Inference for High-dimensional Maximin Effects in Heterogeneous Regression Models Using a Sampling Approach

Zijian Guo

Abstract

Heterogeneity is an important feature of modern data sets and a central task is to extract information from large-scale and heterogeneous data. In this paper, we consider multiple high-dimensional linear models and adopt the definition of maximin effect (Meinshausen, Bühlmann, AoS, 43(4), 1801–1830) to summarize the information contained in this heterogeneous model. We define the maximin effect for a targeted population whose covariate distribution is possibly different from that of the observed data. We further introduce a ridge-type maximin effect to simultaneously account for reward optimality and statistical stability. To identify the high-dimensional maximin effect, we estimate the regression covariance matrix by a debiased estimator and use it to construct the aggregation weights for the maximin effect. A main challenge for statistical inference is that the estimated weights might have a mixture distribution and the resulted maximin effect estimator is not necessarily asymptotic normal. To address this, we devise a novel sampling approach to construct the confidence interval for any linear contrast of high-dimensional maximin effects. The coverage and precision properties of the proposed confidence interval are studied. The proposed method is demonstrated over simulations and a genetic data set on yeast colony growth under different environments.

KEYWORDS: Heterogeneity; stability; Ridge-type maximin effect; covariate shift; confidence interval; adversarial reward.

[0] Z. Guo is an assistant professor at Department of Statistics, Rutgers University. The research of Z. Guo was supported in part by the NSF DMS 1811857, 2015373 and NIH R01GM140463-01.

1 Introduction

Data heterogeneity commonly occurs in modern data analysis, for example, the data being collected from different sub-populations (Lubke and Muthén, 2005; Tsuboi et al., 2009) or the experiments being conducted under different environments (Deusser, 1972; Bloom et al., 2013). A central task is to effectively extract and summarize information from the massive and potentially heterogeneous data. For example, in the regression framework, one important yet challenging question is how to measure the relationship between the outcome variable and a given covariate of interest across different sub-populations and environments. Despite of its importance, there is a lack of inference methods for analyzing heterogeneous data.

To model the data heterogeneity, we consider multiple high-dimensional linear models across a total of L groups of data,

$$Y^{(l)}_{n_l \times 1} = X^{(l)}_{n_l \times p} b^{(l)}_{p \times 1} + \epsilon^{(l)}_{n_l \times 1} \quad \text{for} \quad 1 \leq l \leq L, \tag{1}$$

where for the group l, $b^{(l)} \in \mathbb{R}^p$ denotes the regression vector, n_l denotes the sample size and $Y^{(l)} \in \mathbb{R}^{n_l}, X^{(l)} \in \mathbb{R}^{n_l \times p}$ and $\epsilon^{(l)} \in \mathbb{R}^{n_l}$ denote outcome, design matrix and the model error, respectively. The group labels $\{1, 2, \ldots, L\}$ may correspond to different data sources, for example, different sub-populations or experiment environments. Throughout the paper, the group label l is assumed to be known a priori. Within the group l, $\{(X^{(l)}_{i,\cdot}, Y^{(l)}_i)\}_{1 \leq i \leq n_l}$ are identically and independently generated from the distribution $\mathcal{P}^l$ with $X^{(l)}_{i,\cdot}$ following the covariate distribution $\mathcal{P}^l_X$ and the model errors $\epsilon^{(l)}$ being independent of $X^{(l)}$. Across the groups $\{1, \ldots, L\}$, P^l_X, $b^{(l)} \in \mathbb{R}^p$ and the distribution of $\epsilon^{(l)}$ are allowed to vary.

To identify an effect vector which is representative across all L groups, Meinshausen and Bühlmann (2015) introduced the maximin effect $\beta^* \in \mathbb{R}^p$ as

$$\beta^* := \arg\max_{\beta \in \mathbb{R}^p} R(\beta) \quad \text{with} \quad R(\beta) = \min_{1 \leq l \leq L} \left\{ \mathbf{E}_{\mathcal{P}^l}[Y^{(l)}_1]^2 - \mathbf{E}_{\mathcal{P}^l}[Y^{(l)}_1 - (X^{(l)}_{1,\cdot})^{\intercal}\beta]^2 \right\}. \tag{2}$$

We refer to $R(\beta)$ as the (adversarial) reward since it measures the minimum explained variance of $\beta \in \mathbb{R}^p$ across L groups. The maximin effect $\beta^* \in \mathbb{R}^p$ is defined to maximize this adversarial reward $R(\beta)$. For $L = 1$ (the homogeneous setting), the maximin effect β^* is reduced to the population best linear approximation. For $L \geq 2$, Meinshausen and Bühlmann (2015) showed that β^* can be identified as a weighted average of $\{b^{(l)}\}_{1 \leq l \leq L}$ in (1) with the aggregation weights as solutions to a quadratic optimization problem.

We focus on the model (1) in the high-dimensional setting with $p \gg \max_{1\leq l\leq L}\{n_l\}$. We aim at two goals: a) introduce new concepts of maximin effects to accommodate for possible covariate shift and balance the reward optimality and statistical optimality; b) construct the confidence interval for any linear contrast of high-dimensional maximin effects introduced in the current paper.

1.1 Our results and contribution

We consider a targeted population with covariate distribution $\mathcal{Q}$ and study how to identify the best representation of the heterogeneous regression vectors $\{b^{(l)}\}_{1\leq l\leq L}$ in (1) for this targeted population. Due to the data heterogeneity, the targeted covariate distribution $\mathcal{Q}$ is possibly different from any of the covariate distributions $\{\mathcal{P}_X^l\}_{1\leq l\leq L}$ of the observed data. We introduce in equation (5) the covariate shift maximin effect $\beta^*(\mathcal{Q}) \in \mathbb{R}^p$ as a generalization of the maximin effect in (2). Similarly to β^*, $\beta^*(\mathcal{Q})$ can be identified as a weighted average of $\{b^{(l)}\}_{1\leq l\leq L}$ but the optimal aggregation weights depend on the targeted covariate distribution $\mathcal{Q}$. Specifically, the optimal aggregation weights for $\beta^*(\mathcal{Q})$ are determined by the regression covariance matrix $\Gamma^{\mathcal{Q}} \in \mathbb{R}^{L\times L}$, where $\Gamma^{\mathcal{Q}}_{l,k} = [b^{(l)}]^{\intercal}\Sigma^{\mathcal{Q}} b^{(k)}$ for $1 \leq l,k \leq L$, and $\Sigma^{\mathcal{Q}} = \mathbf{E}X^{\mathcal{Q}}_{1,\cdot}(X^{\mathcal{Q}}_{1,\cdot})^{\intercal}$ with $X^{\mathcal{Q}}_{1,\cdot} \in \mathbb{R}^p$ following the targeted distribution $\mathcal{Q}$.

We further introduce a ridge-type maximin effect which simultaneously accounts for reward optimality and statistical stability. The maximin effect in (2) only maximizes the reward $R(\beta)$, but an equally important goal is to define a summary objective which can be stably estimated. To achieve this, we impose a ridge penalty (with level $\delta > 0$) in constructing the aggregation weights and use these weights to define the ridge-type maximin effect $\beta^*_\delta(\mathcal{Q}) \in \mathbb{R}^p$; see (7). The penalty level δ is useful in balancing reward optimality and stability. In comparison to the non-penalized maximin effect, the ridge-type effect with $\delta > 0$ is a more stable target to make inference for, especially when the regression covariance matrix $\Gamma^{\mathcal{Q}}$ is nearly singular. In numerical studies, with a suitable $\delta > 0$, our proposed confidence intervals for $\beta^*_\delta(\mathcal{Q})$ are much shorter than those for the maximin effect with no penalty; see Figure 4.

Although the reward value of $\beta^*_\delta(\mathcal{Q})$ decreases with an increasing penalty level δ, we show in Theorem 1 that $\beta^*_\delta(\mathcal{Q})$ still satisfies certain reward optimality. Importantly, the reward reduction for any $\delta > 0$ can be estimated in a data-dependent way. This provides a practical method for users to choose the value δ such that the reward of $\beta^*_\delta(\mathcal{Q})$ is comparable to the optimal reward but inference procedures for $\beta^*_\delta(\mathcal{Q})$ are more stable; see Figure 3 and the related discussion.

We construct the confidence interval (CI) for the linear contrast $x_{\text{new}}^{\intercal}\beta^*_{\delta}(\mathcal{Q})$ for any $x_{\text{new}} \in \mathbb{R}^p$ and $\delta \geq 0$. Our proposed method requires an accurate estimator of the regression covariance matrix $\Gamma^{\mathcal{Q}}$. In the covariate-shift setting, $\Sigma^{\mathcal{Q}}$ is estimated by the sample covariance matrix $\widehat{\Sigma}^{\mathcal{Q}} = \sum_{i=1}^{N_{\mathcal{Q}}} X^{\mathcal{Q}}_{i,\cdot}(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}/N_{\mathcal{Q}}$, where $\{X^{\mathcal{Q}}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}$ are i.i.d. samples following the targeted distribution $\mathcal{Q}$. For $1 \leq l \leq L$, $\widehat{b}^{(l)}_{init}$ denotes a reasonable penalized estimator (e.g. Lasso) of $b^{(l)}$. We propose a debiased estimator $\widehat{\Gamma}^{\mathcal{Q}}$ of $\Gamma^{\mathcal{Q}}$ through conducting entry-wise bias correction of the plug-in estimator $[\widehat{b}^{(l)}_{init}]^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}$ for $1 \leq l, k \leq L$. A new projection direction is constructed to account for the possible covariate shift. We establish in Theorem 2 the asymptotic normality of our proposed matrix estimator $\widehat{\Gamma}^{\mathcal{Q}} \in \mathbb{R}^{L\times L}$.

A major challenge of uncertainty quantification for maximin effects is the possible non-normality of the maximin effect estimator. With $\widehat{\Gamma}^{\mathcal{Q}}$, we may estimate the optimal weight $\gamma^* \in \mathbb{R}^L$ by $\widehat{\gamma} \in \mathbb{R}^L$, defined as the maximizer of $\gamma^{\intercal}\widehat{\Gamma}^{\mathcal{Q}}\gamma$ over L-dimensional simplex. Even if $\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}}$ is asymptotically normal, the weight vector error $\widehat{\gamma} - \gamma^*$ can be the mixture of an asymptotic normal distribution and a point mass; see (13). This mixture distribution poses challenges for constructing CIs for the maximin effect using asymptotic normality.

To address this, we propose a novel sampling procedure to construct CIs for maximin effects. For $1 \leq m \leq M$, we sample $\widehat{\Gamma}^{[m]}$ following the asymptotic distribution of $\widehat{\Gamma}^{\mathcal{Q}}$ and use $\widehat{\Gamma}^{[m]}$ to construct a sampled weight vector $\widehat{\gamma}^{[m]} \in \mathbb{R}^L$ and a sampled interval $\text{Int}^{[m]}_{\alpha}(x_{\text{new}})$. We construct the CI for $x_{\text{new}}^{\intercal}\beta^*_{\delta}(\mathcal{Q})$ by taking a union of $\{\text{Int}^{[m]}_{\alpha}(x_{\text{new}})\}_{1\leq m\leq M}$. We establish in Theorem 3 that, for a large sampling number M, there exists a sampled weight vector converging to γ^* at a rate faster than the parametric rate. For such a sampled weight vector, the uncertainty of estimating γ^* is negligible. We establish the coverage property of our proposed CI and study its precision property by comparing its length to an oracle CI with the knowledge of γ^*. Our proposed CI is at most longer than the oracle CI by an order of $\sqrt{\log M}$; see Theorem 4.

We demonstrate the numerical performance of our proposed CI over different targeted distributions $\mathcal{Q}$ and ridge penalty levels δ.

To sum up, the contributions of the current paper are three-folded.

1. We introduce the covariate shift maximin effect and propose the ridge-type maximin effect to simultaneously account for reward optimality and statistical stability.

2. We propose a novel sampling approach to construct CIs for linear contrasts of various maximin effects. To the authors' best knowledge, the proposed CI is the first inference method for the high-dimensional maximin effect.

3. We characterize the dependence of sampling accuracy on the sampling number M. The theoretical analysis of Theorem 3 can be of independent interest.

1.2 Existing literature

The most relevant works are Meinshausen and Bühlmann (2015) and Rothenhäusler et al. (2016). Meinshausen and Bühlmann (2015) introduced the concept of maximin effect and studied its estimation problem and Rothenhäusler et al. (2016) constructed CI for the maximin effect β^* in low dimensions, relying on asymptotic normality results. The current paper focuses on CI construction in high dimensions and constructing CIs with a novel sampling procedure, instead of directly applying asymptotic normality. Beyond high-dimensionality, our proposed methods are valid for two important yet challenging settings (not covered by Rothenhäusler et al. (2016)): the maximin effect β^* is not uniquely defined and the asymptotic limiting distribution of the maximin effect estimator is non-normal. The maximin projection learning was proposed in Shi et al. (2018) for individualized treatment selection. We will show that the maximin projection can be identified as a rescaled version of the covariate shift maximin effect introduced in (5); see Proposition 4 in Section A.1 in the supplement.

Beyond the maximin effect, there is an active literature on analysis of heterogenous data for different purposes. Inference for the shared component of regression functions has been considered under multiple high-dimensional linear models (Liu et al., 2020) and partially linear models (Zhao et al., 2016). In contrast, the regression vectors $\{b^{(l)}\}_{1\leq l\leq L}$ in our model (1) are not required to share any similarity. Peters et al. (2015) and Rothenhäusler et al. (2018) developed invariance principle to identify causal effect with heterogenous data. Distributional robustness has been studied in Gao et al. (2017); Sinha et al. (2017) by minimizing a worst case over a class of distributions. We refer to Bühlmann (2018) as a review of the connection between invariance principle and distributional robustness. Inference in the presence of covariate shift has been studied in Tsuboi et al. (2009); Shimodaira (2000); Sugiyama et al. (2007) with weighting methods.

Inference for a single (homogeneous) high-dimensional linear model was actively investigated in the recent decade. Debiasing, desparsifying or Neyman's Orthogonalization (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Belloni et al., 2014; Chernozhukov et al., 2015; Farrell, 2015; Chernozhukov et al., 2018) were proposed for inference for regression coefficients. Inference problems for linear contrasts and quadratic functionals were studied in Cai and Guo (2017); Athey et al. (2018); Zhu and Bradic (2018);

Cai et al. (2019) and Verzelen and Gassiat (2018); Guo et al. (2019); Cai and Guo (2020); Guo et al. (2019), respectively. In comparison to this literature, the current paper focuses on a completely different task of aggregating heterogeneous regression vectors. The inference target is simultaneously determined by multiple linear models and the targeted covariate distribution. Our proposed estimator $\widehat{\Gamma}^{\mathcal{Q}}$ extends the existing debiasing methodology to the more challenging covariate shift setting; see Remark 3 for detailed comparisons. We shall emphasize that, this debiased estimator only serves as an initial estimator and a further novel sampling approach is required for inference for maximin effects.

Sampling methods have a long history in statistics, such as, bootstrap (Efron, 1979; Efron and Tibshirani, 1994), subsampling (Politis et al., 1999), generalized fiducial inference (Zabell, 1992; Xie and Singh, 2013; Hannig et al., 2016) and repro sampling (Wang and Xie, 2020). As a major difference, we do not directly sample from the original data but instead sample the estimator of the regression covariance matrix. This makes our proposed sampling approach computationally efficient even if a large number of samples are drawn.

Paper Organization. In Section 2, we define maximin effect for a targeted covariate distribution; in Section 3, we introduce the ridge-type maximin effect to balance reward optimality and statistical stability; in Section 4, we illustrate the challenges of inference for high-dimensional maximin effects. Our proposed method is detailed in Section 5 and the theoretical justification is provided in Section 6. In Section 7, we investigate the numerical performance of our proposed method; in Section 8, we apply the proposed method to a genetic data set on yeast colony growth under different environments. We provide conclusion and discussion in Section 9.

Notations. Define $n = \min_{1\leq l\leq L}\{n_l\}$. Let $[p] = \{1, 2, \ldots, p\}$. For a set S, $|S|$ denotes the cardinality of S and S^c denotes its complement. For a vector $x \in \mathbb{R}^p$ and a subset $S \subset [p]$, x_S is the sub-vector of x with indices in S and x_{-S} is the sub-vector with indices in S^c. The ℓ_q norm of a vector x is defined as $\|x\|_q = (\sum_{l=1}^{p} |x_l|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\{1 \leq l \leq p : x_l \neq 0\}|$ and $\|x\|_\infty = \max_{1\leq l\leq p} |x_l|$. For a matrix X, $X_{i,\cdot}$ and $X_{\cdot,j}$ are used to denote its i-th row and j-th column; for a set S, $X_{S,\cdot}$ denotes the sub-matrix of X with row indices belonging to S. For random objects X_1 and X_2, we use $X_1 \overset{d}{=} X_2$ to denote that they are equal in distribution. We use $X \sim \mathcal{Q}$ to denote that the random vector X follows the distribution $\mathcal{Q}$ and $\{X_{i,\cdot}\}_{1\leq i\leq n} \overset{\text{i.i.d.}}{\sim} \mathcal{Q}$ to denote that the sequence random vectors $X_{i,\cdot}$ are independently and identically distributed (i.i.d.) following the distribution $\mathcal{Q}$. For a sequence of random variables X_n indexed by n, we use $X_n \overset{p}{\to} X$ and $X_n \overset{d}{\to} X$ to represent that X_n converges to X in probability and in distribution, respectively. We use c and C

to denote generic positive constants that may vary from place to place. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means that $\exists C > 0$ such that $a_n \leq C b_n$ for all n; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\rightarrow\infty} a_n/b_n = 0$. For a matrix A, we use $\|A\|_F$, $\|A\|_2$ and $\|A\|_\infty$ to denote its Frobenius norm, spectral norm and element-wise maximum norm, respectively. For a symmetric matrix A, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its maximum and minimum eigenvalues, respectively. For a symmetric matrix $A \in \mathbb{R}^{L\times L}$ and its eigen-decomposition $A = U\Lambda U^\intercal$, we define $A_+ = U\Lambda_+ U^\intercal$ with $(\Lambda_+)_{l,l} = \max\{\Lambda_{l,l}, 0\}$ for $1 \leq l \leq L$. For a matrix B, we use $\lambda_j(B)$ to denote its j-th largest singular value.

For a symmetric matrix $D \in \mathbb{R}^{L\times L}$, we use $\text{vecl}(D) \in \mathbb{R}^{L(L+1)/2}$ to denote the vector of stacking the columns of the lower triangle part of D. $\mathcal{I}_L = \{(l,k) : 1 \leq k \leq l \leq L\}$ is the index set of the lower triangular part of D and $[L(L+1)/2]$ is the index set of $\text{vecl}(D)$. We define the one-to-one index mapping π from $\mathcal{I}_L$ to $[L(L+1)/2]$ as

$$\pi(l,k) = \frac{(2L-k)(k-1)}{2} + l \quad \text{for} \quad (l,k) \in \mathcal{I}_L \coloneqq \{(l,k) : 1 \leq k \leq l \leq L\}. \tag{3}$$

For $(l,k) \in \mathcal{I}_L$, we have $[\text{vecl}(D)]_{\pi(l,k)} = D_{l,k}$.

2 Maximin Effect for A Targeted Population

We consider the data $\left\{Y^{(l)}, X^{(l)}\right\}_{1\leq l\leq L}$, where the group labels $\{1, 2, \ldots, L\}$ are known a priori and $\{(Y_i^{(l)}, X_{i,\cdot}^{(l)})\}_{1\leq i\leq n_l} \overset{\text{i.i.d.}}{\sim} \mathcal{P}^l$ inside the group l. To capture the heterogeneity, we allow the distributions $\{\mathcal{P}^l\}_{1\leq l\leq L}$ to be different from each other. We consider the model (1) and the joint distribution $\mathcal{P}^l$ is determined by the covariate distribution $\mathcal{P}_X^l$, the regression vector $b^{(l)}$ and the distribution of $\epsilon_i^{(l)}$. In biology applications, we may take $X^{(l)}$ and $Y^{(l)}$ as the genetic variants and phenotype, respectively. When the group label l corresponds to a specific environment, $b^{(l)}$ captures how genetic variants affect the phenotype under this environment. With an environmental change, the regression vector $b^{(l)}$ might change across groups $\{1, 2, \cdots, L\}$. If the data is collected over different sub-populations, then the covariate distribution $\mathcal{P}_X^l$ is likely to change across $1 \leq l \leq L$. The changes in $b^{(l)}$ and $\mathcal{P}_X^l$ might happen simultaneously.

