemma B2 shows that there exist $\underline{\rho_0} > 0$ and $\overline{\rho_0} > 0$ such that:

$$\underline{\rho_0} \leq \min_i \rho_{min}(\Delta_{\Gamma_0,i}) \leq \max_i \rho_{max}(\Delta_{\Gamma_0,i}) \leq \overline{\rho_0}. \tag{11}$$

Thus

$$\max_i \sum_{k=1}^{n_i}(1-\rho_{i,k}^{-1})^2 \leq \underline{\rho_0}^{-2}\max_i \|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2,$$

and in particular, $\max_{i,k}(1-\rho_{i,k}^{-1})^2 \leq \underline{\rho_0}^{-2}\max_i \|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2$. Thus, if $\max_i \|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2 \to 0$ on $\mathcal{D}_n$, we will have that $\max_{i,k}|1-\rho_{i,k}^{-1}| \to 0$, that is each $\rho_{i,k}$ tends to 1 and so:

$$\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 \lesssim \sum_{k=1}^{n_i}(1-\rho_{i,k}^{-1})^2 \leq \underline{\rho_0}^{-2}\|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2 \lesssim \|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2, \tag{12}$$

where the first inequality is due to $|1-x^{-1}| \lesssim |1-x| \lesssim |1-x^{-1}|$ for $x \to 1$, which enables to bound the first term of $V(p_{0,i}, p_{\beta,\Gamma,i})$.

Now, prove that $\max_i \|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2$ tends to 0 on $\mathcal{D}_n$. We introduce the set $I_{n,\delta} = \{1 \leq i \leq n | \sum_{k=1}^{n_i}(1-\rho_{i,k})^2 \geq \delta\}$, for $\delta > 0$ small. We denote by $|I_{n,\delta}|$ its cardinal. Then for $(\beta,\Gamma) \in \mathcal{D}_n$, since $\sum_{i=1}^n V(p_{0,i}, p_{\beta,\Gamma,i}) \leq c_1 \log(n)$, we have that, on the one hand:

$$\sum_{i=1}^n\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 \leq c_1\log(n),$$

and on the other hand,

$$\begin{aligned}\sum_{i=1}^n\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 &= \sum_{i\in I_{n,\delta}}\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 + \sum_{i\notin I_{n,\delta}}\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 \\ &\gtrsim \delta|I_{n,\delta}| + \sum_{i\notin I_{n,\delta}}\sum_{k=1}^{n_i}\left(1-\frac{1}{\rho_{i,k}}\right)^2\end{aligned}$$

since for $i \notin I_{n,\delta}$, $\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 < \delta$, for $\delta > 0$ small, so each $|1-\rho_{i,k}|$ is less than $\sqrt{\delta}$, and we have that $|1-x^{-1}| \lesssim |1-x| \lesssim |1-x^{-1}|$ for $x \to 1$, so $|1-\rho_{i,k}| \gtrsim |1-\rho_{i,k}^{-1}|$. Then, by using Lemma 10 of Jeong and Ghosal (2021b), we obtain that

$$\sum_{i=1}^n\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 \gtrsim \delta|I_{n,\delta}| + \sum_{i\notin I_{n,\delta}}\frac{1}{\rho_{\max}^2(\Delta_{\Gamma_0,i})}\|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2$$

By (11), we obtain that

$$c_1\log(n) \geq \sum_{i=1}^n\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 \gtrsim \delta|I_{n,\delta}| + \frac{1}{\overline{\rho_0}^2}\sum_{i\notin I_{n,\delta}}\|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2$$

that is equivalent to

$$\frac{\log(n)}{n} \gtrsim \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 \gtrsim \delta\frac{|I_{n,\delta}|}{n} + \frac{1}{n\overline{\rho_0}^2}\sum_{i\notin I_{n,\delta}}\|\Delta_{\Gamma,i}-\Delta_{\Gamma_0,i}\|_F^2$$

In particular, $\delta\dfrac{|I_{n,\delta}|}{n} \lesssim \dfrac{\log(n)}{n}$ for $\delta > 0$ small that implies $|I_{n,\delta}| \lesssim \log(n)$. We have also that $\dfrac{\log(n)}{n} \gtrsim \dfrac{1}{n\underline{\rho_0}^2}\sum_{i\notin I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2$, that is

$$\frac{\log(n)}{n} \gtrsim \frac{1}{n}\sum_{i\notin I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2. \tag{13}$$

Then,

$$\begin{aligned}\frac{1}{n}\sum_{i\notin I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 &= \frac{1}{n}\sum_{i=1}^{n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 - \frac{1}{n}\sum_{i\in I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2\\ &\geq d_n^2(\Gamma,\Gamma_0) - \frac{|I_{n,\delta}|}{n}\max_{1\leq i\leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2\end{aligned}$$