We use $\mathcal{Q}$ to denote a targeted covariate distribution on the p-dimensional covariates and introduce the definition of covariate shift used in the current paper.

Definition 1. *If $\mathcal{P}_X^l \neq \mathcal{Q}$ for some $1 \leq l \leq L$, we refer to this as* **covariate shift***; If $\mathcal{P}_X^l = \mathcal{Q}$ for all $1 \leq l \leq L$, we refer to this as* **no covariate shift**.

By the above definition, the covariate shift happens if any of the covariate distributions $\{\mathcal{P}_X^l\}_{1\leq l\leq L}$ differs from the targeted distribution $\mathcal{Q}$.

We now generalize the maximin effect definition in (2) to aggregate $\{b^{(l)}\}_{1\leq l\leq L}$ for a targeted population $\mathcal{Q}$. We consider the hypothetical data generation mechanism,

$$Y_1^{*,(l)} = [X_{1,\cdot}^{\mathcal{Q}}]^\intercal b^{(l)} + \epsilon_1^{(l)} \quad \text{for} \quad 1 \leq l \leq L, \tag{4}$$

where $b^{(l)}$ and $\epsilon_1^{(l)}$ are the same as those in (1) and $X_{1,\cdot}^{\mathcal{Q}} \sim \mathcal{Q}$. We use the super-index * in $Y_1^{*,(l)}$ to denote that the outcome variable is hypothetical instead of being observed. Under (4), we define the maximin effect with respect to the covariate distribution $\mathcal{Q}$ as

$$\beta^*(\mathcal{Q}) = \arg\max_{\beta\in\mathbb{R}^p} R_{\mathcal{Q}}(\beta) \quad \text{with} \quad R_{\mathcal{Q}}(\beta) = \min_{1\leq l\leq L} \left[\mathbf{E}^*(Y_{1,\cdot}^{*(l)})^2 - \mathbf{E}^*(Y_{1,\cdot}^{*(l)} - (X_{1,\cdot}^{\mathcal{Q}})^\intercal\beta)^2\right]$$

where $\mathbf{E}^*$ is the expectation taken over the joint distribution of $(Y_1^{*(l)}, X_{1,\cdot}^{\mathcal{Q}})$ defined in (4). The maximin effect $\beta^*(\mathcal{Q})$ can be interpreted from an adversarial perspective (Meinshausen and Bühlmann, 2015): in a two-side game, we select an effect vector $\beta \in \mathbb{R}^p$ and the counter agent then chooses the most challenging scenario (from one of the L groups) for this β. Our goal is to select the effect vector $\beta \in \mathbb{R}^p$ such that the worst-case reward $R(\beta)$ returned by the counter agent is maximized. With (4), we simplify the expression of $\beta^*(\mathcal{Q})$ as

$$\beta^*(\mathcal{Q}) = \arg\max_{\beta\in\mathbb{R}^p} R_{\mathcal{Q}}(\beta) \quad \text{with} \quad R_{\mathcal{Q}}(\beta) = \min_{b\in\mathbb{B}} \left[2b^\intercal\Sigma^{\mathcal{Q}}\beta - \beta^\intercal\Sigma^{\mathcal{Q}}\beta\right] \tag{5}$$

where $\mathbb{B} = \{b^{(1)}, \ldots, b^{(L)}\}$ and $\Sigma^{\mathcal{Q}} = \mathbf{E}X_{1,\cdot}^{\mathcal{Q}}(X_{1,\cdot}^{\mathcal{Q}})^\intercal$.

A few remarks are in order for the definition (5). First, if there is no covariate shift, the hypothetical outcome is the same as the observed outcome, that is, $Y_{1,\cdot}^{*(l)} = Y_{1,\cdot}$ in (4) and $\beta^*(\mathcal{Q})$ in (5) is reduced to the maximin effect in (2). Moreover, the definition of $\beta^*(\mathcal{Q})$ is useful in aggregating the regression models for a new or unseen distribution, whose covariate distribution $\mathcal{Q}$ might be different from those of $X_{i,\cdot}^{(l)}$. Second, if the model (1) is correctly specified, the regression vectors $\{b^{(l)}\}_{1\leq l\leq L}$ remain the same even in the presence of covariate shift. In contrast, the maximin effect $\beta^*(\mathcal{Q})$ changes with the target distribution $\mathcal{Q}$.

Third, the definition in (5) decouples two sources of data heterogeneity: heterogeneity among $\{b^{(1)}, \ldots, b^{(L)}\}$ and the covariate shift. With respect to $\mathcal{Q}$, the maximin effect $\beta^*(\mathcal{Q})$ itself is defined as an optimal aggregation of heterogeneous regression vectors $\{b^{(1)}, \ldots, b^{(L)}\}$. In addition, the maximin effect in (5) allows us to explore the heterogeneity in covariate distributions. After calculating $\beta^*(\mathcal{Q}_1), \ldots, \beta^*(\mathcal{Q}_J)$ for different covariate distributions

$\{\mathcal{Q}_1, \cdots, \mathcal{Q}_J\}$ with a positive integer $J > 0$, it is possible to further combine the information contained in these maximin effects; see Figure 7 for an example.

The following proposition shows how to identify the maximin effects $\beta^*(\mathcal{Q})$.

Proposition 1. *If $\lambda_{\min}(\Sigma^{\mathcal{Q}}) > 0$, then $\beta^*(\mathcal{Q})$ defined in (5) is identified as*

$$\beta^*(\mathcal{Q}) = \sum_{l=1}^{L} [\gamma^*(\mathcal{Q})]_l b^{(l)} \quad \textit{with} \quad \gamma^*(\mathcal{Q}) = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \Gamma^{\mathcal{Q}} \gamma \tag{6}$$

where $\Gamma^{\mathcal{Q}}_{lk} = (b^{(l)})^\intercal \Sigma^{\mathcal{Q}} b^{(k)}$ for $1 \leq l, k \leq L$ and $\Delta^L = \{\gamma \in \mathbb{R}^L : \gamma_j \geq 0, \ \sum_{j=1}^{L} \gamma_j = 1\}$ is the simplex over $\mathbb{R}^L$. Furthermore, $\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{B}} \left[2b^\intercal \Sigma^{\mathcal{Q}} \beta - \beta^\intercal \Sigma^{\mathcal{Q}} \beta\right] = [\beta^*(\mathcal{Q})]^\intercal \Sigma^{\mathcal{Q}} \beta^*(\mathcal{Q})$.

The above proposition shows that the covariate shift maximin effect $\beta^*(\mathcal{Q})$ can be identified as a weighted average of $\{b^{(l)}\}_{1 \leq l \leq L}$ and the aggregation weight vector $\gamma^*(\mathcal{Q})$ depends on the covariate distribution $\mathcal{Q}$. The above proposition provides an explicit way of computing $\beta^*(\mathcal{Q})$ when the regression covariance matrix $\Gamma^{\mathcal{Q}} \in \mathbb{R}^{L \times L}$ is known. Proposition 1 is implied by Theorem 1 in Meinshausen and Bühlmann (2015) and the definition of $\beta^*(\mathcal{Q})$ in (5).

Remark 1. In studying the individualized treatment effect, Shi et al. (2018) proposed the maximum projection $\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{1 \leq l \leq L} \beta^\intercal b^{(l)}$. Although the model in Shi et al. (2018) is different from our model (1), the maximin projection $\beta^{*,\mathrm{MP}}$ can be identified as a scaled version of the general maximin effect $\beta^*(\mathcal{Q})$ in (5) with the targeted covariate distribution $\mathcal{Q}$ as the identity design. See more details in Section A.1 in the supplement.

3 Ridge-type Maximin: Optimality and Stability

We define a ridge-type maximin effect to account for both reward optimality and statistical stability. The maximin effect definition in (5) is only optimizing the reward function, with no consideration of the statistical stability. Since our goal is to find a vector β to summarize the heterogeneous regression vectors $\{b^{(l)}\}_{1 \leq l \leq L}$, a better strategy is to define a vector $\beta \in \mathbb{R}^p$ such that its reward $R_{\mathcal{Q}}(\beta)$ is comparable to $R_{\mathcal{Q}}(\beta^*(\mathcal{Q}))$ and β can be estimated stably.

One challenging setting of making stable inference for $\beta^*(\mathcal{Q})$ is that the weight vector $\gamma^*(\mathcal{Q})$ in (6) is not uniquely defined. Through numerical explorations, Rothenhäusler et al. (2016) observed that, if the maximin effect is not uniquely defined, their proposed CIs are typically wide even in the low-dimensional setting with no covariate shift.

To address this, we introduce a new ridge-type maximin effect, for $\delta \geq 0$,

$$\beta^*_\delta(\mathcal{Q}) = \sum_{l=1}^{L} [\gamma^*_\delta(\mathcal{Q})]_l \cdot b^{(l)} \quad \text{with} \quad \gamma^*_\delta(\mathcal{Q}) = \arg\min_{\gamma \in \Delta^L} \left[\gamma^\intercal \Gamma^{\mathcal{Q}} \gamma + \delta \|\gamma\|_2^2\right]. \tag{7}$$

The ridge penalty $\delta\|\gamma\|_2^2$ is imposed to compute the aggregation weights and $\beta^*_{\delta=0}(\mathcal{Q})$ is reduced to $\beta^*(\mathcal{Q})$ in (5). Even if the maximin effect in (5) is not uniquely defined, the ridge-type maximin effect defined in (7) is uniquely defined for any given $\delta > 0$.

When there is no confusion from the context, we will omit $\mathcal{Q}$ in the discussion and write $\Gamma^{\mathcal{Q}}, \beta^*(\mathcal{Q}), \gamma^*(\mathcal{Q}), \beta^*_\delta(\mathcal{Q}), \gamma^*_\delta(\mathcal{Q})$ as $\Gamma, \beta^*, \gamma^*, \beta^*_\delta, \gamma^*_\delta$, respectively.

We consider the special case $L = 2$ and illustrate how the penalty level δ in (7) might affect the statistical stability. For $L = 2$, the solution of (7) can be obtained as $\gamma^*_\delta = ([\gamma^*_\delta]_1, 1 - [\gamma^*_\delta]_1)^\intercal$ with

$$[\gamma^*_\delta]_1 = \min\left\{\max\left\{\frac{\Gamma_{22} + \delta - \Gamma_{12}}{\Gamma_{11} + \Gamma_{22} + 2\delta - 2\Gamma_{12}}, 0\right\}, 1\right\}. \tag{8}$$

If $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12}$ is near zero, it will be challenging to stably estimate the weight $[\gamma^*_{\delta=0}]_1$ since even small estimation errors of $\{\Gamma_{11}, \Gamma_{22}, \Gamma_{12}\}$ might lead to a large estimation error of the weight. In contrast, it is much easier to construct a stable estimator of $[\gamma^*_\delta]_1$ for a positive $\delta > 0$. The following proposition characterizes the property of β^*_δ for this special case.

Proposition 2. *Suppose that $L = 2$ and with $n, p \to \infty$, $\max\{|\Gamma_{11} - \Gamma_{12}|, |\Gamma_{22} - \Gamma_{12}|\} \to 0$ and $\delta = \delta(n, p) \gg \max\{|\Gamma_{11} - \Gamma_{12}| + |\Gamma_{22} - \Gamma_{12}|\}$. Then β^*_δ and γ^*_δ defined in (7) satisfy $[\gamma^*_\delta]_1 \to 1/2$ and $R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*) \to 0$, with $n, p \to \infty$.*

For a positive definite $\Sigma^{\mathcal{Q}}$, $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12} = (b^{(1)} - b^{(2)})^\intercal \Sigma^{\mathcal{Q}} (b^{(1)} - b^{(2)}) \to 0$ corresponds to $\|b^{(1)} - b^{(2)}\|_2 \to 0$. Proposition 2 states that, if $b^{(1)} \approx b^{(2)}$, the ridge-type maximin effect (with $\delta > 0$) automatically aggregates $b^{(1)}$ and $b^{(2)}$ with asymptotically equal weights and the corresponding reward value is asymptotically optimal. The reward optimality is not surprising as in such a special case $b^{(1)} \approx b^{(2)}$, any linear combination leads to a nearly optimal reward. However, the improvement in terms of statistical stability can be significant. In Section 7, we consider the simulation setting 1 with $L = 2$, $p = 500$ and $\Gamma_{1,1} = 3.96$, $\Gamma_{1,2} = 3.97$ and $\Gamma_{2,2} = 4.02$ and have $\gamma^*_{\delta=0} = (1, 0)^\intercal$ and $\gamma^*_{\delta=2} = (0.53, 0.47)^\intercal$. We observe that the CIs for the ridge-type effect decreases with an increasing δ. For $\delta = 2$, the CI lengths are only half of those for $\delta = 0$; see Figure 4 in Section 7 for details.

Beyond the very special setting in Proposition 2, the penalty value δ affects the stability of our proposed estimator in more general settings. We characterize the dependence of our proposed inference method on δ in Theorems 3 and 4 and observe in both theoretical and numerical studies that our proposed procedure is more stable with an increasing penalty level $\delta > 0$; see Figures 6 and 9.

In the following, we discuss what is the effect of δ on the reward value $R_{\mathcal{Q}}(\beta^*_\delta)$. To measure the reward optimality, we introduce the ν-optimal maximin effect:

Definition 2. *A vector $\beta \in \mathbb{R}^p$ is defined to be ν-optimal maximin effect if it satisfies $R_{\mathcal{Q}}[\beta] \geq R_{\mathcal{Q}}[\beta^*(\mathcal{Q})] - \nu$ where the reward function $R_{\mathcal{Q}}$ and the maximin effect $\beta^*(\mathcal{Q})$ are defined in (5) and $\nu = \nu(n,p) > 0$ is a positive number possibly growing with n and p.*

For the ν-optimal maximin effect, the corresponding reward $R_{\mathcal{Q}}[\beta]$ might be worse off than the optimal reward $R_{\mathcal{Q}}[\beta^*(\mathcal{Q})]$ but the reduction is at most ν. The following theorem characterizes the ν-optimality for the ridge-type maximin effect in (7).

Theorem 1. *Suppose $\lambda_L(\mathcal{B}) > 0$ with $\mathcal{B} = \left(b^{(1)}, \ldots, b^{(L)}\right) \in \mathbb{R}^{p \times L}$, then the ridge-type minimizer β^*_δ defined in (7) satisfies*

$$R_{\mathcal{Q}}(\beta^*_\delta) \geq R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2) \geq R_{\mathcal{Q}}(\beta^*) - \delta \cdot (L-1)/(2L), \tag{9}$$

where $R_{\mathcal{Q}}$ and $\beta^* = \beta^*(\mathcal{Q})$ are defined in (5) and $\beta^*_\delta = \beta^*_\delta(\mathcal{Q})$ and $\gamma^*_\delta = \gamma^*_\delta(\mathcal{Q})$ are defined in (7).

Remark 2. The condition $\lambda_L(\mathcal{B}) > 0$ rules out the exact collinearity among the columns of $\mathcal{B}$ but still allows for $\lambda_L(\mathcal{B}) \to 0$ with $n, p \to \infty$, that is, the nearly collinear setting.

The above theorem shows that in terms of rewards, the ridge-type maximin effect β^*_δ is at most worse off than the regular maximin effect β^* by the amount $\nu = 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2)$. In Figure 1, we plot the reward values $R_{\mathcal{Q}}[\beta^*_\delta]$ and the lower bounds $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2)$ over $\delta \in [0,5]$ for three simulation settings detailed in Section 7. For settings 1 and 2, for $0 \leq \delta \leq 5$, $R_{\mathcal{Q}}(\beta^*_\delta)$ is above 95% of the optimal value $R_{\mathcal{Q}}(\beta^*)$. For setting 3, $R_{\mathcal{Q}}(\beta^*_\delta)$ is more sensitive to the choice of δ.

A useful implication of Theorem 1 is to quantify the reward reduction $R_{\mathcal{Q}}[\beta^*] - R_{\mathcal{Q}}[\beta^*_\delta]$ in a data dependent way. Specifically, we estimate γ^*_δ by a consistent estimator $\widehat{\gamma}_\delta$ (see (11)) and estimate an upper bound for the reward reduction by $2\delta(\|\widehat{\gamma}_\delta\|_\infty - \|\widehat{\gamma}_\delta\|_2^2)$. In Figure 1, we plot $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\widehat{\gamma}_\delta\|_\infty - \|\widehat{\gamma}_\delta\|_2^2)$. We will discuss how to choose δ in a data-dependent way in Section 7.1.

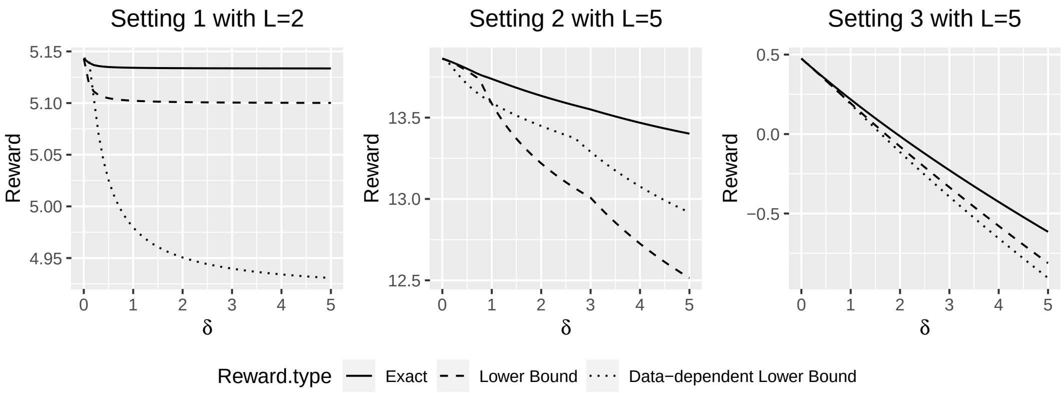


Figure 1: Dependence of reward on δ: the reward type "Exact", "Lower Bound" and "Data-dependent Lower Bound" correspond to $R_{\mathcal{Q}}(\beta^*_\delta)$, $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2)$ and $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\widehat{\gamma}_\delta\|_\infty - \|\widehat{\gamma}_\delta\|_2^2)$, respectively. The simulation settings are specified in Section 7.

4 Challenges of Inference for Maximin Effects

4.1 High-dimensional challenge: bias and variance tradeoff

We briefly review the inference procedure proposed in Rothenhäusler et al. (2016). Denote $\widetilde{b}^{(l)}$ by the ordinary least square estimator computed based on $(X^{(l)}, Y^{(l)})$. Meinshausen and Bühlmann (2015) proposed the magging estimator $\widetilde{\beta} = \sum_{l=1}^{L} \widetilde{\gamma}_l \widetilde{b}^{(l)}$ where the weight vector $\widetilde{\gamma} \in \mathbb{R}^L$ is defined as $\widetilde{\gamma} = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \widetilde{\Gamma} \gamma$ with

$$\widetilde{\Gamma}_{l,k} = [\widetilde{b}^{(l)}]^\intercal \left(\frac{1}{\sum_{l=1}^{L} n_l} \sum_{l=1}^{L} \sum_{i=1}^{n_l} X_{i\cdot}^{(l)} [X_{i\cdot}^{(l)}]^\intercal \right) \widetilde{b}^{(k)} \quad \text{for} \quad 1 \leq l, k \leq L. \tag{10}$$

In the low-dimensional setting with no covariate shift, Rothenhäusler et al. (2016) proposed valid inference procedures by establishing the asymptotic normality of the magging estimator $\widetilde{\beta}$ under certain assumptions; see Theorem 1 in Rothenhäusler et al. (2016).

The estimator $\widetilde{\Gamma}_{l,k}$ in (10) can be viewed as a plug-in estimator of $\Gamma_{l,k} = [b^{(l)}]^\intercal \Sigma b^{(k)}$, which replaces $\Sigma, b^{(l)}$ and $b^{(k)}$ by corresponding reasonable estimators. Despite its effectiveness in low dimensions, the plug-in type estimator of Γ as in (10) is in general not effective in high dimensions. We consider two state-of-the-art estimators as examples. For $p \geq n$, we might take $\widetilde{b}^{(l)}$ in (10) as the Lasso estimator (Tibshirani, 1996). The resulted weight vector $\widetilde{\gamma}$ is not suitable for further statistical inference as the plug-in estimator $\widetilde{\Gamma}_{l,k}$ in (10) inherits the bias from $\widetilde{b}^{(l)}$ and $\widetilde{b}^{(k)}$ (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Cai and Guo, 2020). As an alternative, we might take $\widetilde{b}^{(l)}$ as the coordinate bias-corrected estimator proposed by Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014). However, the plug-in debiased Lasso estimator induces both large bias and variance in estimating $\Gamma_{l,k}$. A main component $[\widetilde{b}^{(l)} - b^{(l)}]^\intercal \Sigma [\widetilde{b}^{(k)} - b^{(k)}]$ of the estimation error $\widetilde{\Gamma}_{l,k} - \Gamma_{l,k}$ is of order p/n. These issues are illustrated in a numerical study in Section D.1 in the supplement.

To address this, we shall propose a debiasing estimator in Section 5.1 to directly estimate the matrix $\Gamma^{\mathcal{Q}} \in \mathbb{R}^{L \times L}$ in the covariate shift setting. In the rest of the current section, we use $\widehat{\Gamma}^{\mathcal{Q}}$ to denote a data-dependent estimator of $\Gamma^{\mathcal{Q}}$ and define $\widehat{\gamma}_\delta \in \mathbb{R}^L$ as the ridge-type weight vector,

$$\widehat{\gamma}_\delta = \arg\min_{\gamma \in \Delta^L} \left[\gamma^\intercal \widehat{\Gamma}^{\mathcal{Q}} \gamma + \delta \|\gamma\|_2^2\right] \quad \text{for} \quad \delta \geq 0. \tag{11}$$

4.2 Maximin effect challenge: mixture distribution

Beyond high dimensionality, a further challenge is that the limiting distribution of the estimated weight vector is not necessarily normal. To illustrate this, we consider the special case with $L = 2$ and $\delta = 0$ and the optimal weight is $(\gamma_1^*, 1 - \gamma_1^*) \in \mathbb{R}^2$ where γ_1^* is defined in (8) with $\delta = 0$. We obtain an explicit solution $(\widehat{\gamma}_1, 1 - \widehat{\gamma}_1) \in \mathbb{R}^2$ of (11) with $\delta = 0$:

$$\widehat{\gamma}_1 = \min\left\{\max\left\{\bar{\gamma}_1, 0\right\}, 1\right\} \quad \text{with} \quad \bar{\gamma}_1 = \frac{\widehat{\Gamma}_{22} - \widehat{\Gamma}_{12}}{\widehat{\Gamma}_{11} + \widehat{\Gamma}_{22} - 2\widehat{\Gamma}_{12}}. \tag{12}$$

We decompose the error $\widehat{\gamma}_1 - \gamma_1^*$ as

$$(\bar{\gamma}_1 - \gamma_1^*) \cdot \mathbf{1}\{0 < \bar{\gamma}_1 < 1\} + (-\gamma_1^*) \cdot \mathbf{1}\{\bar{\gamma}_1 \leq 0\} + (1 - \gamma_1^*) \cdot \mathbf{1}\{\bar{\gamma}_1 \geq 1\}, \tag{13}$$

with $\bar{\gamma}_1$ defined in (12). For the setting $\gamma_1^* = \gamma_1^*(n,p) \asymp 1/\sqrt{n}$, even if $\sqrt{n}(\bar{\gamma}_1 - \gamma_1^*)$ is asymptotically normal, the probability of $\{\bar{\gamma}_1 \leq 0\}$ might not vanish. As a consequence, $\sqrt{n}(\widehat{\gamma}_1 - \gamma_1^*)$ will be a mixture of an asymptotic normal and a single mass at the point $-\sqrt{n}\gamma_1^*$. For a positive $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12}$, the setting $\gamma_1^* \asymp 1/\sqrt{n}$ corresponds to $\left|(b^{(1)} - b^{(2)})^\intercal \Sigma^{\mathcal{Q}} b^{(2)}\right| = c/\sqrt{n}$ for a small positive $c > 0$, that is, the difference vector $b^{(1)} - b^{(2)}$ is nearly orthogonal to $\Sigma^{\mathcal{Q}} b^{(2)}$. By symmetry, when $\left|(b^{(2)} - b^{(1)})^\intercal \Sigma^{\mathcal{Q}} b^{(1)}\right| = c/\sqrt{n}$, $\sqrt{n}(\widehat{\gamma} - \gamma^*)$ might be a mixture of an asymptotic normal distribution and a point mass at $\sqrt{n}(1 - \gamma^*)$. In Figure 2, we illustrate the mixture distribution in (13) by reporting the proportions of 0 and 1 for $\widehat{\gamma}_1$ in (12). When γ_1^* is close to 0, $\widehat{\gamma}_1 = 0$ for more than 15% of 500 simulations; when γ_1^* is close to 1, $\widehat{\gamma}_1 = 1$ for more than 60% of 500 simulations.

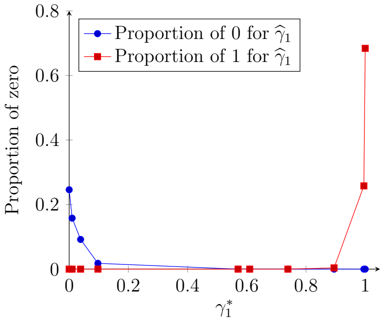


Figure 2: The proportions of 0 and 1 (out of 500 simulations) for $\widehat{\gamma}_1$ with γ_1^* changing from 0 to 1. The simulation setting is the same as that of Figure 5 in Section 7.

The mixture distribution for the weight vector in (13) has posed challenges for constructing CIs for the maximin effect relying on asymptotic normal limiting distribution. Inference procedures relying on asymptotic normality can be under-coverage; see Figure 5 and the related discussion. To address this, we shall devise a novel sampling approach in Section 5.2.

5 Inference for Maximin Effects in High Dimensions

5.1 Estimation of $\Gamma^{\mathcal{Q}}$ in the covariate shift setting

We consider the setting that we have access to a sample $\{X^{\mathcal{Q}}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}} \overset{\text{i.i.d}}{\sim} \mathcal{Q}$. For $1 \leq l \leq L$, we randomly split the data $(X^{(l)}, Y^{(l)})$ into two approximate equal-size subsamples $(X^{(l)}_{A_l,\cdot}, Y^{(l)}_{A_l})$ and $(X^{(l)}_{B_l,\cdot}, Y^{(l)}_{B_l})$, where the index sets A_l and B_l satisfy $A_l \cap B_l = \varnothing$, $A_l \cup B_l = [n_l]$ and $|A_l| = \lfloor n_l/2 \rfloor$. We randomly split the data $\{X^{\mathcal{Q}}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}$ as $X^{\mathcal{Q}}_{A,\cdot}$ and $X^{\mathcal{Q}}_{B,\cdot}$, where the index sets A and B satisfy $A \cap B = \varnothing$, $A \cup B = [N_{\mathcal{Q}}]$ and $|A| = \lfloor N_{\mathcal{Q}}/2 \rfloor$.

The proposed estimator $\widehat{\Gamma}^{\mathcal{Q}}$ is of two steps. In the first step, we estimate $\{b^{(l)}\}_{1\leq l\leq L}$ by applying Lasso (Tibshirani, 1996) to the sub-sample with the index set A_l:

$$\widehat{b}^{(l)}_{init} = \arg\min_{b\in\mathbb{R}^p} \frac{\|Y^{(l)}_{A_l} - X^{(l)}_{A_l,\cdot} b\|_2^2}{2|A_l|} + \lambda_l \sum_{j=1}^{p} \frac{\|X^{(l)}_{A_l,j}\|_2}{\sqrt{|A_l|}} |b_j|, \text{ with } \lambda_l = \sqrt{\frac{(2+c)\log p}{|A_l|}}\sigma_l \tag{14}$$

for some constant $c > 0$. As alternatives, we may use tuning-free penalized estimators, such as scaled Lasso (Sun and Zhang, 2012) or square-root Lasso (Belloni et al., 2011). We construct an initial estimator of $\Gamma^{\mathcal{Q}}_{l,k}$ as $[\widehat{b}^{(l)}_{init}]^{\intercal} \widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init}$ for $1 \leq l, k \leq L$ with $\widehat{\Sigma}^{\mathcal{Q}} = \frac{1}{|B|}\sum_{i\in B} X^{\mathcal{Q}}_{i,\cdot}(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}$.

This plug-in estimator has the error decomposition:

$$
\begin{aligned}
(\widehat{b}^{(l)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init} - (b^{(l)})^\intercal \Sigma^{\mathcal{Q}} b^{(k)} = {} & (\widehat{b}^{(k)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(l)}_{init} - b^{(l)}) + (\widehat{b}^{(l)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(k)}_{init} - b^{(k)}) \\
& - (\widehat{b}^{(l)}_{init} - b^{(l)})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(k)}_{init} - b^{(k)}) + (b^{(l)})^\intercal (\widehat{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}) b^{(k)}.
\end{aligned}
$$

In the second step, we correct the plug-in estimator by accurately estimating $(\widehat{b}^{(k)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(l)}_{init} - b^{(l)})$ and $(\widehat{b}^{(l)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(k)}_{init} - b^{(k)})$. We detail how to approximate $(\widehat{b}^{(k)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(l)}_{init} - b^{(l)})$ and a similar procedure will be proposed to approximate $(\widehat{b}^{(l)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(k)}_{init} - b^{(k)})$. With $\widehat{\Sigma}^{(l)} = \frac{1}{|B_l|} \sum_{i \in B_l} X^{(l)}_{i,\cdot} [X^{(l)}_{i,\cdot}]^\intercal$ and $\widehat{u}^{(l,k)} \in \mathbb{R}^p$ denoting a projection direction to be constructed, we approximate $(\widehat{b}^{(k)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(l)}_{init} - b^{(l)})$ by $-\frac{1}{|B_l|} [\widehat{u}^{(l,k)}]^\intercal [X^{(l)}_{B_l,\cdot}]^\intercal (Y^{(l)}_{B_l} - X^{(l)}_{B_l,\cdot} \widehat{b}^{(l)}_{init})$ and the approximation error is

$$
[\widehat{\Sigma}^{(l)} \widehat{u}^{(l,k)} - \widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init}]^\intercal (\widehat{b}^{(l)}_{init} - b^{(l)}) - \frac{1}{|B_l|} [\widehat{u}^{(l,k)}]^\intercal [X^{(l)}_{B_l,\cdot}]^\intercal \epsilon^{(l)}_{B_l}. \tag{15}
$$

We construct $\widehat{u}^{(l,k)}$ as follows,

$$
\widehat{u}^{(l,k)} = \operatorname*{arg\,min}_{u \in \mathbb{R}^p} u^\intercal \widehat{\Sigma}^{(l)} u \quad \text{subject to } \|\widehat{\Sigma}^{(l)} u - \omega^{(k)}\|_\infty \le \|\omega^{(k)}\|_2 \mu_l \tag{16}
$$

$$
\left| [\omega^{(k)}]^\intercal \widehat{\Sigma}^{(l)} u - \|\omega^{(k)}\|_2^2 \right| \le \|\omega^{(k)}\|_2^2 \mu_l \tag{17}
$$

where $\mu_l \asymp \sqrt{\log p / |B_l|}$ and

$$
\omega^{(k)} = \widetilde{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init} \in \mathbb{R}^p \quad \text{with} \quad \widetilde{\Sigma}^{\mathcal{Q}} = \frac{1}{|A|} \sum_{i \in A} X^{\mathcal{Q}}_{i,\cdot} (X^{\mathcal{Q}}_{i,\cdot})^\intercal. \tag{18}
$$

We provide intuitions on how the constraints in (16) and (17) ensure a small approximation error in (15). The objective $u^\intercal \widehat{\Sigma}^{(l)} u$ in (16) is proportional to the variance of the second term in (15). The constraint set in (16) implies $\widehat{\Sigma}^{(l)} \widehat{u}^{(l,k)} - \widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init} \approx \widehat{\Sigma}^{(l)} \widehat{u}^{(l,k)} - \omega^{(k)} \approx \mathbf{0}$, which guarantees the first term of (15) to be small. The additional constraint (17) is seemingly useless to control the approximation error in (15). However, this additional constraint ensures that the second term in (15) dominates the first term in (15), which is critical in constructing an asymptotically normal estimator of $\Gamma^{\mathcal{Q}}_{l,k}$. The additional constraint (17) is particularly useful in the covariate shift setting, that is, $\Sigma^{(l)} \neq \Sigma^{\mathcal{Q}}$ for some $1 \le l \le L$. Finally, we construct the bias-corrected estimator of $\Gamma^{\mathcal{Q}}_{l,k}$ as

$$
\widehat{\Gamma}^{\mathcal{Q}}_{l,k} = (\widehat{b}^{(l)}_{init})^\intercal \widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init} + [\widehat{u}^{(l,k)}]^\intercal \frac{1}{|B_l|} [X^{(l)}_{B_l,\cdot}]^\intercal (Y^{(l)}_{B_l} - X^{(l)}_{B_l,\cdot} \widehat{b}^{(l)}_{init}) + [\widehat{u}^{(k,l)}]^\intercal \frac{1}{|B_k|} [X^{(k)}_{B_k,\cdot}]^\intercal (Y^{(k)}_{B_k} - X^{(k)}_{B_k,\cdot} \widehat{b}^{(k)}_{init}) \tag{19}
$$

where $\widehat{u}^{(l,k)}$ and $\widehat{u}^{(k,l)}$ are defined in the optimization algorithm (16) and (17). Then we estimate the maximin weight vector by $\widehat{\gamma}_\delta$ defined in (11).

If $\Sigma^{\mathbb{Q}}$ is known, we modify $\widehat{\Gamma}^{\mathbb{Q}}_{l,k}$ in (19) by replacing $\widehat{\Sigma}^{\mathbb{Q}}$ by $\Sigma^{\mathbb{Q}}$ and $\omega^{(k)}$ in (18) by $\omega^{(k)} = \Sigma^{\mathbb{Q}}\widehat{b}^{(k)}_{init}$. This modified estimator (with known $\Sigma^{\mathbb{Q}}$) is of a smaller variance as there is no uncertainty of estimating $\Sigma^{\mathbb{Q}}$; see Figure 6 for numerical comparisons.

A few simplifications can be made when there is no covariate shift. For $1 \leq l \leq L$, we estimate $b^{(l)}$ by applying Lasso to the whole data set $(X^{(l)}, Y^{(l)})$: $\widehat{b}^{(l)}_{init} = \arg\min_{b\in\mathbb{R}^p} \|Y^{(l)} - X^{(l)}b\|_2^2/(2n_l) + \lambda_l \sum_{j=1}^{p} \|X^{(l)}_j\|_2/\sqrt{n_l} \cdot |b_j|$ with $\lambda = \sqrt{(2+c)\log p/n_l}\sigma_l$ for some constant $c > 0$. We slightly abuse the notation by using $\widehat{b}^{(l)}_{init}$ to denote the Lasso estimator based on the non-split data. Since $\Sigma^{(l)} = \Sigma^{\mathbb{Q}}$ for $1 \leq l \leq L$, we define

$$\widehat{\Sigma} = \frac{1}{\sum_{l=1}^{L} n_l + N_{\mathbb{Q}}} \left(\sum_{l=1}^{L} \sum_{i=1}^{n_l} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^\intercal + \sum_{i=1}^{N_{\mathbb{Q}}} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^\intercal \right)$$

and estimate $\Gamma_{l,k}$ as

$$\widehat{\Gamma}^{\mathbb{Q}}_{l,k} = (\widehat{b}^{(l)}_{init})^\intercal \widehat{\Sigma}\widehat{b}^{(k)}_{init} + (\widehat{b}^{(l)}_{init})^\intercal \frac{1}{n_k}[X^{(k)}]^\intercal(Y^{(k)} - X^{(k)}\widehat{b}^{(k)}_{init}) + (\widehat{b}^{(k)}_{init})^\intercal \frac{1}{n_l}[X^{(l)}]^\intercal(Y^{(l)} - X^{(l)}\widehat{b}^{(l)}_{init}). \quad (20)$$

This estimator can be viewed as a special case of (19) by taking $\widehat{u}^{(l,k)}$ and $\widehat{u}^{(k,l)}$ as $\widehat{b}^{(k)}_{init}$ and $\widehat{b}^{(l)}_{init}$, respectively. Neither the optimization in (16) and (17) nor the sample splitting is needed for constructing the debiased estimator in the no covariate shift setting.

Remark 3. Verzelen and Gassiat (2018); Cai and Guo (2020); Guo et al. (2019) considered inference for quadratic functionals in a single high-dimensional linear model while Guo et al. (2019) proposed debiased estimators of $[b^{(j)}]^\intercal b^{(k)}$ for $1 \leq j, k \leq L$. The main challenge in the covariate shift setting is that $\Sigma^{(l)} \neq \Sigma^{\mathbb{Q}}$ for some $1 \leq l \leq L$ and $\omega^{(k)} = \widetilde{\Sigma}^{\mathbb{Q}}\widehat{b}^{(k)}_{init} \in \mathbb{R}^p$ can be an arbitrarily dense vector. To address this, the novel projection direction is constructed in (16) and (17). Importantly, the focus of the current paper is not on inference for entries of $\Gamma^{\mathbb{Q}}$ as in Verzelen and Gassiat (2018); Guo et al. (2019); Cai and Guo (2020); Guo et al. (2019), but to use $\widehat{\Gamma}^{\mathbb{Q}} \in \mathbb{R}^{L\times L}$ as an initial estimator for the sampling procedure in Section 5.2, which requires quantification of correlations among the entries of $\widehat{\Gamma}^{\mathbb{Q}}$.

5.2 A sampling approach for the maximin effect inference

We propose a sampling procedure to address the mixture distribution of the maximin effects discussed in Section 4.2. For our proposed estimator $\widehat{\Gamma}^{\mathbb{Q}}$ in (19), we will show in Theorem

2 that the stacked long vector $\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}}) \in \mathbb{R}^{L(L+1)/2}$ can be approximated by a multivariate Gaussian random vector $S^* \in \mathbb{R}^{L(L+1)/2}$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}$. The index mapping π in (3) maps the matrix indexes $(l_1, k_1), (l_2, k_2) \in \mathcal{I}_L$ to the corresponding stacked vector index $\pi(l_1, k_1), \pi(l_2, k_2) \in [L(L+1)/2]$. The term $\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ approximates the covariance between $\widehat{\Gamma}^{\mathcal{Q}}_{l_1,k_1} - \Gamma^{\mathcal{Q}}_{l_1,k_1}$ and $\widehat{\Gamma}^{\mathcal{Q}}_{l_2,k_2} - \Gamma^{\mathcal{Q}}_{l_2,k_2}$. We estimate $\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ by $\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined as

$$\begin{aligned}
&\frac{\widehat{\sigma}^2_{l_1}}{|B_{l_1}|}(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = l_1)\right] \\
&\quad + \frac{\widehat{\sigma}^2_{k_1}}{|B_{k_1}|}(\widehat{u}^{(k_1,l_1)})^{\intercal}\widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = k_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = k_1)\right] \\
&+\frac{1}{|B|N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left((\widehat{b}^{(l_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (\widehat{b}^{(l_1)}_{init})^{\intercal}\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^{\intercal}\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_2)}_{init}\right)
\end{aligned} \tag{21}$$

where $\bar{\Sigma}^{\mathcal{Q}} = \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} X^{\mathcal{Q}}_{i,\cdot}(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}$ and $\widehat{\sigma}^2_l = \|Y^{(l)} - X^{(l)}\widehat{b}^{(l)}\|_2^2/n_l$ for $1 \leq l \leq L$.