By Lemma B3, with Assumptions A7, A8 and A9, for $\Gamma_1,\Gamma_2 \in \mathcal{H}$, we have that

$$\max_i \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2 \lesssim \|\Gamma_1 - \Gamma_2\|_F^2 \lesssim d_n^2(\Gamma_1,\Gamma_2).$$

Thus, there exist some constants $M_1 > 0$ and $M_2 > 0$:

$$\begin{aligned}\frac{1}{n}\sum_{i\notin I_{n,\delta}} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 &\geq M_1 \max_{1\leq i\leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 - \frac{|I_{n,\delta}|}{n}\max_{1\leq i\leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2\\ &\geq \left(M_1 - \frac{|I_{n,\delta}|}{n}\right)\max_{1\leq i\leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2\\ &\geq \left(M_1 - M_2\frac{\log(n)}{n}\right)\max_{1\leq i\leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 \end{aligned}\tag{14}$$

since $|I_{n,\delta}| \lesssim \log(n)$. Finally, by combining (13) and (14), since $\dfrac{\log(n)}{n} \to 0$, we obtain that, on $\mathcal{D}_n$,

$$\max_{1\leq i\leq n} \|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 \lesssim \frac{\log(n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Hence, by using Equation (12), we obtain that:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} V(p_{0,i}, p_{\beta,\Gamma,i}) &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{n_i}(1-\rho_{i,k})^2 + \frac{1}{n}\sum_{i=1}^{n}\|\Delta_{\Gamma_0,i}^{1/2}\Delta_{\Gamma,i}^{-1}\left(f_i(X_i\beta) - f_i(X_i\beta_0)\right)\|_2^2\\ &\lesssim \frac{1}{n}\sum_{i=1}^{n}\|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\Delta_{\Gamma_0,i}^{1/2}\|_{sp}^2\|\Delta_{\Gamma,i}^{-1}\|_{sp}^2\|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2\end{aligned}$$

since $\|Ax\|_2 \leq \|A\|_{sp}\|x\|_2$. Then, by Lemma B2, we have that $\|\Delta_{\Gamma_0,i}^{1/2}\|_{sp}^2 = \rho_{max}(\Delta_{\Gamma_0,i}) \lesssim 1$ for each $i \in \{1,\ldots,n\}$. Moreover, on $\mathcal{D}_n$,

$$\begin{aligned}\|\Delta_{\Gamma,i}^{-1}\|_{sp} = \rho_{max}(\Delta_{\Gamma,i}^{-1}) &= \rho_{max}(\Delta_{\Gamma_0,i}^{-1/2}\Delta_{\Gamma_0,i}^{1/2}\Delta_{\Gamma,i}^{-1}\Delta_{\Gamma_0,i}^{1/2}\Delta_{\Gamma_0,i}^{-1/2})\\ &\leq \rho_{max}(\Delta_{\Gamma_0,i}^{1/2}\Delta_{\Gamma,i}^{-1}\Delta_{\Gamma_0,i}^{1/2})\|\Delta_{\Gamma_0,i}^{-1/2}\|_{sp}^2 \text{ by Lemma B1}\\ &\lesssim \rho_{max}(\Delta_{\Gamma_0,i}^{1/2}\Delta_{\Gamma,i}^{-1}\Delta_{\Gamma_0,i}^{1/2}) = \max_k \rho_{i,k} \lesssim 1\end{aligned}$$

since $\|\Delta_{\Gamma_0,i}^{-1/2}\|_{sp}^2 = \rho_{max}(\Delta_{\Gamma_0,i}^{-1}) = \rho_{min}(\Delta_{\Gamma_0,i})^{-1} \leq \underline{\rho_0}^{-1}$ by Lemma B2, and since each $\rho_{i,k}$ tends to 1 on $\mathcal{D}_n$. Then, using Assumption A1, that is each f_i is $\overline{K_n}$-Lipschitz, we deduce that:

$$\frac{1}{n}\sum_{i=1}^{n} V(p_{0,i}, p_{\beta,\Gamma,i}) \lesssim d_n^2(\Gamma,\Gamma_0) + \frac{K_n^2}{n}\|X(\beta-\beta_0)\|_2^2.$$

Let us now focus on the term $K(p_{0,i}, p_{\beta,\Gamma,i})$. We have shown that on $\mathcal{D}_n$ each $|1-\rho_{i,k}|$ is small for each i and k, so: $\log(\rho_{i,k}) = \log(1-(1-\rho_{i,k})) \sim -(1-\rho_{i,k}) - \dfrac{(1-\rho_{i,k})^2}{2}$, and so $-\log(\rho_{i,k}) - (1-\rho_{i,k}) \sim \dfrac{(1-\rho_{i,k})^2}{2}$. Thus, by using Inequality (12):