With the estimated $\widehat{\mathbf{V}}$, we sample $S^{[m]} \in \mathbb{R}^{L(L+1)/2}$ for $1 \leq m \leq M$ as

$$S^{[m]} \sim \mathcal{N}\left(\mathbf{0}, \widehat{\mathbf{V}} + d_0/n \cdot \mathbf{I}\right) \quad \text{with} \quad d_0 = \max\left\{\tau \cdot \max_{(l,k)\in\mathcal{I}_L}\left\{n \cdot \widehat{\mathbf{V}}_{\pi(l,k),\pi(l,k)}\right\}, 1\right\}, \tag{22}$$

where $\tau > 0$ is a positive tuning parameter (default set as 0.2) and $\mathbf{I}$ is the identity matrix of the same dimension as $\widehat{\mathbf{V}}$. The rescaled diagonal entry $n \cdot \widehat{\mathbf{V}}_{\pi(l,k),\pi(l,k)}$ will be shown in Proposition 3 to be of a constant order under regularity conditions and hence d_0 is a constant depending on the maximum of $n \cdot \widehat{\mathbf{V}}$ and a pre-specified constant τ. The sampling distribution in (22) is slightly noisier than that with $\widehat{\mathbf{V}}$ and the enlargement in the generating covariance is $d_0/n \cdot \mathbf{I}$. This slightly enlarged covariance in the sampling distribution is useful from two perspectives: firstly, the covariance matrix $\widehat{\mathbf{V}} + d_0/n \cdot \mathbf{I}$ is positive-definite even if $\widehat{\mathbf{V}}$ is nearly singular; secondly, for the setting of nearly null $b^{(k)}$ and $b^{(l)}$, $\widehat{\mathbf{V}}_{\pi(l,k),\pi(l,k)}$ might not accurately quantify the uncertainty of $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ since the remaining bias of $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ in (19) may be the dominating term. Even in this setting, the enlarged variance $\widehat{\mathbf{V}}_{\pi(l,k),\pi(l,k)} + d_0/n$ provides an upper bound for the uncertainty of our proposed $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$.

For samples $\{S^{[m]}\}_{1\leq m\leq M}$ generated in (22), we construct $\widehat{\Gamma}^{[m]} \in \mathbb{R}^{L\times L}$ as

$$\widehat{\Gamma}^{[m]}_{l,k} = \begin{cases} \widehat{\Gamma}^{\mathcal{Q}}_{l,k} - S^{[m]}_{\pi(l,k)} & \text{if} \quad 1 \leq l \leq k \leq L \\ \widehat{\Gamma}^{[m]}_{k,l} & \text{otherwise} \end{cases} \tag{23}$$

where the index mapping π is defined in (3). The lower triangle part of $\widehat{\Gamma}^{[m]}$ is constructed with the sampled $S^{[m]}$ while the upper triangle part is constructed by symmetry.

For $1 \leq m \leq M$, the sampled weight vector $\widehat{\gamma}_{\delta}^{[m]}$ is constructed as

$$\widehat{\gamma}_{\delta}^{[m]} = \underset{\gamma \in \Delta^L}{\arg\min} \left[\gamma^{\intercal} \widehat{\Gamma}_{+}^{[m]} \gamma + \delta \|\gamma\|_2^2\right] \quad \text{for} \quad \delta \geq 0. \tag{24}$$

We provide some intuitions on this sampling step. By (23), we can write

$$\mathrm{vecl}(\widehat{\Gamma}^{[m]}) \overset{d}{\approx} \mathrm{vecl}(\Gamma^{\mathcal{Q}}) + S^* - S^{[m]} \quad \text{for } 1 \leq m \leq M,$$

where $\overset{d}{\approx}$ indicates approximately equal in distribution. With a large sampling number M, there exists $1 \leq m^* \leq M$ such that $S^{[m^*]}$ is sufficiently close to S^*. As a consequence, $\widehat{\Gamma}^{[m^*]}$ almost recovers the true regression covariance matrix $\Gamma^{\mathcal{Q}}$; see Theorem 3.

5.3 CI construction for $x_{\mathrm{new}}^{\intercal}\beta_{\delta}^*$

For the group l with $1 \leq l \leq L$, we adopt the existing debiased estimator of $x_{\mathrm{new}}^{\intercal} b^{(l)}$ in Cai et al. (2019),

$$\widehat{x_{\mathrm{new}}^{\intercal} b^{(l)}} = x_{\mathrm{new}}^{\intercal} \widehat{b}_{init}^{(l)} + [\widehat{v}^{(l)}]^{\intercal} \frac{1}{n_l} (X^{(l)})^{\intercal} (Y^{(l)} - X^{(l)} \widehat{b}_{init}^{(l)}) \tag{25}$$

where $\widehat{b}_{init}^{(l)}$ is the Lasso estimator on $(X^{(l)}, Y^{(l)})$ and $\widehat{v}^{(l)} \in \mathbb{R}^p$ is defined as

$$\widehat{v}^{(l)} = \underset{v \in \mathbb{R}^p}{\arg\min}\, v^{\intercal} \frac{1}{n_l} (X^{(l)})^{\intercal} X^{(l)} v \quad \text{s.t.} \quad \max_{w \in \mathcal{C}(x_{\mathrm{new}})} \left| \langle w, \frac{1}{n_l} (X^{(l)})^{\intercal} X^{(l)} v - x_{\mathrm{new}} \rangle \right| \leq \eta_l \tag{26}$$

with $\mathcal{C}(x_{\mathrm{new}}) = \{e_1, e_2, \cdots, e_p, x_{\mathrm{new}}/\|x_{\mathrm{new}}\|_2\}$ and $\eta_l \asymp \|x_{\mathrm{new}}\|_2 \sqrt{\log p / n_l}$. The main idea of (25) is to correct the bias of $x_{\mathrm{new}}^{\intercal} \widehat{b}_{init}^{(l)}$ and if x_{new} is taken as the j-th Euclidean bias, this estimator is reduced to the debiased estimator of β_j in Zhang and Zhang (2014); Javanmard and Montanari (2014); see more detailed discussions in Cai et al. (2019).

We study inference for $x_{\mathrm{new}}^{\intercal}\beta_{\delta}^*$, which is a weighted average of $x_{\mathrm{new}}^{\intercal} b^{(1)}, \cdots, x_{\mathrm{new}}^{\intercal} b^{(L)}$. We aggregate $\{\widehat{x_{\mathrm{new}}^{\intercal} b^{(l)}}\}_{1 \leq l \leq L}$ with the sampled weight $\widehat{\gamma}_{\delta}^{[m]}$ in (24):

$$\widehat{x_{\mathrm{new}}^{\intercal} \beta}^{[m]} = \sum_{l=1}^{L} [\widehat{\gamma}_{\delta}^{[m]}]_l \cdot \widehat{x_{\mathrm{new}}^{\intercal} b^{(l)}} \quad \text{for} \quad 1 \leq m \leq M. \tag{27}$$

For $1 \leq m \leq M$, we construct the sampled interval centering at $\widehat{x_{\rm new}^{\intercal}\beta}^{[m]}$,

$$\mathrm{Int}_{\alpha}^{[m]}(x_{\rm new}) = \left(\widehat{x_{\rm new}^{\intercal}\beta}^{[m]} - z_{1-\alpha/2}\widehat{\rm se}^{[m]}(x_{\rm new}), \widehat{x_{\rm new}^{\intercal}\beta}^{[m]} + z_{1-\alpha/2}\widehat{\rm se}^{[m]}(x_{\rm new})\right)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and

$$\widehat{\rm se}^{[m]}(x_{\rm new}) = 1.01\sqrt{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}} \quad \text{with } \widehat{v}^{(l)} \text{ defined in (26).}$$

Here, the estimated standard error is slightly enlarged by a factor of 1.01 to offset the finite sample bias. Then we propose the CI for $x_{\rm new}^{\intercal}\beta_{\delta}^{*}$ as

$$\mathrm{CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right) = \cup_{m=1}^{M}\mathrm{Int}_{\alpha}^{[m]}(x_{\rm new}). \tag{28}$$

We refer to this proposed CI as Sampling Aggregation for Ridge-type maximin effects, shorthanded as SAR. A few remarks are in order. Firstly, the intuition of (28) is that there exists $1 \leq m^{*} \leq M$ such that $\widehat{\gamma}_{\delta}^{[m^{*}]}$ is fairly close to the truth γ_{δ}^{*} and the uncertainty of the sampled point estimator $\widehat{x_{\rm new}^{\intercal}\beta}^{[m^{*}]}$ mainly comes out of $\{\widehat{x_{\rm new}^{\intercal}b^{(l)}}\}_{1\leq l\leq L}$ instead of the weight estimator. Not every sampled interval $\mathrm{Int}_{\alpha}^{[m]}(x_{\rm new})$ covers $x_{\rm new}^{\intercal}\beta_{\delta}^{*}$ properly since the uncertainty of $\widehat{\gamma}_{\delta}^{[m]}$ is not quantified in the construction of $\mathrm{Int}_{\alpha}^{[m]}(x_{\rm new})$.

Secondly, although the CI in (28) is a union of M intervals, its length is still well controlled, mainly due to the fact that the centers of the intervals $\{\mathrm{Int}_{\alpha}^{[m]}(x_{\rm new})\}_{1\leq m\leq M}$ are close to $x_{\rm new}^{\intercal}\beta_{\delta}^{*}$ and standard errors $\widehat{\rm se}^{[m]}(x_{\rm new})$ are of order $\|x_{\rm new}\|_2/\sqrt{n}$; see (37) in Theorem 4 and the "Efficiency Ratio" plots in Section 7.

Thirdly, the proposed method is computationally efficient, in the sense that, the samples $S^{[m]}$ are generated after implementing high-dimensional optimization. After sampling each $S^{[m]}$, we mainly solve a L-dimension optimization problem in (24). This significantly reduces the computation cost in comparison to sampling from the original data and conducting the high-dimensional optimization problems for each sampled data.

As a next step, we can extend the inference procedure in (28) to test the null hypothesis of $x_{\rm new}^{\intercal}\beta^{*}$ by $\phi_{\alpha} = \mathbf{1}\left(0 \not\in \mathrm{CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right)\right)$. If $x_{\rm new}$ is taken as the j-th Euclidean basis, then this is to test the significance of the j-th variable.

Algorithm 1 summarizes the construction of $\mathrm{CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right)$, with the corresponding tuning parameters selection presented in Section 7.1.

Algorithm 1 Sampling Aggregation for Ridge-type maximin effects (SAR)

Input: Data $\{X^{(l)}, Y^{(l)}\}_{1\leq l\leq L}$, $X^{\mathcal{Q}}$; loading $x_{\rm new} \in \mathbb{R}^p$; level $\alpha \in (0,1)$; sampling size M; tuning parameters $\delta \geq 0$, $\tau > 0$, $\mu_l > 0$, $\lambda_l > 0$ and $\eta_l > 0$ for $1 \leq l \leq L$.

Output: Confidence interval $\mathrm{CI}_\alpha\left(x_{\rm new}^\intercal \beta^*_\delta\right)$; point estimator $\widehat{x_{\rm new}^\intercal \beta^*_\delta}$.

```
for $l \leftarrow 1$ to $L$ do
    Compute $\widehat{b}^{(l)}$ in (14) with $\lambda_l > 0$ and $\widehat{\sigma}_l^2 = \|Y^{(l)} - X^{(l)}\widehat{b}^{(l)}\|_2^2/n_l$;
    Compute $\widehat{v}^{(l)}$ in (26) with $\eta_l > 0$ and $\widehat{x_{\rm new}^\intercal b^{(l)}}$ in (25);
end for                                                    ▷ Construction of Initial Estimators
for $(l,k) \leftarrow \mathcal{I}_L = \{(l,k) : 1 \leq k \leq l \leq L\}$ do
    Compute $\widehat{u}^{(l,k)}$ in (16),(17) with $\mu_l > 0$;
    Compute $\widehat{u}^{(k,l)}$ in (16),(17) with $\mu_k > 0$;
    Compute $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ in (19);
end for                                                    ▷ Estimation of $\Gamma^{\mathcal{Q}}$
Compute $\widehat{\gamma}_\delta$ in (11) with $\delta \geq 0$;
Compute $\widehat{x_{\rm new}^\intercal \beta^*_\delta} = \sum_{l=1}^{L} [\widehat{\gamma}_\delta]_l \cdot \widehat{x_{\rm new}^\intercal b^{(l)}}$;      ▷ Point estimation
for $(l_1,k_1),(l_2,k_2) \leftarrow \mathcal{I}_L$ do
    Compute $\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ in (21);
end for                                                    ▷ Uncertainty quantification of $\widehat{\Gamma}^{\mathcal{Q}}$
for $m \leftarrow 1,2,\ldots,M$ do
    Sample $\widehat{\Gamma}^{[m]}$ in (22), (23) with $\widehat{\mathbf{V}}$ and $\tau > 0$;
    Compute $\widehat{\gamma}_\delta^{[m]}$ in (24) with $\delta \geq 0$;
    Construct the interval $\mathrm{Int}_\alpha^{[m]}(x_{\rm new})$ with $\widehat{\gamma}_\delta^{[m]}$;
end for                                                    ▷ Sampling for Ridge-type Maximin
Construct $\mathrm{CI}_\alpha\left(x_{\rm new}^\intercal \beta^*_\delta\right)$ in (28).    ▷ Confidence Interval Aggregation
```

6 Theoretical Justification

6.1 Properties of $\widehat{\Gamma}^{\mathcal{Q}}$

Before stating the main results, we introduce the assumptions for the model (1).

(A1) For $1 \leq l \leq L$, the data $\{X^{(l)}_{i,\cdot}, \epsilon^{(l)}_i\}_{1\leq i\leq n_l}$ are independently and identically generated, where $\epsilon^{(l)}_1$ is Gaussian with mean 0 and covariance σ_l^2 and independent of $X^{(l)}_{1,\cdot}$, and $X_{1,\cdot}^{(l)} \in \mathbb{R}^p$ is sub-gaussian with $\Sigma^{(l)} = \mathbf{E} X_{1,\cdot}^{(l)} [X_{1,\cdot}^{(l)}]^{\intercal}$ satisfying $c_0 \leq \lambda_{\min}\left(\Sigma^{(l)}\right) \leq \lambda_{\max}\left(\Sigma^{(l)}\right) \leq C_0$ for positive constants $C_0 > c_0 > 0$. The data $\{X_{i,\cdot}^{\mathcal{Q}}\}_{1\leq i \leq N_{\mathcal{Q}}} \overset{i.i.d.}{\sim} \mathcal{Q}$ and $X_{1,\cdot}^{\mathcal{Q}}$ is sub-gaussian with $\Sigma^{\mathcal{Q}} = \mathbf{E} X_{1,\cdot}^{\mathcal{Q}} [X_{1,\cdot}^{\mathcal{Q}}]^{\intercal}$ satisfying $c_1 \leq \lambda_{\min}\left(\Sigma^{\mathcal{Q}}\right) \leq \lambda_{\max}\left(\Sigma^{\mathcal{Q}}\right) \leq C_1$ for positive constants $C_1 > c_1 > 0$. L is finite and $\max_{1\leq l \leq L} \|b^{(l)}\|_2 \leq C$ for a positive constant $C > 0$.

(A2) Define $s = \max_{1\leq l\leq L} \|b^{(l)}\|_0$ and $n = \min_{1\leq l \leq L} n_l$. $n \asymp \max_{1\leq l\leq L} n_l$ and the model complexity parameters (s, n, p) satisfy $(s \log p)^2/n \to 0$.

Assumption (A1) is commonly assumed for the theoretical analysis of high-dimensional linear models; c.f. Bühlmann and van de Geer (2011). The Gaussian error $\epsilon_1^{(l)}$ assumption can be relaxed to sub-gaussian with a more refined analysis (Javanmard and Montanari, 2014; Cai et al., 2019). The positive definite $\Sigma^{(l)}$ guarantees the restricted eigenvalue condition with a high probability (Bickel et al., 2009; Zhou, 2009). The model complexity condition $n \gg (s \log p)^2$ in (A2) is commonly assumed in the CI construction for high-dimensional linear models (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014) and has been shown in Cai and Guo (2017) as the minimum sample size requirement for constructing adaptive CIs for single regression coefficients. The boundedness assumptions on L and $\|b^{(l)}\|_2$ are mainly imposed to simplify the presentation, so is the assumption $n \asymp \max_{1\leq l\leq L} n_l$. The boundedness assumption on $\|b^{(l)}\|_2$ can be implied by a finite second order moment of the outcome variable and a positive definite $\Sigma^{(l)}$. In our analysis, we shall point out the dependence on L, $\|b^{(l)}\|_2$ and n_l whenever possible. Define $\mathbf{V} = (\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)})_{(l_1,k_1)\in\mathcal{I}_L,(l_2,k_2)\in\mathcal{I}_L}$ as

$$\begin{aligned}\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)} = & \frac{\sigma_{l_1}^2}{|B_{l_1}|} (\widehat{u}^{(l_1,k_1)})^{\intercal} \widehat{\Sigma}^{(l_1)} \left[\widehat{u}^{(l_2,k_2)} \mathbf{1}(l_2 = l_1) + \widehat{u}^{(k_2,l_2)} \mathbf{1}(k_2 = l_1)\right] \\ & + \frac{\sigma_{k_1}^2}{|B_{k_1}|} (\widehat{u}^{(k_1,l_1)})^{\intercal} \widehat{\Sigma}^{(k_1)} \left[\widehat{u}^{(l_2,k_2)} \mathbf{1}(l_2 = k_1) + \widehat{u}^{(k_2,l_2)} \mathbf{1}(k_2 = k_1)\right] \\ & \frac{1}{|B|} (\mathbf{E}[b^{(l_1)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}} [b^{(k_1)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}} [b^{(l_2)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}} [b^{(k_2)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}} - (b^{(l_1)})^{\intercal} \Sigma^{\mathcal{Q}} b^{(k_1)} (b^{(l_2)})^{\intercal} \Sigma^{\mathcal{Q}} b^{(k_2)})\end{aligned} \tag{29}$$

where the index mapping π and the matrix index set $\mathcal{I}_L$ are defined in (3).

The following theorem shows that our proposed estimator $\text{vecl}(\widehat{\Gamma}^{\mathcal{Q}})$ can be approximately by a multivariate Gaussian random vector with zero mean and covariance $\mathbf{V}$.

Theorem 2. *Consider the model* (1). *Suppose Condition* (A1) *holds and* $\frac{s \log p}{\min\{n, N_{\mathcal{Q}}\}} \to 0$ *with* $n = \min_{1\leq l\leq L} n_l$ *and* $s = \max_{1\leq l \leq L} \|b^{(l)}\|_0$. *Then the proposed estimator* $\widehat{\Gamma}^{\mathcal{Q}} \in \mathbb{R}^{L\times L}$ *in* (19) *satisfies* $\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}} = D + \mathrm{Rem}$, *where there exists* $D^* \in \mathbb{R}^{L\times L}$ *and* $S^* \in \mathbb{R}^{L(L+1)/2}$ *such that* $D^* \overset{d}{=} D$, $\|\mathrm{vecl}(D^*) - S^*\|_2 = o(N_{\mathcal{Q}}^{-2/3})$ *almost surely, and*

$$S^* \mid X^{\mathcal{Q}}_{A,\cdot}, \{X^{(l)}, \epsilon^{(l)}_{A_l}\}_{1\leq l\leq L} \sim \mathcal{N}(0, \mathbf{V})$$

with the covariance matrix $\mathbf{V}$ *defined in* (29)*; for* $1 \leq l,k \leq L$, *the reminder term* $\mathrm{Rem}_{l,k}$ *satisfies, with probability larger than* $1-\min\{n,p\}^{-c}$,

$$|\mathrm{Rem}_{l,k}| \lesssim (1 + \|\omega^{(k)}\|_2 + \|\omega^{(l)}\|_2)\frac{s\log p}{n} + (\|b^{(k)}\|_2 + \|b^{(l)}\|_2)\sqrt{\frac{s(\log p)^2}{nN_{\mathcal{Q}}}}, \tag{30}$$

where $c > 0$ *is a positive constant and* $\omega^{(k)}$ *and* $\omega^{(l)}$ *are defined in* (18)*.*

The above theorem shows that $\mathrm{vecl}(\widehat{\Sigma}^{\mathcal{Q}}) - \mathrm{vecl}(\Sigma^{\mathcal{Q}})$ can be approximated in distribution by S^* and the approximation error decomposes of two parts: the reminder terms $\mathrm{Rem}_{l,k}$ controlled in (30) and the distribution approximation error $\mathrm{vecl}(D^*) - S^*$ of the order $N_{\mathcal{Q}}^{-2/3}$. We control the diagonal of the covariance matrix $\mathbf{V}$ in the following proposition.

Proposition 3. *Suppose that the assumptions of Theorem 2 hold. Then with probability larger than* $1-\min\{n,p\}^{-c}$*, the diagonal element* $\mathbf{V}_{\pi(l,k),\pi(l,k)}$ *in* (29) *for* $(l,k) \in \mathcal{I}_L$ *satisfies,*

$$\frac{\|\omega^{(l)}\|_2^2}{n_k} + \frac{\|\omega^{(k)}\|_2^2}{n_l} \lesssim \mathbf{V}_{\pi(l,k),\pi(l,k)} \lesssim \frac{\|\omega^{(l)}\|_2^2}{n_k} + \frac{\|\omega^{(k)}\|_2^2}{n_l} + \frac{\|b^{(l)}\|_2^2\|b^{(k)}\|_2^2}{N_{\mathcal{Q}}} \tag{31}$$

where $c > 0$ *is a positive constant and* $\omega^{(l)}$ *and* $\omega^{(k)}$ *are defined in* (18)*. If* $\Sigma^{\mathcal{Q}}$ *is known, then with probability larger than* $1-\min\{n,p\}^{-c}$, $n\cdot \mathbf{V}_{\pi(l,k),\pi(l,k)} \lesssim \|b^{(k)}\|_2^2 + \|b^{(l)}\|_2^2 + s\log p/n$*. If* $\Sigma^{\mathcal{Q}}$ *is unknown, then with probability larger than* $1-\min\{n,p\}^{-c}$,

$$n\cdot \mathbf{V}_{\pi(l,k),\pi(l,k)} \lesssim \left(1 + \frac{p}{N_{\mathcal{Q}}}\right)^2 \left(\|b^{(k)}\|_2^2 + \|b^{(l)}\|_2^2 + s\frac{\log p}{n}\right) + \frac{n}{N_{\mathcal{Q}}}\|b^{(l)}\|_2^2\|b^{(k)}\|_2^2. \tag{32}$$

The above proposition controls the variance $\mathbf{V}_{\pi(l,k),\pi(l,k)}$ of $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ for the covariate shift setting. For known $\Sigma^{\mathcal{Q}}$, we show that the diagonal elements of $\mathbf{V}$ is of order $1/n$ if $\max_{1\leq l\leq L}\|b^{(l)}\|_2$ is bounded and $s \lesssim n/\log p$. If the matrix $\Sigma^{\mathcal{Q}}$ is estimated from the data, then (32) shows that the diagonal elements of $\mathbf{V}$ is of order $1/n$ under the additional assumption $N_{\mathcal{Q}} \gtrsim \max\{p,n\}$. This requires a relatively large sample size $N_{\mathcal{Q}}$ of the unlabelled covariate data for the targeted population while the sizes of the labelled data $\{n_l\}_{1\leq l\leq L}$ are allowed to be much smaller than p. The condition $N_{\mathcal{Q}} \gtrsim \max\{p,n\}$ is mainly imposed for bounding the diagonal entries of $n \cdot \mathbf{V}$ but not for establishing the asymptotic normality in Theorem 2. We have explored the dependence of the finite sample performance on N; see Figure 6 for more details.

Now we combine Theorem 2 and Proposition 3 to discuss when the approximation errors in Theorem 2 are negligible in comparison to the sampling uncertainty in (22). Specifically, if (A2) holds and

$$N_{\mathcal{Q}} \gg \max\{s(\log p)^2, n^{3/4}, \sqrt{ns[\log \max\{N_{\mathcal{Q}}, p\}]^3}\}, \tag{33}$$

then $\sqrt{n}\|\mathrm{Rem}\|_\infty / d_0 \xrightarrow{p} 0$ and $\sqrt{n}\|\mathrm{vecl}(D^*) - S^*\|_\infty / d_0 \xrightarrow{p} 0$ with d_0 defined in (22), that is, the approximation errors in Theorem 2 are negligible. If $N_{\mathcal{Q}} \gtrsim n$, (33) is reduced to $N_{\mathcal{Q}} \gg s[\log \max\{N_{\mathcal{Q}}, p\}]^3$. The sample size condition on $N_{\mathcal{Q}}$ is mild as the amount of unlabelled data can be large in practical applications such as electronic health record data and genome-wide association study.

For the setting with no covariate shift, we establish theoretical results for the estimator in (21), which are similar to Theorem 2 and Proposition 3. The details are presented in Section A.2 in the supplement.

6.2 Sampling accuracy

We justify the sampling step in Section 5.2. For $L > 0$ and $\alpha_0 \in (0, 1/2)$, define

$$C^*(L, \alpha_0) = \mathrm{Vol}(L(L+1)/2) \cdot \frac{1}{2\sqrt{2\pi}} \prod_{i=1}^{\frac{L(L+1)}{2}} [n \cdot \lambda_i(\mathbf{V}) + 3d_0/2]^{-1} \exp\left(-F^{-1}_{\chi^2_r}(1-\alpha_0)\right) \tag{34}$$

where $\mathrm{Vol}(L(L+1)/2)$ denotes the volume of a unit ball in $L(L+1)/2$ dimensions, $1 \le r \le L(L+1)/2$ is the rank of $\mathbf{V}$ defined in (29) and $F^{-1}_{\chi^2_r}(1-\alpha_0)$ denotes the $1-\alpha_0$ quantile of the χ^2 distribution with degree of freedom r.

Theorem 3. *Consider the model* (1)*. Suppose Conditions* (A1), (A2) *and* (33) *hold. For* $\alpha_0 \in (0, 1/2)$*, define the sampling accuracy as* $\mathrm{err}_n(M) = \left[\frac{2\log n}{C^*(L,\alpha_0)M}\right]^{\frac{2}{L(L+1)}}$ *with* $C^*(L, \alpha_0)$ *in* (34)*. If* $\mathrm{err}_n(M) \le c\min\{1, \sqrt{n}(\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)\}$ *for* δ *in* (7) *and a small positive constant* $c > 0$*, then with probability larger than* $1 - \alpha_0 - \min\{N_{\mathcal{Q}}, n, p\}^{-c_1}$ *for some* $c_1 > 0$*, there exists* $1 \le m^* \le M$ *such that*

$$\|\widehat{\gamma}_\delta^{[m^*]} - \gamma_\delta^*\|_2 \le \frac{2\sqrt{2}\mathrm{err}_n(M)}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \cdot \frac{1}{\sqrt{n}} \tag{35}$$

By further assuming $N_{\mathcal{Q}} \gtrsim \max\{n,p\}$*, then with probability larger than* $1-\min\{n,p\}^{-c}$*,* $C^*(L,\alpha_0) \geq c$ *for a positive constant* $c>0$*.*

The above theorem characterizes the dependence of the sampling accuracy $\mathrm{err}_n(M)$ on the sampling number M. We further establish the dependence of the best estimation error $\|\widehat{\gamma}_\delta^{[m^*]} - \gamma_\delta^*\|_2$ on the penalty level δ and the sampling accuracy $\mathrm{err}_n(M)$. We shall choose a large sampling number M such that $\mathrm{err}_n(M)$ converges to zero and is much smaller than $\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta$. Then (35) guarantees that at least one sampled maximin weight $\widehat{\gamma}_\delta^{[m^*]}$ converges to the true weight γ_δ^* at a rate faster than $1/\sqrt{n}$. The best estimation error in (35) reveals that a larger value of δ reduces the uncertainty of our sampled weight estimator $\widehat{\gamma}_\delta^{[m^*]}$. If $\Gamma^{\mathcal{Q}}$ is (nearly) singular, we shall choose a positive penalty level $\delta>0$ such that $\delta+\lambda_{\min}(\Gamma^{\mathcal{Q}}) \gg \mathrm{err}_n(M)$. When $C^*(L,\alpha_0) \geq c > 0$ and $\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta$ is of a constant order, we can simply choose $M \gg \log n$. In practice, we set $M=500$ and observe reliable inference results.

6.3 Statistical inference for maximin effects

The following decomposition reveals the intuition on the validity of our proposed CI,

$$\widehat{x_{\text{new}}^{\intercal}\beta}^{[m]} - x_{\text{new}}^{\intercal}\beta = \sum_{l=1}^{L}([\widehat{\gamma}_\delta^{[m]}]_l - [\gamma_\delta^*]_l)\cdot \widehat{x_{\text{new}}^{\intercal}b^{(l)}} + \sum_{l=1}^{L}[\gamma_\delta^*]_l \cdot (\widehat{x_{\text{new}}^{\intercal}b^{(l)}} - x_{\text{new}}^{\intercal}b^{(l)}). \tag{36}$$

With $m=m^*$ in Theorem 3, we shall just quantify the uncertainty of $\sum_{l=1}^{L}[\gamma_\delta^*]_l \cdot (\widehat{x_{\text{new}}^{\intercal}b^{(l)}} - x_{\text{new}}^{\intercal}b^{(l)})$ since the uncertainty of $\sum_{l=1}^{L}(\widehat{\gamma}_\delta^{[m]}]_l - [\gamma_\delta^*]_l)\cdot \widehat{x_{\text{new}}^{\intercal}b^{(l)}}$ is negligible. The following theorem establishes the properties of $\mathrm{CI}(x_{\text{new}}^{\intercal}\beta_\delta^*)$ in (28).

Theorem 4. *Consider the model* (1)*. Suppose Conditions* (A1), (A2) *and* (33) *hold. If the sampling number* M *is chosen such that* $\mathrm{err}_n(M) \ll \min\{1, \lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta\}$ *with* $\mathrm{err}_n(M)$ *defined in Theorem 3 and* δ *in the definition* (7)*, then* $\mathrm{CI}_\alpha(x_{\text{new}}^{\intercal}\beta_\delta^*)$ *in* (28) *satisfies*

$$\lim_{n,p\to\infty} \mathbf{P}\left(x_{\text{new}}^{\intercal}\beta_\delta^* \in \mathrm{CI}\left(x_{\text{new}}^{\intercal}\beta_\delta^*\right)\right) \geq 1-\alpha-\alpha_0,$$

where α *is the pre-specified significance level and* $\alpha_0 \in (0,1/2)$ *is a small positive number. By further assuming* $\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta \gg \sqrt{d_0\log M/n}$ *with* d_0 *in* (22)*, then with probability larger than* $1-\min\{n,p\}^{-c} - M^{-c} - \exp(-t^2)$*,*

$$\mathcal{L}\left(\mathrm{CI}\left(x_{\text{new}}^{\intercal}\beta_\delta^*\right)\right) \lesssim \frac{\sqrt{d_0\log M/n}}{\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta}\cdot\left(\frac{\log n\cdot\|x_{\text{new}}\|_2}{\sqrt{n}} + \max_{1\leq l\leq L}|x_{\text{new}}^{\intercal}b^{(l)}|\right) + (1+t)\frac{\|x_{\text{new}}\|_2}{\sqrt{n}} \tag{37}$$

where $\mathcal{L}\left(\mathrm{CI}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right)$ denotes the interval length and $c > 0$, $t > 0$ are positive constants.

The above theorem justifies the validity of our proposed CI. The small positive value $\alpha_0 \in (0, 1/2)$ is the price paid for establishing the sampling accuracy in Theorem 3. The value α_0 is the probability of $\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}})$ being at the tail of Gaussian distribution.

The upper bound in (37) reveals that the CI length tends to decrease with a larger δ. A major question for the union-type CI in (28) is its conservativeness. We compare its length with an oracle CI relying on known γ_{δ}^{*}, whose length is of order $\|x_{\text{new}}\|_2/\sqrt{n}$. Then we have

$$\frac{\mathcal{L}\left(\mathrm{CI}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right)}{\|x_{\text{new}}\|_2/\sqrt{n}} \lesssim 1 + \sqrt{d_0 \log M} \cdot \frac{\log n/\sqrt{n} + \max_{1\leq l\leq L}|x_{\text{new}}^{\intercal}b^{(l)}|/\|x_{\text{new}}\|_2}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta},$$

which is further upper bounded by $1+\sqrt{d_0 \log M}/(\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)$. Since Proposition 3 implies that d_0 is of a constant order, then the length ratio is at most $\sqrt{\log M}/(\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)$ in terms of order of magnitude, where $\sqrt{\log M}$ comes out of aggregating M confidence intervals. A sharper bound can be established in certain interesting settings. For example, if elements of x_{new} are of the same order, then $\max_{1\leq l\leq L}|x_{\text{new}}^{\intercal}b^{(l)}|/\|x_{\text{new}}\|_2 \lesssim \sqrt{s/p}$ by the sparsity of $b^{(l)}$. The length of our proposed CI is of the same order as the oracle CI under the mild condition $\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta \gg \sqrt{d_0 \log M}(\log n/\sqrt{n} + \sqrt{s/p})$. In Section 7, we examine the finite sample length of the proposed CI and compare it with some oracle benchmark relying on asymptotic normality; see "Efficiency Ratio" plots in Section 7.

7 Simulation Results

7.1 Algorithm 1 implementation with tuning parameters selection

We choose the value of δ through balancing reward optimality and statistical stability. Since $\lambda_{\min}(\Gamma^{\mathcal{Q}})$ may serve as a stability indicator for the non-penalized maximin effect, we then use $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ as a data-dependent stability indicator. If $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ is above a positive constant (say 0.5), the uniquely defined maximin effect can be estimated relatively stably.

The rule of thumb is to choose δ such that $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}}) + \delta$ is above a positive threshold (say 0.5) and the estimated reward corresponding to δ is comparable to the estimated optimal reward (corresponding to $\delta = 0$). We demonstrate how to choose δ across three simulation settings, whose detailed configurations are presented in Section 7.2. In Figure 3, $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ are 0.10, 0.09 and 0.63 from the leftmost panel to the rightmost. This indicates that estimation of non-penalized maximin effect is more stable for setting 3 and less stable for settings 1 and 2. In Figure 3, we plot the estimated rewards for $0 \leq \delta \leq 5$ with $\widehat{\gamma}_{\delta}$ constructed in a single data set with $n = 500$ and $N_{\mathbb{Q}} = 2,000$. For settings 1 and 2, even for $\delta = 2$, the estimated awards are above 95% of the estimated optimal rewards (corresponding to $\delta = 0$). For settings 1 and 2, we recommend $\delta = 2$ as a good balance between reward optimality and stability. For setting 3, since the estimated reward is more sensitive to δ and $\lambda_{\min}(\widehat{\Gamma}^{\mathbb{Q}})$ is relatively large, we recommend the non-penalized maximin effect with $\delta = 0$.

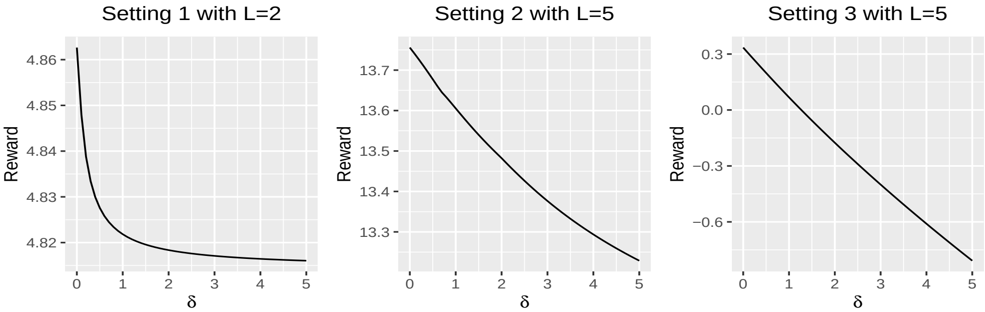


Figure 3: Plot of estimated rewards. For settings $1, 2$ and 3, the optimal rewards $5.14, 13.86$ and 0.48 are estimated by $4.86, 13.75$ and 0.33, respectively.

Existing methods are used to compute $\{\widehat{b}^{(l)}_{init}, \widehat{x_{\rm new}^{\intercal} b^{(l)}}\}_{1 \le l \le L}$: $\widehat{b}^{(l)}_{init}$ is implemented by the R-package `glmnet` (Friedman et al., 2010) with tuning parameters $\{\lambda_l\}_{1 \le l \le L}$ chosen by cross-validation; $\widehat{x_{\rm new}^{\intercal} b^{(l)}}$ is implemented using the online code of the paper (Cai et al., 2019) with its built-in tuning parameter selection. For $\widehat{u}^{(l,k)}$ in (16) and (17), we solve its dual problem

$$\widehat{h} = \operatorname*{arg\,min}_{h \in \mathbb{R}^{p+1}} \frac{1}{4} h^{\intercal} H^{\intercal} \widehat{\Sigma}^{(l)} H h + (\omega^{(k)})^{\intercal} H h / \|\omega^{(k)}\|_2 + \lambda \|h\|_1 \text{ with } H = \left[\frac{\omega^{(k)}}{\|\omega^{(k)}\|_2}, \mathbf{I}_{p \times p} \right],$$

where we adopt the notation $0/0 = 0$. The objective value of this dual problem is unbounded from below when $H^{\intercal} \widehat{\Sigma}^{(l)} H$ is singular and λ is near zero. We choose the smallest $\lambda > 0$ such that the dual problem is bounded from below. We then construct $\widehat{u}^{(l,k)} = -\frac{1}{2}(\widehat{h}_{-1} + \widehat{h}_1 \omega^{(k)} / \|\omega^{(k)}\|_2)$. We set $\tau = 0.2$ in (22) and $M = 500$. Although sample splitting is needed for establish asymptotic normality of $\widehat{\Gamma}^{\mathbb{Q}}$ in the covariate-shift setting, we construct the estimator $\widehat{\Gamma}^{\mathbb{Q}}$ using the full sample and observe reliable inference results in simulations. The algorithm implementation with tuning parameter selection can be found at `https://github.com/zijguo/Maximin-Inference`.

7.2 Numerical results

Simulation Settings. Throughout the simulation, we generate the data $\{X^{(l)}, Y^{(l)}\}_{1 \le l \le L}$ following (1), where, for the l-th group, $\{X^{(l)}_{i,\cdot}\}_{1 \le i \le n_l} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma^{(l)})$ and $\{\epsilon^{(l)}_i\}_{1 \le i \le n_l} \overset{\text{i.i.d.}}{\sim}$

$\mathcal{N}(0, \sigma_l^2)$. The dimension p is set as 500 and $N_{\mathcal{Q}}$ is set as $2,000$ by default. In the covariate shift setting, we assume that we either have access to $N_{\mathcal{Q}}$ i.i.d. samples $\{X^{\mathcal{Q}}_{i,\cdot}\}_{1 \leq i \leq N_{\mathcal{Q}}} \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma^{\mathcal{Q}})$ or we know $\Sigma^{\mathcal{Q}}$ a priori. For $1 \leq l \leq L$, we take $n_l = n$, $\sigma_l = 1$ and $\Sigma^{(l)} = \Sigma$, where $\Sigma_{j,k} = 0.6^{|j-k|}$ for $1 \leq j, k \leq p$; we conduct 500 simulations and report the average measures. We consider the following covariate shift settings.