$$\begin{aligned}
\frac{1}{n}K(p_{0,i}, p_{\beta,\Gamma,i}) &= \frac{1}{2n}\sum_{i=1}^{n}\left[-\sum_{k=1}^{n_i}\log(\rho_{i,k}) - \sum_{k=1}^{n_i}(1-\rho_{i,k}) + \|\Delta_{\Gamma,i}^{-1/2}\left(f_i(X_i\beta) - f_i(X_i\beta_0)\right)\|_2^2\right] \\
&\lesssim \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{1}{2}\|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 + \|\Delta_{\Gamma,i}^{-1/2}\|_{sp}^2\|f_i(X_i\beta) - f_i(X_i\beta_0)\|_2^2\right) \\
&\lesssim d_n^2(\Gamma,\Gamma_0) + \frac{K_n^2}{n}\|X(\beta-\beta_0)\|_2^2,
\end{aligned}$$

by Assumption A1 and since $\|\Delta_{\Gamma,i}^{-1}\|_{sp} \lesssim 1$ on $\mathcal{D}_n$. Thus, we obtain finally that $\frac{1}{n}K(p_{0,i}, p_{\beta,\Gamma,i})$ and $\frac{1}{n}V(p_{0,i}, p_{\beta,\Gamma,i})$ are bounded above by $d_n^2(\Gamma,\Gamma_0) + \dfrac{K_n^2}{n}\|X(\beta-\beta_0)\|_2^2$ up to a multiplicative constant. Then, for c_1 large enough,

$$\begin{aligned}
\Pi(\mathcal{D}_n) &= \Pi\left((\beta,\Gamma)\in\mathcal{B}\times\mathcal{H}\,\middle|\,\frac{1}{n}\sum_{i=1}^{n}K(p_{0,i}, p_{\beta,\Gamma,i}) \le c_1\frac{\log(n)}{n}, \frac{1}{n}\sum_{i=1}^{n}V(p_{0,i}, p_{\beta,\Gamma,i}) \le c_1\frac{\log(n)}{n}\right) \\
&\ge \Pi\left((\beta,\Gamma)\in\mathcal{B}\times\mathcal{H}\,\middle|\,d_n^2(\Gamma,\Gamma_0) + \frac{K_n^2}{n}\|X(\beta-\beta_0)\|_2^2 \le 2\frac{\log(n)}{n}\right) \\
&\ge \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,d_n^2(\Gamma,\Gamma_0) \le \frac{\log(n)}{n}\right)\Pi\left(\beta\in\mathcal{B}\,\middle|\,\frac{K_n^2}{n}\|X(\beta-\beta_0)\|_2^2 \le \frac{\log(n)}{n}\right) \\
&\ge \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,d_n^2(\Gamma,\Gamma_0) \le \frac{\log(n)}{n}\right)\Pi\left(\beta\in\mathcal{B}\,\middle|\,\frac{K_n^2}{n}\|X\|_*^2\|\beta-\beta_0\|_1^2 \le \frac{\log(n)}{n}\right)
\end{aligned}$$

since $\|X\theta\|_2 \le \|X\|_*\|\theta\|_1$. For the first term, we have that:

$$\begin{aligned}
\Pi\left(\Gamma\in\mathcal{H}\,\middle|\,d_n^2(\Gamma,\Gamma_0) \le \frac{\log(n)}{n}\right) &= \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,\frac{1}{n}\sum_{i=1}^{n}\|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 \le \frac{\log(n)}{n}\right) \\
&\ge \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,\max_i\|\Delta_{\Gamma,i} - \Delta_{\Gamma_0,i}\|_F^2 \le \frac{\log(n)}{n}\right) \\
&\ge \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,\|\Gamma-\Gamma_0\|_F^2 \le \frac{1}{\overline{\rho_Z}^4}\frac{\log(n)}{n}\right) \\
&= \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,\|\Gamma-\Gamma_0\|_F \le \frac{1}{\overline{\rho_Z}^2}\sqrt{\frac{\log(n)}{n}}\right)
\end{aligned}$$

by Lemma B3 and Assumption A9. Thus, by Assumption A3:

$$\begin{aligned}
\|\Gamma-\Gamma_0\|_F &= \|\Gamma_0^{1/2}(\Gamma_0^{-1/2}\Gamma\Gamma_0^{-1/2} - Id)\Gamma_0^{1/2}\|_F \\
&\le \|\Gamma_0\|_{sp}\|\Gamma_0^{-1/2}\Gamma\Gamma_0^{-1/2} - Id\|_F \\
&\le \overline{\rho_\Gamma}\|\Gamma_0^{-1/2}\Gamma\Gamma_0^{-1/2} - Id\|_F
\end{aligned}$$