Setting 1 (covariate shift with $L = 2$): $b^{(1)}_j = j/40$ for $1 \leq j \leq 10$, $b^{(1)}_j = 1$ for $j = 22, 23$, $b^{(1)}_j = 1/2$ for $j = 498$, $b^{(1)}_j = -1/2$ for $j = 499, 500$ and $b^{(1)}_j = 0$ otherwise; $b^{(2)}_j = b^{(1)}_j$ for $1 \leq j \leq 10, j = 22, 23$, $b^{(2)}_j = 1$ for $j = 500$ and $b^{(2)}_j = 0$ otherwise; $\Sigma^{\mathcal{Q}}_{i,i} = 1.5$ for $1 \leq i \leq 500$, $\Sigma^{\mathcal{Q}}_{i,j} = 0.9$ for $1 \leq i \neq j \leq 5$, $\Sigma^{\mathcal{Q}}_{i,j} = 0.9$ for $499 \leq i \neq j \leq 500$ and $\Sigma^{\mathcal{Q}}_{i,j} = \Sigma_{i,j}$ otherwise; $[x_{\rm new}]_j = 1$ for $498 \leq j \leq 500$ and $[x_{\rm new}]_j = 0$ otherwise.

Setting 2 (covariate shift with $L \geq 2$): $b^{(1)}_j = j/10$ for $1 \leq j \leq 10$, $b^{(1)}_j = (10 - j)/10$ for $11 \leq j \leq 20$, $b^{(1)}_j = 1/5$ for $j = 21$, $b^{(1)}_j = 1$ for $j = 22, 23$ and $b^{(1)}_j = 0$ for $24 \leq j \leq 500$; For $2 \leq l \leq L$, $b^{(l)}_j = b^{(1)}_j + 0.1 \cdot (l - 1)/\sqrt{300}$ for $1 \leq j \leq 10$, $b^{(2)}_j = -0.3 \cdot (l - 1)/\sqrt{300}$ for $11 \leq j \leq 20$, $b^{(l)}_j = 0.5 \cdot (l - 1)$ for $j = 21$, $b^{(l)}_j = 0.2 \cdot (j - 1)$ for $j = 22, 23$ and $b^{(2)}_j = 0$ for $24 \leq j \leq 500$; $\Sigma^{\mathcal{Q}}_{i,i} = 1.1$ for $1 \leq i \leq 500$, $\Sigma^{\mathcal{Q}}_{i,j} = 0.75$ for $1 \leq i \neq j \leq 6$ and $\Sigma^{\mathcal{Q}}_{i,j} = \Sigma_{i,j}$ otherwise; $[x_{\rm new}]_j = 1$ for $21 \leq j \leq 23$ and $[x_{\rm new}]_j = 0$ otherwise.

Setting 3 (covariate shift with $L \geq 2$): $b^{(1)}$, $b^{(2)}$ and $\Sigma^{\mathcal{Q}}$ are the same as Setting 2; for $l \geq 3$, $\{b^{(l)}_j\}_{1 \leq j \leq 6}$ are independently generated following standard normal and $b^{(l)}_j = 0$ for $7 \leq j \leq 500$; $x_{\rm new} \sim \mathcal{N}(\mathbf{0}, \Sigma^{\rm new})$ with $\Sigma^{\rm new}_{i,j} = 0.5^{1+|i-j|}/25$ for $1 \leq i, j \leq 500$.

Dependence on n and $\delta > 0$. For setting 1, the Root Mean Square Error (RMSE) of $\widehat{x_{\rm new}^{\intercal}\beta^*_{\delta}}$ in Algorithm 1 is reported in Table 1. The RMSE decreases with an increasing n or a larger δ value. For $n = 300, 500$, the RMSE for $\delta = 2$ (the recommended value) is almost half of that for $\delta = 0$.

n	$\delta = 0$	$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$	$\delta = 5$
100	0.158	0.153	0.139	0.132	0.128	0.127	0.127	0.126
200	0.121	0.114	0.095	0.088	0.085	0.084	0.084	0.083
300	0.116	0.104	0.079	0.070	0.067	0.066	0.066	0.065
500	0.106	0.091	0.063	0.057	0.055	0.055	0.055	0.054

Table 1: Root Mean Square Error of $\widehat{x_{\rm new}^{\intercal}\beta^*_{\delta}}$ for Setting 1.

The inference results for setting 1 are reported in Figure 4. We plot the empirical coverage and length of our proposed CIs over $\delta \in \{0, 0.1, 0.5, 1, 2, 3, 4, 5\}$. The corresponding values of $x_{\rm new}^{\intercal}\beta^*_{\delta}$ are $\{0.1, 0.164, 0.192, 0.196, 0.198, 0.1986, 0.1989, 0.199\}$. In Figure 4, for $n = 300, 500$, with increasing δ from 0 to 2 (the recommended value), the empirical coverage levels drop from 100% to the desired 95%; the CI lengths are reduced by around 50%. For $\delta = 2$, the empirical coverage is around 85% for $n = 100$ and 92.5% for $n = 200$. This supports the discussion in Section 3 that the ridge-type maximin effect is a more stable inference target.

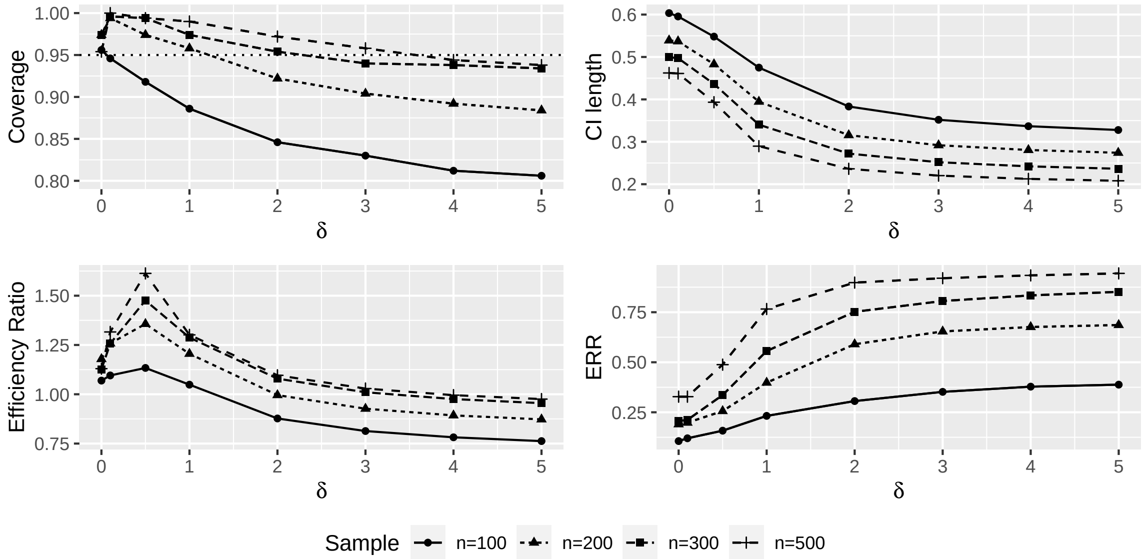


Figure 4: Dependence on δ and n: setting 1 (covariate shift) with unknown $\Sigma^{\mathbb{Q}}$. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to the normal interval with an oracle SE; "ERR" represents the empirical rejection rate out of 500 simulations.

We compare the length of our proposed CI with that of a CI relying on asymptotic normality and the oracle value of standard error (SE). We refer to it as the normal interval with an oracle SE. We construct its center as $\widehat{x_{\text{new}}^{\intercal}\beta_{\delta}^{*}}$ in Algorithm 1 and calculate its length by multiplying the estimated SE of $\widehat{x_{\text{new}}^{\intercal}\beta_{\delta}^{*}}$ by a factor of $2 \cdot 1.96$. Since the SE is estimated after running 500 simulations, this CI can only be computed in oracle settings. We define the "Efficiency Ratio" as the ratio of the length of our proposed CI over that of the normal interval with an oracle SE. Since the limiting distribution of $\widehat{x_{\text{new}}^{\intercal}\beta^{*}}$ is not necessarily normal, the normal interval with an oracle SE might not even be valid (see Figure 5 for an example). The "Efficiency Ratio" is mainly used to give some hints about the efficiency of the proposed CI. In Figure 4, the Efficiency Ratios are below 1.5 in most cases; for $\delta = 2$, the ratio is below 1 for $n = 100$ (corresponding to under-coverage) but is slightly above 1 for $n = 200, 300, 500$.

We report the empirical rejection rate (ERR) of the test $\phi_{\alpha} = \mathbf{1}\left(0 \notin \mathrm{CI}_{\alpha}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right)$, where ERR is defined as the proportion of null hypothesis $x_{\text{new}}^{\intercal}\beta_{\delta}^{*} = 0$ being rejected out of the 500 simulations. Under the null hypothesis, ERR is an empirical measure of the type I error; under the alternative hypothesis, ERR is an empirical measure of the power. In Figure 4, ERR, as a measure of the empirical power, increases with a larger n and δ.

The results for settings 2 and 3 with $L > 2$ are similar to those for setting 1 and are presented in Section D.2 in the supplement.

Irregularity of maximin effects. We consider the scenario where the distribution of $\widehat{\gamma}$ is likely to be a mixture distribution, as in (13). We set $L = 2$; set the loading as $[x_{\rm new}]_j = j/5$ for $1 \leq j \leq 5$ and $[x_{\rm new}]_j = 0$ for $6 \leq j \leq 500$; set $b^{(1)} \in \mathbb{R}^p$ as $b^{(1)}_j = j/40$ for $1 \leq j \leq 10$, $b^{(1)}_j = (10 - j)/40$ for $11 \leq j \leq 20$, $b^{(1)}_{21} = 0.2, b^{(1)}_{22} = b^{(1)}_{23} = 1$ and $b^{(1)}_j = 0$ otherwise; set $b^{(2)} \in \mathbb{R}^p$ as $b^{(2)}_j = b^{(1)}_j + \text{perb}/\sqrt{300}$ for $1 \leq j \leq 10$, $b^{(2)}_j = 0$ for $11 \leq j \leq 20$ and $b^{(2)}_{21} = 0.5, b^{(2)}_{22} = b^{(2)}_{23} = 0.2$ and $b^{(2)}_j = 0$ otherwise. Here, settings 1 to 10 correspond to the values of perb taken as $\{1, 1.125, 1.25, 1.5, 3.75, 4, 5, 7, 10, 12\}$. As reported in Figure 5, the constructed CI achieves the desired coverage level when the sample size reaches 300 and the CI length decreases with a larger sample size. An interesting observation is that, for perturbation settings 9 and 10, although the Efficiency Ratio is around 1.5, the corresponding coverage levels are just achieving the desired 95% level, which indicates the under-coverage of the CI relying on the asymptotic normality and the oracle SE.

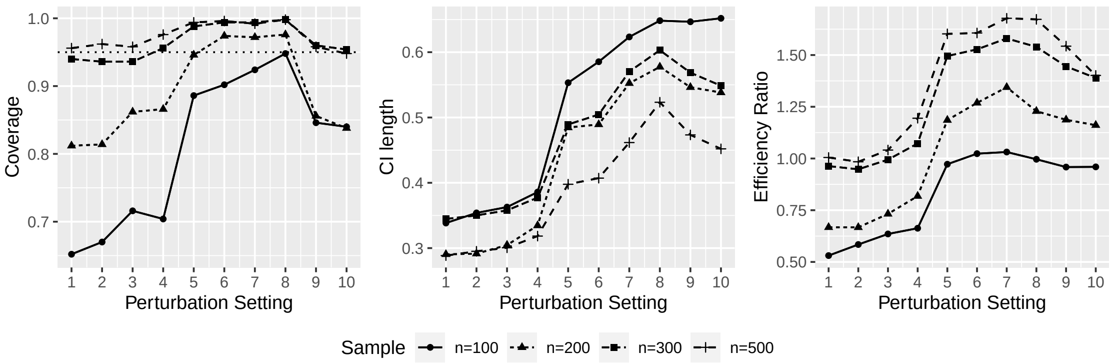
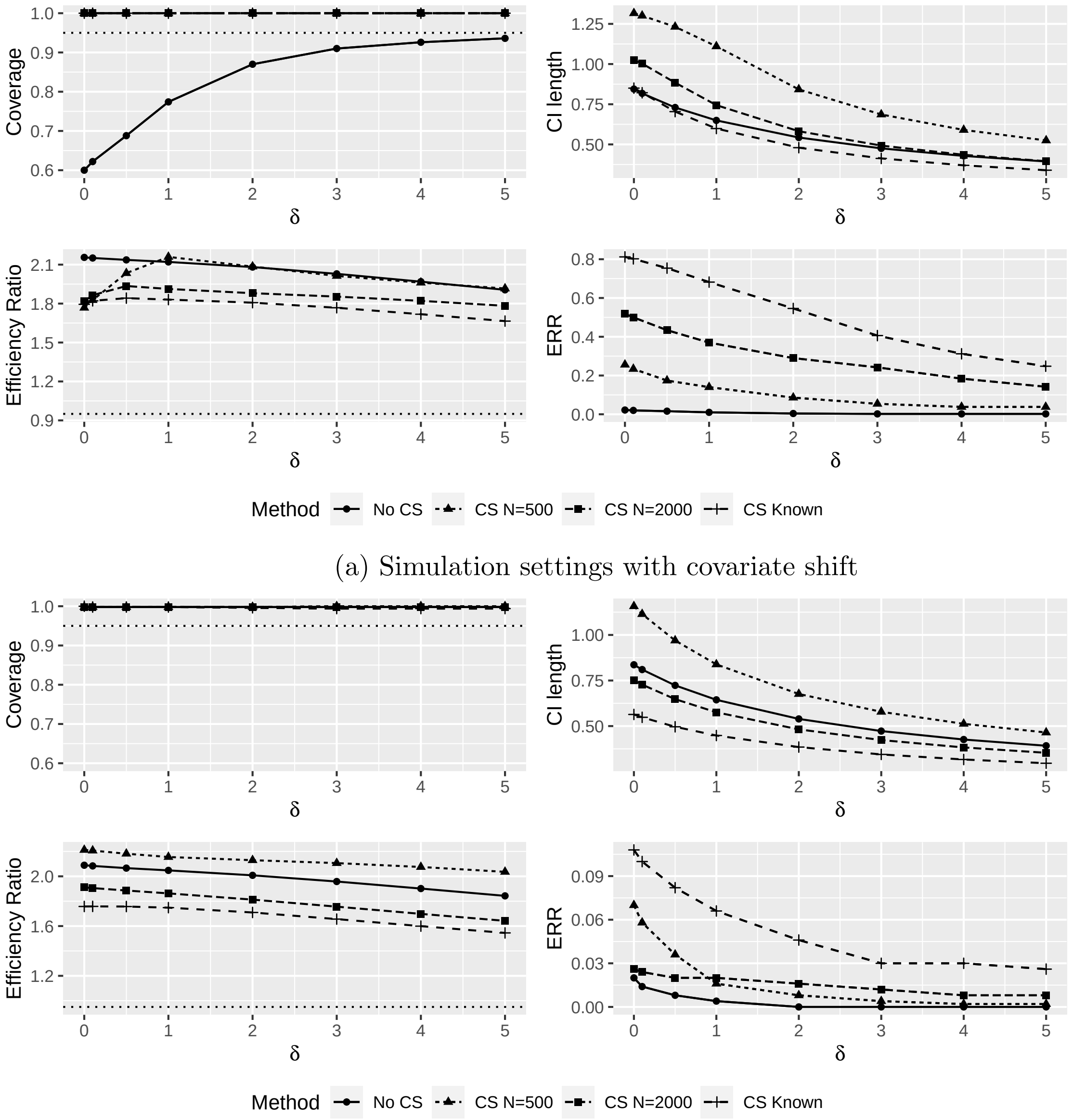


Figure 5: Inference for the maximin effects: settings 1 to 10 correspond to perb being taken as $\{1, 1.125, 1.25, 1.5, 3.75, 4, 5, 7, 10, 12\}$. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to the normal interval with an oracle SE.

Comparison of algorithms with or without covariate shift. We consider covariate shift algorithm (Algorithm 1) with $N_{\mathbb{Q}} = 500, 2{,}000$ or known $\Sigma^{\mathbb{Q}}$ and compare it with the proposed algorithm assuming no covariate shift, which replaces the general estimator $\widehat{\Gamma}^{\mathbb{Q}}$ in Algorithm 1 by (21). We set $L = 2$; generate $b^{(1)} \in \mathbb{R}^p$ as $b^{(1)}_j = j/40$ for $1 \leq j \leq 10$, $b^{(1)}_j = 1$ for $j = 22, 23$ and $b^{(1)}_j = 1/2$ for $j = 498$, $b^{(1)}_j = -1/2$ for $j = 499, 500$ and $b^{(1)}_j = 0$ otherwise; generate $b^{(2)} \in \mathbb{R}^p$ as $b^{(2)}_j = b^{(1)}_j$ for $1 \leq j \leq 10, j = 22, 23$, $b^{(2)}_j = 1$ for $j = 500$ and $b^{(2)}_j = 0$ otherwise. Set the loading $[x_{\rm new}]_j = 1$ for $j = 499, 500$ and $[x_{\rm new}]_j = 0$ otherwise. For the covariate shift setting, we generate $\Sigma^{\mathbb{Q}}$ as $\Sigma^{\mathbb{Q}}_{i,i} = 1.5$ for $1 \leq i \leq 500$, $\Sigma^{\mathbb{Q}}_{i,j} = 0.6$ for $1 \leq i \neq j \leq 5$, $\Sigma^{\mathbb{Q}}_{i,j} = -0.9$ for $499 \leq i \neq j \leq 500$ and $\Sigma^{\mathbb{Q}}_{i,j} = \Sigma_{i,j}$ otherwise.



(a) Simulation settings with covariate shift

(b) Simulation settings with no covariate shift

Figure 6: Comparison of covariate shift and no covariate shift algorithms with $n = 500$. The methods “No CS”, “CS N=500”, “CS N=2000”, “CS Known” represent algorithms assuming no covariate shift, Algorithm 1 (covariate shift) with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. “Coverage” and “CI Length” stand for the empirical coverage and the average length of our proposed CI; “Efficiency Ratio” represents the ratio of the length of CI in (28) to the normal interval with an oracle SE; “ERR” represents the empirical rejection rate out of 500 simulations.

The top of Figure 6 corresponds to the simulation settings with covariate shift and $n = 500$. The no covariate shift algorithm does not achieve the 95% coverage due to the bias of assuming no covariate shift. In contrast, all three covariate shift algorithms achieve the 95% coverage level and the CI lengths decrease with an increasing $N_{\mathcal{Q}}$, where the known $\Sigma^{\mathcal{Q}}$ algorithm can be viewed as $N_{\mathcal{Q}} = \infty$. Correspondingly, the empirical power, measured by ERR, increases with a larger $N_{\mathcal{Q}}$. The bottom of Figure 6 corresponds to the setting with no covariate shift. All algorithm achieves the desired coverage levels. Since the no covariate shift algorithm can be viewed as the setting with $N_{\mathcal{Q}} = 1,000$, the corresponding lengths of the constructed CIs decrease with a larger $N_{\mathcal{Q}}$. The results for sample sizes $\{100, 200, 300\}$ are similar to those for $n = 500$ and are reported in Section D.3 in the supplement.

8 Real Data Applications

We apply our proposed method to a genome-wide association study (Bloom et al., 2013), which studies the yeast colony growth under 46 different growth media. The study is based on $n = 1,008$ Saccharomyces cerevisiae segregants crossbred from a laboratory strain and a wine strain. A set of $p = 4,410$ genetic markers have been selected out of the total 11,623 genetic markers (Bloom et al., 2013). The outcome variables of interest are the end-point colony sizes under different growth media. To demonstrate our method, we consider the colony sizes under five growth media (i.e. $L = 5$): “Ethanol”, “Lactate”, “Lactose”, “Sorbitol” and “Trehalose”. Our model (1) can be applied here with $L = 5$ and each $1 \leq l \leq 5$ corresponds to one growth media (environment). The goal is to find a vector $\beta^*_{\delta}(\mathcal{Q})$ to summarize the relation between the colony growth size and genetic markers across different growth media.

All of these outcome variables are normalized to have variance 1, with corresponding variance explained by the genetic markers $0.60, 0.69, 0.68, 0.51$ and 0.66. We make inference for ridge-type maximin effects for no covariate shift setting and two covariate shift settings. For covariate shift setting 1, we set $\Sigma^{\mathcal{Q}} = \mathrm{I}$; for setting 2, we set $\Sigma^{\mathcal{Q}}_{j,j} = 1$ for $1 \leq j \leq p$ and $\Sigma^{\mathcal{Q}}_{j,l} = 0.75$ for $j \neq l$ and $j, l \in \mathcal{S}_0$, where $\mathcal{S}_0$ is a set of 14 pre-selected important genetic markers. Since $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ is used as a stability indicator for the estimator $\widehat{\beta^*_{\delta=0}}$, the results in Table 2 suggest the use of some positive δ values, e.g. $\delta = 0.5$ and $\delta = 1$. We report the estimated reward (as a minimum of the explained variance across $L = 5$ groups) in Table 2, where $\widehat{R}$ denotes replacing the expectation in (5) with its sample average. The estimated reward decreases by 10% to 15% when δ is increased from 0 to 0.5.

	$\widehat{R}(\widehat{\beta^*_{\delta=0}})$	$\widehat{R}(\widehat{\beta^*_{\delta=0.1}})$	$\widehat{R}(\widehat{\beta^*_{\delta=0.5}})$	$\widehat{R}(\widehat{\beta^*_{\delta=1}})$	$\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$
No Covariate Shift	0.470	0.462	0.421	0.398	0.032
Covariate Shift Setting 1	0.359	0.346	0.302	0.276	0.082
Covariate Shift Setting 2	0.186	0.173	0.164	0.143	0.110

Table 2: Estimated rewards $\widehat{R}(\widehat{\beta^*_{\delta}})$ and the stability indicator $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ for the real data.

In Figure 7, we plot our proposed CI for $[\beta^*_{\delta}(\mathcal{Q})]_j$ for $j \in \mathcal{S}$, where $\mathcal{S} \subset \mathcal{S}_0$ is a set of ten pre-selected regression indexes. We vary δ across $\{0, 0.1, 0.5, 1\}$ and $\mathcal{Q}$ across {no covariate shift, covariate shift setting 1 and 2}. We observe that the constructed CIs get shorter with a larger δ. Additionally, with a different targeted distribution $\mathcal{Q}$, the inference results change but we observe a consistent patten across different targeted covariate distributions. In particular, the variable corresponding to index 2 is the most significant and the corresponding CIs are below zero across different δ; the variable with index 3 becomes significant for $\delta \geq 0.1$; the variables with indexes $\{1, 9, 10\}$ are significant until δ reaches 1 while all remaining variables are not significant even for $\delta = 1$.

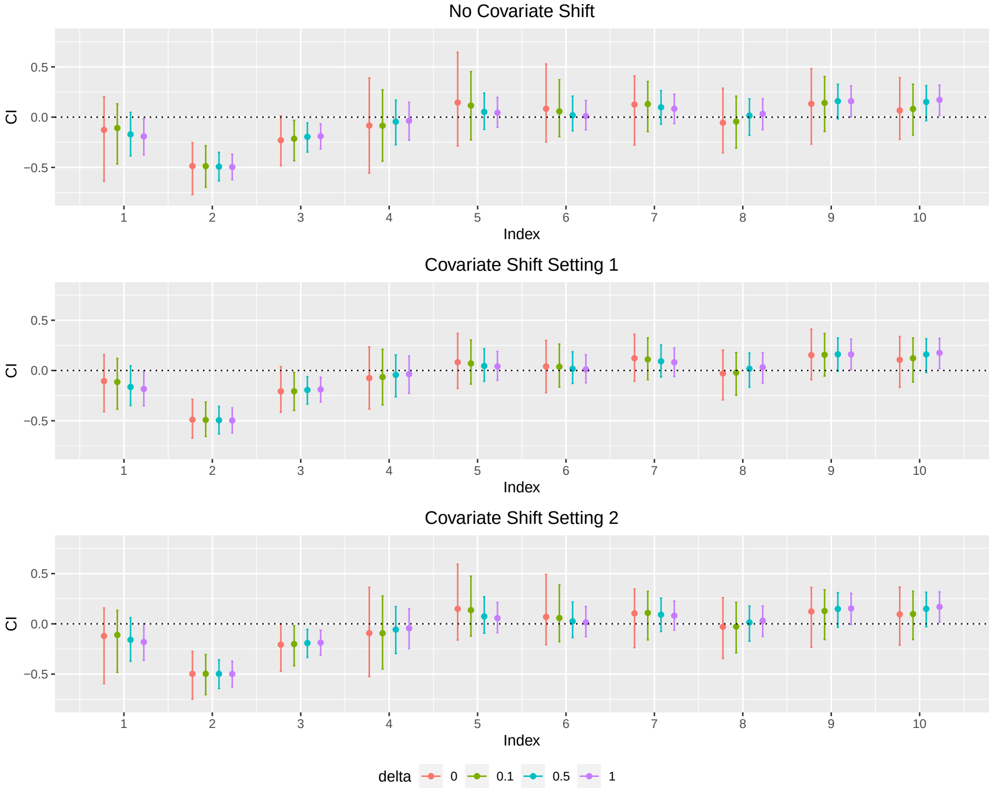


Figure 7: Plot of CI for $[\beta^*_{\delta}(\mathcal{Q})]_j$ for $j \in \mathcal{S}$, where $\mathcal{S}$ is a set of ten genetic markers. δ is varied across $\{0, 0.1, 0.5, 1\}$ and $\mathcal{Q}$ is varied across {no covariate shift, covariate shift 1, 2}.

9 Conclusion and Discussions

We introduce covariate shift and ridge-type maximin effects as effective summaries of heterogeneous regression vectors in high dimensions. Our proposed sampling approach is a computationally efficient uncertainty quantification method for high-dimensional maximin effects. Beyond linear models in (1), an interesting direction is the model aggregation of more complex models in high dimensions, for example, the sparse additive models (Ravikumar et al., 2009; Meier et al., 2009). Another important question is on model aggregation when the linear models in (1) are possibly misspecified (Wasserman, 2014; Bühlmann and van de Geer, 2015). Both questions are left for future research.

Supplement

The supplement contains all proofs and additional methods, theories and simulation results.

References

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Belloni, A., V. Chernozhukov, and L. Wang (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*(4), 791–806.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak (2013). Finding the sources of missing heritability in a yeast cross. *Nature 494*(7436), 234–237.

Bühlmann, P. (2018). Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*.

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

Bühlmann, P. and S. van de Geer (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics 9*(1), 1449–1473.

Cai, T., T. Cai, and Z. Guo (2019). Optimal statistical inference for individualized treatment effects in high-dimensional models. *arXiv preprint arXiv:1904.12891*.

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684.

Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics 45*(2), 615–646.

Cai, T. T. and Z. Guo (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B 82*(2), 391–419.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ. 7*(1), 649–688.

Deusser, E. (1972). Heterogeneity of ribosomal populations in escherichia coli cells grown in different media. *Molecular and General Genetics MGG 119*(3), 249–258.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap.* CRC press.

Einmahl, U. (1989). Extensions of results of komlós, major, and tusnády to the multivariate case. *Journal of multivariate analysis 28*(1), 20–68.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Gao, R., X. Chen, and A. J. Kleywegt (2017). Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*.

Guo, Z., C. Renaux, P. Bühlmann, and T. T. Cai (2019). Group inference in high dimensions with applications to hierarchical testing. *arXiv preprint arXiv:1909.01503*.

Guo, Z., W. Wang, T. T. Cai, and H. Li (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association 114*(525), 358–369.

Hannig, J., H. Iyer, R. C. S. Lai, and T. C. M. Lee (2016). Generalized fiducial inference: A review and new results. *Journal of American Statistical Association.* To appear. Accepted in March 2016. doidoi:10.1080/01621459.2016.1165102.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Liu, M., Y. Xia, T. Cai, and K. Cho (2020). Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *arXiv preprint arXiv:2004.00816*.

Lubke, G. H. and B. Muthén (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods 10*(1), 21.

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

Meinshausen, N. and P. Bühlmann (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics 43*(4), 1801–1830.

Peters, J., P. Bühlmann, and N. Meinshausen (2015). Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Science & Business Media.

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

Rothenhäusler, D., N. Meinshausen, and P. Bühlmann (2016). Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data*, pp. 255–277. Springer.

Rothenhäusler, D., N. Meinshausen, P. Bühlmann, and J. Peters (2018). Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*.

Shi, C., R. Song, W. Lu, and B. Fu (2018). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical methodology 80*(4), 681.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference 90*(2), 227–244.

Sinha, A., H. Namkoong, R. Volpi, and J. Duchi (2017). Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

Sugiyama, M., M. Krauledat, and K.-R. MÃžller (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research 8*(May), 985–1005.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 101*(2), 269–284.

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

Tsuboi, Y., H. Kashima, S. Hido, S. Bickel, and M. Sugiyama (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing 17*, 138–155.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press.

Verzelen, N. and E. Gassiat (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli 24*(4B), 3683–3710.

Wang, P. and M. Xie (2020). Repro sampling method for statistical inference of high dimensional linear models. *Research Manuscript*.

Wasserman, L. (2014). Discussion:" a significance test for the lasso". *The Annals of Statistics 42*(2), 501–508.

Xie, M. and K. Singh (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review 81*, 3–39.

Ye, F. and C.-H. Zhang (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research 11*, 3519–3540.

Zabell, S. L. (1992). Ra fisher and fiducial argument. *Statistical Science 7*(3), 369–387.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhao, T., G. Cheng, and H. Liu (2016). A partially linear framework for massive heterogeneous data. *Annals of statistics 44*(4), 1400.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*.

Zhu, Y. and J. Bradic (2018). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 1–18.

A Additional Methods and Theories

In Section A.1, we discuss the connection to the maximin projection learning problem in Shi et al. (2018). In Section A.2, we present additional results for no covariate shift setting.

A.1 Connection to maximin projection

Follow the model setting in Shi et al. (2018): for the group l with $1 \leq l \leq L$ and $1 \leq i \leq n_l$, the outcome $Y_i^{(l)} \in \mathbb{R}$, the binary treatment $A_i^{(l)} \in \{0,1\}$ and the baseline covariates $X_{i,\cdot}^{(l)} \in \mathbb{R}^p$, are modeled by

$$Y_i^{(l)} = h_l(X_{i,\cdot}^{(l)}) + A_i^{(l)} \cdot [(b^{(l)})^\intercal X_{i,\cdot}^{(l)} + c] + e_i^{(l)}$$

where $h_l : \mathbb{R}^p \to \mathbb{R}$, $b^{(l)} \in \mathbb{R}^p$, $c \in \mathbb{R}$ and the error $e_i^{(l)}$ satisfies $\mathbf{E}(e_i^{(l)} \mid X_{i,\cdot}^{(l)}, A_i^{(l)}) = 0$. In studying the individualized treatment effect, Shi et al. (2018) has proposed the maximum projection

$$\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{1 \leq l \leq L} \beta^\intercal b^{(l)}. \tag{38}$$

Proposition 4. *The maximum projection $\beta^{*,\mathrm{MP}}$ in (38) satisfies*

$$\beta^{*,\mathrm{MP}} = \frac{1}{\|\beta^*(\mathrm{I})\|_2}\beta^*(\mathrm{I}) \quad \textit{with} \quad \beta^*(\mathrm{I}) = \sum_{l=1}^{L} [\gamma^*(\mathrm{I})]_l b^{(l)}$$

where $\gamma^*(\mathrm{I}) = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \Gamma^{\mathrm{I}} \gamma$ and $\Gamma^{\mathrm{I}}_{lk} = (b^{(l)})^\intercal b^{(k)}$ for $1 \leq l, k \leq L$.

Through comparing the above proposition with Proposition 1, we note that the maximin projection is proportional to the general maximin effect defined in (5) with $\Sigma^{\mathbb{Q}} = \mathrm{I}$ and hence the identification of $\beta^*(\mathrm{I})$ is instrumental in identifying $\beta^{*,\mathrm{MP}}$. We refer to Shi et al. (2018) for more details on maximin projection in the low-dimensional setting.

A.2 Inference for $\Gamma^{\mathbb{Q}}$ with No Covariate Shift

We consider the special setting with no covariate shift. The corresponding results are similar to Theorem 2 and Proposition 3 which are derived for the general setting allowing for covariate shift. We define the covariance matrices $\mathbf{V}^{(j)} = (\mathbf{V}^{(j)}_{\pi(l_1,k_1),\pi(l_2,k_2)})_{(l_1,k_1)\in\mathcal{I}_L,(l_2,k_2)\in\mathcal{I}_L} \in$

$\mathbb{R}^{L(L+1)/2\times L(L+1)/2}$ for $j=1,2$ as

$$\begin{aligned}\mathbf{V}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)} &= \frac{\sigma^2_{l_1}}{n^2_{l_1}}[b^{(k_1)}]^{\intercal}[X^{(l_1)}]^{\intercal}\left[X^{(l_2)}b^{(k_2)}\mathbf{1}(l_2=l_1)+X^{(k_2)}b^{(l_2)}\mathbf{1}(k_2=l_1)\right] \\ &\quad+\frac{\sigma^2_{k_1}}{n^2_{l_1}}[b^{(l_1)}]^{\intercal}[X^{(k_1)}]^{\intercal}\left[X^{(l_2)}b^{(k_2)}\mathbf{1}(l_2=k_1)+X^{(k_2)}b^{(l_2)}\mathbf{1}(k_2=k_1)\right]\end{aligned} \tag{39}$$

$$\mathbf{V}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \frac{\mathbf{E}[b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}-(b^{(l_1)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_1)}(b^{(l_2)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_2)}}{\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}} \tag{40}$$

We establish the properties of the estimator $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ in (21) for no covariate shift setting.

Proposition 5. *Consider the model* (1)*. Suppose Condition* (A1) *holds and* $s\log p/n\to 0$ *with* $n=\min_{1\le l\le L}n_l$ *and* $s=\max_{1\le l\le L}\|b^{(l)}\|_0$*. If* $\{X^{(l)}_{i,\cdot}\}_{1\le i\le n_l}\overset{i.i.d.}{\sim}\mathcal{Q}$ *for* $1\le l\le L$*, then the estimator* $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ *defined in* (21) *satisfies*

$$\widehat{\Gamma}^{\mathcal{Q}}_{l,k}-\Gamma^{\mathcal{Q}}_{l,k}=D^{(1)}_{l,k}+D^{(2)}_{l,k}+\mathrm{Rem}_{l,k},$$

where $\mathrm{vecl}(D^{(1)})\mid\{X^{(l)}\}_{1\le l\le L}\sim\mathcal{N}(0,\mathbf{V}^{(1)})$ *with* $\mathbf{V}^{(1)}$ *defined in* (39) *and the components of* $D^{(1)}$ *are uncorrelated with the components of* $D^{(2)}$*; there exists* $D^{(2),*}\in\mathbb{R}^{L\times L}$ *and* $T^*\sim\mathcal{N}(0,\mathbf{V}^{(2)})$ *with* $\mathbf{V}^{(2)}$ *defined in* (40) *such that* $D^{(2),*}\overset{d}{=}D^{(2)}$, $\mathrm{vecl}(D^{(2),*})-T^*=o((\sum_{l=1}^{L}n_l+N_{\mathcal{Q}})^{-2/3})$ *almost surely; for* $1\le l,k\le L$*, the reminder term* $\mathrm{Rem}_{l,k}$ *satisfies with probability larger than* $1-\min\{n,p\}^{-c}$ *for a constant* $c>0$,

$$|\mathrm{Rem}_{l,k}|\lesssim(1+\|b^{(k)}\|_2+\|b^{(l)}\|_2)\cdot\frac{s\log p}{n}. \tag{41}$$

With probability larger than $1-p^{-c}$ *for a positive constant* $c>0$*, for any* $(l,k)\in\mathcal{I}_L$,

$$\mathbf{V}^{(1)}_{\pi(l,k),\pi(l,k)}\asymp\frac{\|b^{(k)}\|_2^2+\|b^{(k)}\|_2^2}{n}\quad\textit{and}\quad\mathbf{V}^{(2)}_{\pi(l,k),\pi(l,k)}\lesssim\frac{\|b^{(l)}\|_2^2\|b^{(k)}\|_2^2}{\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}}. \tag{42}$$

Proposition 5 shows that $\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}})-\mathrm{vecl}(\Gamma^{\mathcal{Q}})$ is approximated by the summation of two random vectors $\mathrm{vecl}(D^{(1)})$ and $\mathrm{vecl}(D^{(2)})$. We show that $\mathrm{vecl}(D^{(1)})$ and $\mathrm{vecl}(D^{(2)})$ are approximated by (conditionally or unconditionally) multivariate Gaussian separately, which is sufficiently for our analysis of characterizing the sampling accuracy.

In general, the theoretical results results for the general covariate shift setting holds for the more special no covariate shift setting. We simply replace D and $\mathbf{V}$ in Theorem 2 with

$D^{(1)} + D^{(2)}$ and $\mathbf{V}^{(1)} + \mathbf{V}^{(2)}$ in Proposition 5, respectively. We estimate $\mathbf{V}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ in (39) and $\mathbf{V}^{(2)}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ in (40) by

$$\begin{aligned}\widehat{\mathbf{V}}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)} &= \frac{\widehat{\sigma}^2_{l_1}}{n^2_{l_1}}[b^{(l_1)}]^{\intercal}[X^{(l_1)}]^{\intercal}X^{(l_1)}\left[b^{(l_2)}\mathbf{1}(l_2 = l_1) + b^{(k_2)}\mathbf{1}(k_2 = l_1)\right] \\ &+ \frac{\widehat{\sigma}^2_{k_1}}{n^2_{l_1}}[b^{(k_1)}]^{\intercal}[X^{(k_1)}]^{\intercal}X^{(k_1)}\left[b^{(l_2)}\mathbf{1}(l_2 = k_1) + b^{(k_2)}\mathbf{1}(k_2 = k_1)\right]\end{aligned} \tag{43}$$

$$\begin{aligned}\widehat{\mathbf{V}}^{(2)}_{\pi(l_1,k_1),\pi(l_2,k_2)} &= \frac{\sum_{i=1}^{N_{\mathcal{Q}}}\left((\widehat{b}^{(l_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (\widehat{b}^{(l_1)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_2)}_{init}\right)}{(\sum_{l=1}^{L} n_l + N_{\mathcal{Q}})^2} \\ &+ \frac{\sum_{l=1}^{L}\sum_{i=1}^{n_l}\left((\widehat{b}^{(l_1)}_{init})^{\intercal}X^{(l)}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^{\intercal}X^{(l)}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^{\intercal}X^{(l)}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^{\intercal}X^{(l)}_{i,\cdot} - (\widehat{b}^{(l_1)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_2)}_{init}\right)}{(\sum_{l=1}^{L} n_l + N_{\mathcal{Q}})^2}\end{aligned} \tag{44}$$

with $\widehat{\Sigma} = \frac{1}{N+N_{\mathcal{Q}}}\left(\sum_{l=1}^{L}\sum_{i=1}^{n_l} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^{\intercal} + \sum_{i=1}^{N_{\mathcal{Q}}} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^{\intercal}\right)$.