By using Lemma B4, we obtain that:

$$\begin{aligned}
\Pi\left(\Gamma\in\mathcal{H}\,\middle|\,d_n^2(\Gamma,\Gamma_0) \le \frac{\log(n)}{n}\right) &\ge \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,\|\Gamma_0^{-1/2}\Gamma\Gamma_0^{-1/2} - Id\|_F \le \overline{\rho_\Gamma}^{-1}\overline{\rho_Z}^{-2}\sqrt{\frac{\log(n)}{n}}\right) \\
&\ge \Pi\left(\Gamma\in\mathcal{H}\,\middle|\,\bigcap_{k=1}^{q}\left\{1 \le \rho_k(\Gamma_0^{-1/2}\Gamma\Gamma_0^{-1/2}) \le 1 + \overline{\rho_\Gamma}^{-1}\overline{\rho_Z}^{-2}\sqrt{\frac{\log(n)}{qn}}\right\}\right)
\end{aligned}$$

Then, denoting by $A = \Gamma_0^{-1/2}\Gamma\Gamma_0^{-1/2} \in \mathbb{R}^{q\times q}$, since $\Gamma \sim \mathcal{IW}_q(d,\Sigma)$, we known that $A \sim \mathcal{IW}_q(d, \Gamma_0^{-1/2}\Sigma\Gamma_0^{-1/2})$, and so $A^{-1} \sim \mathcal{W}_q(d, \Gamma_0^{1/2}\Sigma^{-1}\Gamma_0^{1/2})$. Then, using Lemma 6.3 of Ning et al. (2020)) on the eigenvalues of a Wishart distribution, we obtain that, for large n and since q is fixed:

$$\Pi\left(\Gamma \in \mathcal{H}\middle| d_n^2(\Gamma,\Gamma_0) \le \frac{\log(n)}{n}\right) \ge \left(\frac{a_1 t e^2 d}{8\sqrt{\pi}}\right)^{-q}\left(\frac{2dq}{ea_1 t}\right)^{-dq/2}\left(\frac{d}{2e}\right)^{-q^2/2}(\det(\Gamma_0^{1/2}\Sigma^{-1}\Gamma_0^{1/2}))^{-d/2}\times \exp\left(-\frac{a_1(1+t)\mathrm{Tr}(\Gamma_0^{-1/2}\Sigma\Gamma_0^{-1/2})}{2}\right)$$

with $a_1 = \left(1+\overline{\rho_\Gamma}^{-1}\overline{\rho_Z}^{-2}\sqrt{\frac{\log(n)}{qn}}\right)^{-1}$ and $t = \overline{\rho_\Gamma}^{-1}\overline{\rho_Z}^{-2}\sqrt{\frac{\log(n)}{qn}}$.

Finally, for n large enough, we have that:

$$\log\left(\Pi\left(\Gamma \in \mathcal{H}\middle| d_n^2(\Gamma,\Gamma_0) \le \frac{\log(n)}{n}\right)\right) \gtrsim -\log(n).$$

Concerning the second term $\Pi\left(\beta \in \mathcal{B}\middle|\frac{K_n^2}{n}\|X\|_*^2\|\beta-\beta_0\|_1^2 \le \frac{\log(n)}{n}\right)$ in the lower bound of $\Pi(\mathcal{D}_n)$, by defining $\mathcal{B}_{S_0,n} = \left\{\beta_{S_0} \in \mathbb{R}^{s_0}\middle|\frac{K_n}{\sqrt{n}}\|X\|_*\|\beta_{S_0}-\beta_{0,S_0}\|_1 \le \sqrt{\frac{\log(n)}{n}}\right\}$ we have that:

$$\begin{aligned}\Pi\left(\beta \in \mathcal{B}\middle|\frac{K_n^2}{n}\|X\|_*^2\|\beta-\beta_0\|_1^2 \le \frac{\log(n)}{n}\right) &\ge \Pi\left(S=S_0, \beta \in \mathcal{B}\middle|\frac{K_n}{\sqrt{n}}\|X\|_*\|\beta-\beta_0\|_1 \le \sqrt{\frac{\log(n)}{n}}\right)\\ &\ge \frac{\pi_p(s_0)}{\binom{p}{s_0}}\int_{\mathcal{B}_{S_0,n}} g_{S_0}(\beta_{S_0})d\beta_{S_0}\\ &\ge \frac{\pi_p(s_0)}{\binom{p}{s_0}}e^{-\lambda\|\beta_0\|_1}\int_{\mathcal{B}_{S_0,n}} g_{S_0}(\beta_{S_0}-\beta_{0,S_0})d\beta_{S_0}\end{aligned}$$

because g_S is the Laplace distribution so satisfy the inequality $g_{S_0}(\beta_{S_0}) \ge e^{-\lambda\|\beta_0\|_1}g_{S_0}(\beta_{S_0}-\beta_{0,S_0})$. Then, since $s_0 > 0$ by Assumption A4 and using the equation (6.2) of Castillo et al. (2015), we obtain that:

$$\begin{aligned}\int_{\mathcal{B}_{S_0,n}} g_{S_0}(\beta_{S_0}-\beta_{0,S_0})d\beta_{S_0} &\ge e^{-\lambda\frac{\sqrt{\log(n)}}{K_n\|X\|_*}}\frac{\left(\lambda\frac{\sqrt{\log(n)}}{K_n\|X\|_*}\right)^{s_0}}{s_0!}\\ &\ge e^{-L_3\sqrt{\frac{\log(n)}{n}}}\frac{\left(\frac{\sqrt{\log(n)}}{L_1 p^{L_2}}\right)^{s_0}}{s_0!}\end{aligned}$$

by Assumption A6. We deduce that, by using $\binom{p}{s_0}s_0! \le p^{s_0}$:

$$
\begin{aligned}
&\Pi\left(\beta \in \mathcal{B} \middle| \frac{K_n^2}{n}\|X\|_*^2\|\beta-\beta_0\|_1^2 \leq \frac{\log(n)}{n}\right) \\
&\geq \frac{\pi_p(s_0)}{\binom{p}{s_0}} e^{-\lambda\|\beta_0\|_1} e^{-L_3\sqrt{\frac{\log(n)}{n}}} \frac{\left(\frac{\sqrt{\log(n)}}{L_1 p^{L_2}}\right)^{s_0}}{s_0!} \\
&\geq \pi_p(s_0)\log(n)^{s_0/2}\exp\left(-(L_2-1)s_0\log(p) - L_3\sqrt{\frac{\log(n)}{n}} - s_0\log(L_1) - \lambda\|\beta_0\|_1\right) \\
&\geq \pi_p(s_0)\exp\left(-\tilde{C}s_0\log(p)\right)
\end{aligned}
$$

for a constant $\tilde{C}$ since $s_0 + \sqrt{\frac{\log(n)}{n}} + s_0\log(p) \lesssim s_0\log(p)$ and $\lambda\|\beta_0\|_1 \lesssim s_0\log(p)$ by Assumption A2. Finally, there exists a constant L such that:

$$
\begin{aligned}
\Pi(\mathcal{D}_n) &\geq \Pi\left(\Gamma \in \mathcal{H} \middle| d_n^2(\Gamma,\Gamma_0) \leq \frac{\log(n)}{n}\right)\Pi\left(\beta \in \mathcal{B} \middle| \frac{K_n^2}{n}\|X\|_*^2\|\beta-\beta_0\|_1^2 \leq \frac{\log(n)}{n}\right) \\
&\geq \pi_p(s_0)e^{-L(s_0\log(p)+\log(n))}
\end{aligned}
$$

Thus, we have shown that $e^{-(1+C)c_1\log(n)}\Pi(\mathcal{D}_n) \geq \pi_p(s_0)e^{-M(s_0\log(p)+\log(n))}$ for some constant M, as required to conclude the proof of Lemma 1.

A.2 Proof of Lemma 2

Let $(\beta_1,\Gamma_1) \in \mathcal{B}\times\mathcal{H}$ such that $R_n(p_0,p_1) \geq \epsilon_n^2$. First, for testing H_0: $p = p_0$ against H_1: $p = p_1$, consider the most powerful test $\overline{\varphi}_n = \mathbb{1}_{\Lambda_n(\beta_1,\Gamma_1)\geq 1}$ given by the Neyman-Pearson lemma, where $\Lambda_n(\beta_1,\Gamma_1) = \dfrac{p_1}{p_0}$ be the likelihood ratio of p_1 and p_0. Thus,

$$
\begin{aligned}
\mathbb{E}_0[\overline{\varphi}_n] = \mathbb{P}_0(\sqrt{\Lambda_n(\beta_1,\Gamma_1)} \geq 1) &= \int \mathbb{1}_{\sqrt{p_1(y)}\geq\sqrt{p_0(y)}}p_0(y)dy \\
&\leq \int \sqrt{p_0(y)p_1(y)}dy = e^{-nR_n(p_0,p_1)} \leq e^{-n\epsilon_n^2}
\end{aligned}
$$

by assumption on (β_1,Γ_1). This proves the first result of the lemma.

Then, for the second part of the lemma, note that:

$$
\mathbb{E}_1[1-\overline{\varphi}_n] = \mathbb{P}_1(\sqrt{\Lambda_n(\beta_1,\Gamma_1)} \leq 1) \leq \int \sqrt{p_0(y)p_1(y)}dy \leq e^{-n\epsilon_n^2} \tag{15}
$$

However, by using Cauchy-Schwarz inequality:

$$
\begin{aligned}
\mathbb{E}_{\beta,\Gamma}[1-\overline{\varphi}_n] &= \int(1-\overline{\varphi}_n(y))\frac{p_{\beta,\Gamma}(y)}{p_1(y)}dp_1(y) \\
&\leq \mathbb{E}_1[1-\overline{\varphi}_n]^{1/2}\mathbb{E}_1\left[\left(\frac{p_{\beta,\Gamma}}{p_1}\right)^2\right]^{1/2} \\
&\leq e^{-n\epsilon_n^2/2}\mathbb{E}_1\left[\left(\frac{p_{\beta,\Gamma}}{p_1}\right)^2\right]^{1/2}
\end{aligned}
$$

by Equation (15). Therefore, the test $\overline{\varphi}_n$ can also have exponentially small error of type II at other alternatives if we can controlled the second term: we want to show that $\mathbb{E}_1\left[\left(\frac{p_{\beta,\Gamma}}{p_1}\right)^2\right]^{1/2} \leq e^{7n\epsilon_n^2/16}$ for every $(\beta,\Gamma) \in \mathcal{F}_{1,n}$.

Recall that here $p_{\beta,\Gamma} = \prod_{i=1}^n \mathcal{N}_{n_i}(f_i(X_i\beta), \Delta_{\Gamma,i})$, where $\Delta_{\Gamma,i} = Z_i\Gamma Z_i^\top + \sigma^2 I_{n_i}$. By denoting $\Delta^*_{\Gamma,i} = \Delta_{\Gamma,i}^{-1/2}\Delta_{\Gamma_1,i}\Delta_{\Gamma,i}^{-1/2}$, then for $(\beta,\Gamma)\in\mathcal{F}_{1,n}$, if $2\Delta^*_{\Gamma,i} - Id$ and $2Id - \Delta^{*^{-1}}_{\Gamma,i}$ are non-singular matrices for every $i\in\{1,\cdots,n\}$, we can show that:

$$\begin{aligned}\mathbb{E}_1\left[\left(\frac{p_{\beta,\Gamma}}{p_1}\right)^2\right] =& \prod_{i=1}^n\left[\det(\Delta^*_{\Gamma,i})^{1/2}\det(2Id-\Delta^{*^{-1}}_{\Gamma,i})^{-1/2}\right]\times\\ &\exp\left\{\sum_{i=1}^n \|(2\Delta^*_{\Gamma,i}-Id)^{-1/2}\Delta_{\Gamma,i}^{-1/2}(f_i(X_i\beta)-f_i(X_i\beta_1))\|_2^2\right\}. \end{aligned} \tag{16}$$

Let us now prove that these matrices are non-singular. We have, for all $k\le n_i$,

$$\max_{1\le i\le n}\|\Delta^*_{\Gamma,i}-Id\|_{sp} = \max_{1\le i\le n}\rho_{max}(|\Delta^*_{\Gamma,i}-Id|)\ge\max_{1\le i\le n}|\rho_k(\Delta^*_{\Gamma,i})-1|. \tag{17}$$

Note that

$$\begin{aligned}\max_{1\le i\le n}\|\Delta^*_{\Gamma,i}-Id\|_{sp} &= \max_{1\le i\le n}\|\Delta_{\Gamma,i}^{-1/2}(\Delta_{\Gamma_1,i}-\Delta_{\Gamma,i})\Delta_{\Gamma,i}^{-1/2}\|_{sp}\\ &\le \max_{1\le i\le n}\|\Delta_{\Gamma,i}^{-1}\|_{sp}\|\Delta_{\Gamma_1,i}-\Delta_{\Gamma,i}\|_F\\ &\le \max_{1\le i\le n}\|\Delta_{\Gamma,i}^{-1}\|_{sp}d_n(\Gamma,\Gamma_1)\\ &\le \frac{\epsilon_n^2}{2J_n}\end{aligned}$$

by Lemma B3 and since $(\beta,\Gamma)\in\mathcal{F}_{1,n}$. Thus, for all $k\le n_i$, $\max_{1\le i\le n}|\rho_k(\Delta^*_{\Gamma,i})-1|\le\frac{\epsilon_n^2}{2J_n}$. We deduce that

$$1-\frac{\epsilon_n^2}{2J_n}\le\min_{1\le i\le n}\rho_{\min}(\Delta^*_{\Gamma,i})\le\max_{1\le i\le n}\rho_{\max}(\Delta^*_{\Gamma,i})\le 1+\frac{\epsilon_n^2}{2J_n}. \tag{18}$$

Therefore, since $\frac{\epsilon_n^2}{2J_n}\underset{n\to\infty}{\longrightarrow}0$ by Assumption A4, and for all $k\le n_i$, $\rho_k(2\Delta^*_{\Gamma,i}-Id) = 2\rho_k(\Delta^*_{\Gamma,i})-1$ and $\rho_k(2Id-\Delta^{*^{-1}}_{\Gamma,i}) = 2-\rho_k(\Delta^{*^{-1}}_{\Gamma,i}) = 2-\rho_k^{-1}(\Delta^*_{\Gamma,i})$, we deduce that $2\Delta^*_{\Gamma,i}-Id$ and $2Id-\Delta^{*^{-1}}_{\Gamma,i}$ are non-singular on $\mathcal{F}_{1,n}$ for every $i\in\{1,\cdots,n\}$.

For concluding the proof, it remains to bound the right side term of (16). By using (18) and the inequalities $(1-x^2)/(1-2x)\le 1+3x$ for $x>0$ small, and $1+x\le e^x$, we obtain for n large enough:

$$\begin{aligned}\det(\Delta^*_{\Gamma,i})^{1/2}\det(2Id-\Delta^{*^{-1}}_{\Gamma,i})^{-1/2} &= \left(\prod_{k=1}^{n_i}\rho_k(\Delta^*_{\Gamma,i})\right)^{1/2}\left(\prod_{k=1}^{n_i}2-\rho_k^{-1}(\Delta^*_{\Gamma,i})\right)^{-1/2}\\ &= \left(\prod_{k=1}^{n_i}\frac{\rho_k(\Delta^*_{\Gamma,i})}{2-\rho_k^{-1}(\Delta^*_{\Gamma,i})}\right)^{1/2}\\ &\le \left(\frac{1-\frac{\epsilon_n^4}{4J_n^2}}{1+\frac{\epsilon_n^2}{J_n}}\right)^{n_i/2} \le \left(1+3\frac{\epsilon_n^2}{2J_n}\right)^{n_i/2}\le\exp\left(3\frac{n_i\epsilon_n^2}{4J_n}\right)\le e^{3\epsilon_n^2/4}\end{aligned}$$

Moreover, for n large enough,

$$\begin{aligned}
&\sum_{i=1}^{n} \|(2\Delta^*_{\Gamma,i} - Id)^{-1/2}\Delta_{\Gamma,i}^{-1/2}(f_i(X_i\beta) - f_i(X_i\beta_1))\|_2^2 \\
&\leq \max_{1\leq i\leq n} \|(2\Delta^*_{\Gamma,i} - Id)^{-1}\|_{sp} \max_{1\leq i\leq n} \|\Delta_{\Gamma,i}^{-1}\|_{sp} \sum_{i=1}^{n} \|f_i(X_i\beta) - f_i(X_i\beta_1)\|_2^2 \\
&\leq 2\gamma_n \frac{n\epsilon_n^2}{16\gamma_n} = \frac{n\epsilon_n^2}{8}
\end{aligned}$$

since $(\beta,\Gamma) \in \mathcal{F}_{1,n}$. Finally, by using (16), we conclude that $\mathbb{E}_1\left[\left(\frac{p_{\beta,\Gamma}}{p_1}\right)^2\right]^{1/2} \leq e^{3n\epsilon_n^2/8}e^{n\epsilon_n^2/16} = e^{7n\epsilon_n^2/16}$, and so $\sup_{(\beta,\Gamma)\in\mathcal{F}_{1,n}} \mathbb{E}_{\beta,\Gamma}[1-\overline{\varphi}_n] \leq e^{-n\epsilon_n^2/16}$, which concludes the proof.

Appendix B Useful lemmas

Lemma B1. *For A and B two matrices, $\rho_{min}(B)\|A\|_{sp}^2 \leq \rho_{max}(ABA^\top) \leq \rho_{max}(B)\|A\|_{sp}^2$.*

Proof. By the Courant–Fischer–Weyl min-max principle,

$$\begin{aligned}
\rho_{max}(ABA^\top) &= \max_{x\neq 0} \frac{\langle ABA^\top x, x\rangle}{\|x\|^2} \\
&= \max_{x\neq 0} \frac{\langle BA^\top x, A^\top x\rangle}{\|x\|^2} \\
&= \max_{x\neq 0} \frac{\langle BA^\top x, A^\top x\rangle}{\|A^\top x\|^2} \frac{\|A^\top x\|^2}{\|x\|^2} \\
&\leq \rho_{max}(B) \max_{x\neq 0} \frac{\|A^\top x\|^2}{\|x\|^2} \\
&= \rho_{max}(B)\rho_{max}(AA^\top) \\
&= \rho_{max}(B)\|A\|_{sp}^2
\end{aligned}$$

We obtain the other inequality with similar arguments. □

Lemma B2. *Grant Assumptions A3 and A9. Thus, $\Delta_{\Gamma_0,i} := Z_i\Gamma_0 Z_i^\top + \sigma^2 I_{n_i}$ satisfies:*

$$1 \lesssim \min_i \rho_{min}(\Delta_{\Gamma_0,i}) \leq \max_i \rho_{max}(\Delta_{\Gamma_0,i}) \lesssim 1$$

Proof. By the Weyl's inequality, for $1 \leq i \leq n$,

$$\rho_{min}(\Delta_{\Gamma_0,i}) \geq \rho_{min}(Z_i\Gamma_0 Z_i^\top) + \sigma^2 \geq \sigma^2$$

since $Z_i\Gamma_0 Z_i^\top$ is a positive definite matrix. Thus $\min_i \rho_{min}(\Delta_{\Gamma_0,i}) \geq \sigma^2$, otherwise, $\min_i \rho_{min}(\Delta_{\Gamma_0,i}) \gtrsim 1$. For the other inequality, by the Weyl's inequality, we have that

$$\rho_{max}(\Delta_{\Gamma_0,i}) \leq \rho_{max}(Z_i\Gamma_0 Z_i^\top) + \sigma^2.$$

Then, by Lemma B1, we have that:

$$\rho_{max}(Z_i\Gamma_0 Z_i^\top) \leq \rho_{max}(\Gamma_0)\|Z_i\|_{sp}^2$$

and by Assumptions A3 and A9,

$$\max_i \rho_{max}(\Delta_{\Gamma_0,i}) \leq \rho_{max}(\Gamma_0) \max_i \|Z_i\|_{sp}^2 + \sigma^2 \lesssim 1.$$

□

Lemma B3. *For $\Gamma_1, \Gamma_2 \in \mathcal{H}$, under Assumptions A7, A8 and A9, we have that*

$$\max_i \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2 \lesssim \|\Gamma_1 - \Gamma_2\|_F^2 \lesssim d_n^2(\Gamma_1, \Gamma_2) = \frac{1}{n}\sum_{i=1}^n \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2.$$

Proof. First,

$$\|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2 = \|Z_i(\Gamma_1 - \Gamma_2)Z_i^\top\|_F^2 \le \|Z_i\|_{sp}^4 \|\Gamma_1 - \Gamma_2\|_F^2,$$

since $\|AB\|_F \le \|A\|_{sp}\|B\|_F$, and by Assumption A9, we have that

$$\max_{1\le i\le n} \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2 \lesssim \|\Gamma_1 - \Gamma_2\|_F^2.$$

Then, by Assumption A8, for each i such that $n_i \ge q$, $Z_i^\top Z_i$ is invertible and

$$\|\Gamma_1 - \Gamma_2\|_F^2 = \|(Z_i^\top Z_i)^{-1} Z_i^\top Z_i(\Gamma_1 - \Gamma_2)Z_i^\top Z_i(Z_i^\top Z_i)^{-1}\|_F^2 \le \|Z_i(\Gamma_1 - \Gamma_2)Z_i^\top\|_F^2 \|(Z_i^\top Z_i)^{-1} Z_i^\top\|_{sp}^4.$$

Then, by Assumption A7, we have that

$$\begin{aligned}
\max_{1\le i\le n} \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2 \lesssim \|\Gamma_1 - \Gamma_2\|_F^2 &\le \frac{1}{\sum_{i=1}^n \mathbb{1}_{n_i \ge q}} \sum_{i:n_i\ge q} \|Z_i(\Gamma_1 - \Gamma_2)Z_i^\top\|_F^2 \|(Z_i^\top Z_i)^{-1} Z_i^\top\|_{sp}^4 \\
&\lesssim \frac{1}{n} \sum_{i:n_i\ge q} \|Z_i(\Gamma_1 - \Gamma_2)Z_i^\top\|_F^2 \|(Z_i^\top Z_i)^{-1} Z_i^\top\|_{sp}^4 \\
&\lesssim \frac{1}{n}\sum_{i=1}^n \|Z_i(\Gamma_1 - \Gamma_2)Z_i^\top\|_F^2 = \frac{1}{n}\sum_{i=1}^n \|\Delta_{\Gamma_1,i} - \Delta_{\Gamma_2,i}\|_F^2 = d_n^2(\Gamma_1, \Gamma_2).
\end{aligned}$$

where the last inequality uses assumptions A8 and A9. □

Lemma B4. *For a positive definite symmetric matrix $A \in \mathbb{R}^{q\times q}$ such as its eigenvalues satisfy $1 \le \rho_1(A) \le ... \le \rho_q(A) \le 1 + \frac{\epsilon}{\sqrt{q}}$, then $\|A - I_q\|_F \le \epsilon$.*

Proof. Observe that

$$\begin{aligned}
\|A - I_q\|_F \le \epsilon &\Leftrightarrow \mathrm{Tr}((A - I_q)^2) \le \epsilon^2 \quad \text{since } A \text{ is symmetric} \\
&\Leftrightarrow \sum_{k=1}^q \rho_k(A - I_q)^2 \le \epsilon^2 \\
&\Leftrightarrow \sum_{k=1}^q (\rho_k(A) - 1)^2 \le \epsilon^2.
\end{aligned}$$

By assumption, for $1 \le k \le q$, we have that $0 \le \rho_k(A) - 1 \le \frac{\epsilon}{\sqrt{q}}$ and then $\max_{1\le k\le q}(\rho_k(A) - 1)^2 \le \frac{\epsilon^2}{q}$. Hence, since $\sum_{k=1}^q (\rho_k(A) - 1)^2 \le q \times \max_{1\le k\le q}(\rho_k(A) - 1)^2 \le \epsilon^2$ and so $\|A - I_q\|_F \le \epsilon$. □