The sampling accuracy results in Theorem 3 also hold for the no covariate shift by replacing the constant $C^*(L, \alpha_0)$ in (34) with

$$C^*_1(L, \alpha_0) = \mathrm{Vol}(L(L+1)/2) \cdot \frac{1}{2\sqrt{2\pi}} \cdot \frac{\exp\left(-2F^{-1}_{\chi^2_{r_1}}(1-\frac{\alpha_0}{2}) - F^{-1}_{\chi^2_{r_2}}(1-\frac{\alpha_0}{2})\right)}{\prod_{i=1}^{\frac{L(L+1)}{2}}\left[n\cdot\lambda_i(\mathbf{V}^{(1)}+\mathbf{V}^{(2)})+3d_0/2\right]} \tag{45}$$

where $\mathrm{Vol}(L(L+1)/2)$ denotes the volume of a unit ball in $L(L+1)/2$ dimensions, $1 \leq r_1, r_2 \leq L(L+1)/2$ denote the ranks of $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$, respectively.

Proposition 6. *Consider the model* (1) *with no covariate shift. Suppose Conditions* (A1) *and* (A2) *hold. For* $0 < \alpha_0 < 1/2$, *we define* $\mathrm{err}_n(M) = \left[\frac{2\log n}{C^*_1(L,\alpha_0)M}\right]^{\frac{2}{L(L+1)}}$ *with the positive constant* $C^*_1(L, \alpha_0)$ *defined in* (45). *If* $\mathrm{err}_n(M) \leq c$ *for a small positive constant* $c > 0$, *then there exists a positive constant* $c_1 > 0$ *such that*

$$\mathbf{P}\left(\min_{1\leq m\leq M}\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F/\sqrt{2} \leq \mathrm{err}_n(M)/\sqrt{n}\right) \geq 1 - \alpha_0 - n^{-c_1} - p^{-c_1}. \tag{46}$$

By further assuming $\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta > 2\sqrt{2}\mathrm{err}_n(M)/\sqrt{n}$ *with* δ *defined in* (7), *then with probability larger than* $1 - \alpha_0 - n^{-c_1} - p^{-c_1}$, *there exists* $1 \leq m^* \leq M$ *such that*

$$\|\widehat{\gamma}^{[m^*]}_{\delta} - \gamma^*_{\delta}\|_2 \leq \frac{\sqrt{2}\|\widehat{\Gamma}^{[m^*]} - \Gamma^{\mathcal{Q}}\|_2}{\lambda_{\min}(\widehat{\Gamma}^{[m^*]}) + \delta}\|\gamma^*_{\delta}\|_2 \leq \frac{2\sqrt{2}\mathrm{err}_n(M)}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \cdot \frac{1}{\sqrt{n}} \tag{47}$$

With probability larger than $1 - \min\{n, p\}^{-c}$, $C^*_1(L, \alpha_0) \geq c$ *for a positive constant* $c > 0$.

In comparison to Theorem 3 for the general covariate shift setting, the assumption (33) can be removed for the simpler setting assuming no covariate shift.

B Proofs

In this section, we present the proofs of Theorems 1 to 4 and Propositions 1 to 6. The proofs of additional lemmas are postponed to Section C.

B.1 High probability events

We introduce the following events to facilitate the proofs,

$$
\begin{aligned}
&\mathcal{G}_0 = \left\{ \left\| \frac{1}{n_l} [X^{(l)}]^{\intercal} \epsilon^{(l)} \right\|_{\infty} \lesssim \sqrt{\frac{\log p}{n_l}} \quad \text{for } 1 \leq l \leq L \right\}; \\
&\mathcal{G}_1 = \left\{ \max\left\{ \|\widehat{b}^{(l)}_{init} - b^{(l)}\|_2, \frac{1}{\sqrt{n_l}} \|X^{(l)}(\widehat{b}^{(l)}_{init} - b^{(l)})\|_2 \right\} \lesssim \sqrt{\|b^{(l)}\|_0 \frac{\log p}{n_l}} \sigma_l \quad \text{for } 1 \leq l \leq L \right\}, \\
&\mathcal{G}_2 = \left\{ \|\widehat{b}^{(l)}_{init} - b^{(l)}\|_1 \lesssim \|b^{(l)}\|_0 \sqrt{\frac{\log p}{n_l}} \sigma_l, \|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l^c}\|_1 \leq C \|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l}\|_1 \quad \text{for } 1 \leq l \leq L \right\}, \\
&\mathcal{G}_3 = \left\{ |\widehat{\sigma}_l^2 - \sigma_l^2| \lesssim \|b^{(l)}\|_0 \frac{\log p}{n_l} + \sqrt{\frac{\log p}{n_l}} \quad \text{for } 1 \leq l \leq L \right\},
\end{aligned} \tag{48}
$$

where $\mathcal{S}_l \subset [p]$ denotes the support of $b^{(l)}$ for $1 \leq l \leq L$ and $C > 0$ is a positive constant.

Recall that, for $1 \leq l \leq L$, $\widehat{b}^{(l)}_{init}$ is the Lasso estimator defined in (14) with $\lambda_l = \sqrt{(2+c)\log p / |A_l|}\sigma_l$ for some constant $c > 0$ or the Lasso estimator based on the non-split data; $\widehat{\sigma}_l^2 = \|Y^{(l)} - X^{(l)}\widehat{b}^{(l)}\|_2^2 / n_l$ for $1 \leq l \leq L$. Then under Condition (A1), we can establish

$$
\mathbf{P}\left(\cap_{j=0}^{3} \mathcal{G}_j\right) \geq 1 - \min\{n, p\}^{-c}, \tag{49}
$$

for some $c > 0$. The above high-probability statement follows from the existing literature results on the theoretical analysis of Lasso estimator. We shall point to the exact literature results. The control of the probability of $\mathcal{G}_0$ follows from Lemma 6.2 of Bühlmann and van de Geer (2011). Regarding the events $\mathcal{G}_1$ and $\mathcal{G}_2$, the control of $\|\widehat{b}^{(l)}_{init} - b^{(l)}\|_1$, $\|\widehat{b}^{(l)}_{init} - b^{(l)}\|_2$ and $\frac{1}{\sqrt{n_l}} \|X^{(l)}(\widehat{b}^{(l)}_{init} - b^{(l)})\|_2$ can be found in Theorem 3 of Ye and Zhang (2010), Theorem 7.2 of Bickel et al. (2009) or Theorem 6.1 of Bühlmann and van de Geer (2011); the control of $\|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l^c}\|_1 \leq C \|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l}\|_1$ can be found in Corollary B.2 of Bickel et al. (2009) or

Lemma 6.3 of Bühlmann and van de Geer (2011). For the event $\mathcal{G}_3$, its probability can be controlled as Theorem 2 or (20) in Sun and Zhang (2012).

We further define the following events,

$$\begin{aligned}\mathcal{G}_4 &= \left\{\|\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\|_2 \lesssim \sqrt{\frac{p}{N}} + \frac{p}{N}\right\} \\ \mathcal{G}_5 &= \left\{\max_{\mathcal{S}\subset[p],|\mathcal{S}|\leq s}\ \max_{\|w_{\mathcal{S}^c}\|_1\leq C\|w_{\mathcal{S}}\|_1} \left|\frac{w^{\intercal}\left(\frac{1}{N}\sum_{i=1}^{N} X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}\right)w}{w^{\intercal}E(X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}])w} - 1\right| \lesssim \frac{s\log p}{N}\right\} \\ \mathcal{G}_6(w,v,t) &= \left\{\left|w^{\intercal}\left(\widehat{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\right)v\right| + \left|w^{\intercal}\left(\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\right)v\right| \lesssim t\frac{\|(\Sigma^{\mathcal{Q}})^{1/2}w\|_2\|(\Sigma^{\mathcal{Q}})^{1/2}v\|_2}{\sqrt{N_{\mathcal{Q}}}}\right\}\end{aligned} \tag{50}$$

where $t > 0$ is any positive constant and $w, v \in \mathbb{R}^p$ are pre-specified vectors.

If $X^{\mathcal{Q}}_{i,\cdot}$ is sub-gaussian, then we have

$$\mathbf{P}\left(\mathcal{G}_4 \cap \mathcal{G}_5\right) \geq 1 - p^{-c} \tag{51}$$

$$\mathbf{P}\left(\mathcal{G}_6(w,v,t)\right) \geq 1 - 2\exp(-ct^2) \tag{52}$$

for some positive constant $c > 0$. Since $X^{\mathcal{Q}}_{i,\cdot}$ is sub-gaussian, it follows from equation (5.26) of Vershynin (2012) that the event $\mathcal{G}_4$ holds with probability larger than $1-\exp(-cp)$ for some positive constant $c > 0$.; it follows from Theorem 1.6 of Zhou (2009) that the event $\mathcal{G}_5$ holds with probability larger than $1 - p^{-c}$ for some positive constant $c > 0$. The proof of (52) follows from Lemma 10 in the supplement of Cai and Guo (2020).

B.2 Proofs of Theorem 1 and Proposition 2

For $\mathcal{B} = \begin{pmatrix} b^{(1)} & b^{(2)} & \ldots & b^{(L)}\end{pmatrix} \in \mathbb{R}^{p\times L}$, we define its SVD as

$$\mathcal{B} = U\Lambda V^{\intercal} \quad \text{with} \quad U \in \mathbb{R}^{p\times L}, \Lambda \in \mathbb{R}^{L\times L}, \text{ and } V \in \mathbb{R}^{L\times L}$$

where $\Lambda_{1,1} \geq \ldots \geq \Lambda_{L,L} > 0$. We define $\Delta = \delta \cdot U\Lambda^{-2}U^{\intercal} \in \mathbb{R}^{p\times p}$. Then

$$[\mathcal{B}]^{\intercal}\Delta\mathcal{B} = \delta\mathrm{I} \quad \text{and} \quad \mathcal{B}^{\intercal}(\Sigma^{\mathcal{Q}} + \Delta)\mathcal{B} = \Gamma^{\mathcal{Q}} + \delta\cdot\mathrm{I}. \tag{53}$$

It follows from Proposition 1 and the definition of β^*_{δ} in (7) that

$$\beta^*_{\delta} = \max_{\beta\in\mathbb{R}^p}\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta - \beta^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta\right]$$

and

$$\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta_{\delta}^{*}-[\beta_{\delta}^{*}]^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta_{\delta}^{*}\right]=[\beta_{\delta}^{*}]^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta_{\delta}^{*}=[\gamma_{\delta}^{*}]^{\intercal}\left(\Gamma^{\mathcal{Q}}+\delta\cdot\mathrm{I}\right)\gamma_{\delta}^{*}\quad(54)$$

Now we compute the lower bound for $R_{\mathcal{Q}}(\beta_{\delta}^{*})=\min_{b\in\mathbb{B}}\left[2b^{\intercal}\Sigma^{\mathcal{Q}}\beta_{\delta}^{*}-[\beta_{\delta}^{*}]^{\intercal}\Sigma^{\mathcal{Q}}\beta_{\delta}^{*}\right]$.

We apply (53) and $\beta_{\delta}^{*}=\boldsymbol{\mathcal{B}}\gamma_{\delta}^{*}$ and establish

$$\begin{aligned}&\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta_{\delta}^{*}-[\beta_{\delta}^{*}]^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta_{\delta}^{*}\right]\\&=\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta_{\delta}^{*}-[\beta_{\delta}^{*}]^{\intercal}\Sigma^{\mathcal{Q}}\beta_{\delta}^{*}\right]-\delta\|\gamma_{\delta}^{*}\|_2^2\\&\leq\min_{b\in\mathbb{B}}\left[2b^{\intercal}\Sigma^{\mathcal{Q}}\beta_{\delta}^{*}-[\beta_{\delta}^{*}]^{\intercal}\Sigma^{\mathcal{Q}}\beta_{\delta}^{*}\right]+2\max_{b\in\mathbb{B}}b^{\intercal}\Delta\beta_{\delta}^{*}-\delta\|\gamma_{\delta}^{*}\|_2^2\\&=R_{\mathcal{Q}}(\beta_{\delta}^{*})+2\delta\max_{\gamma\in\Delta^{L}}\gamma^{\intercal}\gamma_{\delta}^{*}-\delta\|\gamma_{\delta}^{*}\|_2^2\end{aligned}\quad(55)$$

where the last inequality follows from Proposition 1 and the function $b^{\intercal}\Delta\beta_{\delta}^{*}$ is linear in b, $\beta_{\delta}^{*}=\boldsymbol{\mathcal{B}}\gamma_{\delta}^{*}$ and (53). We combine (54) and (55) and establish

$$\begin{aligned}R_{\mathcal{Q}}(\beta_{\delta}^{*})&\geq[\gamma_{\delta}^{*}]^{\intercal}\left(\Gamma^{\mathcal{Q}}+\delta\cdot\mathrm{I}\right)\gamma_{\delta}^{*}-2\delta\max_{\gamma\in\Delta^{L}}\gamma^{\intercal}\gamma_{\delta}^{*}+\delta\|\gamma_{\delta}^{*}\|_2^2\\&\geq[\gamma^{*}]^{\intercal}\Gamma^{\mathcal{Q}}\gamma^{*}+2\delta\|\gamma_{\delta}^{*}\|_2^2-2\delta\max_{\gamma\in\Delta^{L}}\gamma^{\intercal}\gamma_{\delta}^{*}\\&=R_{\mathcal{Q}}(\beta^{*})-2\delta\left(\max_{\gamma\in\Delta^{L}}\gamma^{\intercal}\gamma_{\delta}^{*}-\|\gamma_{\delta}^{*}\|_2^2\right)\end{aligned}\quad(56)$$

where the second inequality follows from the definition of γ^{*}. Note that $\max_{\gamma\in\Delta^{L}}\gamma^{\intercal}\gamma_{\delta}^{*}-\|\gamma_{\delta}^{*}\|_2^2\geq 0$ and

$$\max_{\gamma\in\Delta^{L}}\gamma^{\intercal}\gamma_{\delta}^{*}-\|\gamma_{\delta}^{*}\|_2^2=\|\gamma_{\delta}^{*}\|_{\infty}-\|\gamma_{\delta}^{*}\|_2^2$$

We use $j^{*}\in[L]$ to denote the index such that $[\gamma_{\delta}^{*}]_{j^{*}}=\|\gamma_{\delta}^{*}\|_{\infty}$. Then we have

$$\begin{aligned}\|\gamma_{\delta}^{*}\|_{\infty}-\|\gamma_{\delta}^{*}\|_2^2&=[\gamma_{\delta}^{*}]_{j^{*}}-[\gamma_{\delta}^{*}]_{j^{*}}^2-\sum_{l\neq j^{*}}[\gamma_{\delta}^{*}]_l^2\\&\leq[\gamma_{\delta}^{*}]_{j^{*}}-[\gamma_{\delta}^{*}]_{j^{*}}^2-\frac{1}{L-1}\left(\sum_{l\neq j^{*}}[\gamma_{\delta}^{*}]_l\right)^2\\&=[\gamma_{\delta}^{*}]_{j^{*}}-[\gamma_{\delta}^{*}]_{j^{*}}^2-\frac{1}{L-1}(1-[\gamma_{\delta}^{*}]_{j^{*}})^2\end{aligned}\quad(57)$$

We take the maximum value of the right hand side with respect to $[\gamma^*_\delta]_{j^*}$ over the domain $[1/L, 1]$. Then we obtain

$$\max_{\frac{1}{L} \leq [\gamma^*_\delta]_{j^*} \leq 1} [\gamma^*_\delta]_{j^*} - [\gamma^*_\delta]^2_{j^*} - \frac{1}{L-1}(1 - [\gamma^*_\delta]_{j^*})^2 = \frac{1}{4}\left(1 - \frac{1}{L}\right)$$

where the maximum value is achieved at $[\gamma^*_\delta]_{j^*} = \frac{1+\frac{1}{L}}{2}$. Combined with (56) and (57), we establish

$$R_{\mathcal{Q}}(\beta^*_\delta) \geq R_{\mathcal{Q}}(\beta^*) - \frac{\delta}{2} \cdot \left(1 - \frac{1}{L}\right).$$

B.3 Proof of Proposition 2

We establish $[\gamma^*_\delta]_1 \to 1/2$ by (8) and the condition $\delta = \delta(n,p) \gg \max\{|\Gamma_{11} - \Gamma_{12}| + |\Gamma_{22} - \Gamma_{12}|\}$. For a positive definite $\Sigma^{\mathcal{Q}}$, we note that $\lambda_{\min}(\Gamma) \to 0$ implies that $\lambda_L(\mathcal{B}) \to 0$ as $n, p \to \infty$. In the following discussion, we shall use the notation $n \to \infty$ and $p = p(n) \to \infty$.

For $L = 2$, as $[\gamma^*_\delta]_1 \to 1/2$, we have

$$\max_{\gamma \in \Delta^L} \gamma^\intercal \gamma^*_\delta - \|\gamma^*_\delta\|_2^2 = \|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2 \to 0. \tag{58}$$

We consider two separate cases:

1. If $\lambda_L(\mathcal{B}) > 0$ for a given n, then it follows from (56) and (58) that for any given small positive $\epsilon_0 > 0$, there exists N_0 such that for $n \geq N_0$ and $\lambda_L(\mathcal{B}) > 0$, then $|R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*)| \leq \epsilon_0$.

2. If $\lambda_L(\mathcal{B}) = 0$ for a given n, then $b^{(1)}$ and $b^{(2)}$ are collinear. Since $\lambda_{\min}(\Gamma) \to 0$, we have $b^{(2)} = (1 + c_n) b^{(1)}$ with $c_n \to 0$. Hence $\beta^*_\delta = b^{(1)}(1 + [\gamma^*_\delta]_2 c_n)$ and

   $$R_{\mathcal{Q}}(\beta^*_\delta) = (1 + [\gamma^*_\delta]_2 c_n) \min\{1, 1 + c_n\} \, 2[b^{(1)}]^\intercal \Sigma^{\mathcal{Q}} b^{(1)} - (1 + [\gamma^*_\delta]_2 c_n)^2 [b^{(1)}]^\intercal \Sigma^{\mathcal{Q}} b^{(1)}.$$

   For any $[\gamma^*_\delta]_2 \in [0,1]$, we have $R_{\mathcal{Q}}(\beta^*_\delta) \to [b^{(1)}]^\intercal \Sigma^{\mathcal{Q}} b^{(1)}$ for any $\delta \geq 0$. For any given small positive $\epsilon_0 > 0$, there exists N_1 such that for $n \geq N_1$ and $\lambda_L(\mathcal{B}) = 0$, then $|R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*)| \leq \epsilon_0$.

For any $\delta \geq 0$ and $\epsilon_0 > 0$, we take $N_* = \max\{N_0, N_1\}$ and have shown that for $n \geq N_*$, $|R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*)| \leq \epsilon_0$, which implies $R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*) \to 0$.

B.4 Proof of Theorem 2

We decompose the error $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}-\Gamma^{\mathcal{Q}}_{l,k}$ as

$$\begin{aligned}\widehat{\Gamma}^{\mathcal{Q}}_{l,k}-\Gamma^{\mathcal{Q}}_{l,k}=&\frac{1}{|B_l|}(\widehat{u}^{(l,k)})^{\intercal}[X^{(l)}_{B_l,\cdot}]^{\intercal}\epsilon^{(l)}_{B_l}+\frac{1}{|B_k|}(\widehat{u}^{(k,l)})^{\intercal}[X^{(k)}_{B_k,\cdot}]^{\intercal}\epsilon^{(k)}_{B_k}\\&+(b^{(l)})^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}}-\Sigma^{\mathcal{Q}})b^{(k)}-(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}^{(k)}_{init}-b^{(k)})\\&+(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})+(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(l)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(k,l)})^{\intercal}(\widehat{b}^{(k)}_{init}-b^{(k)})\end{aligned}\tag{59}$$

We define $D_{l,k}=D^{(a)}_{l,k}+D^{(b)}_{l,k}$ with

$$D^{(a)}_{l,k}=\frac{1}{|B_l|}(\widehat{u}^{(l,k)})^{\intercal}[X^{(l)}_{B_l,\cdot}]^{\intercal}\epsilon^{(l)}_{B_l}+\frac{1}{|B_k|}(\widehat{u}^{(k,l)})^{\intercal}[X^{(k)}_{B_k,\cdot}]^{\intercal}\epsilon^{(k)}_{B_k}$$

and

$$D^{(b)}_{l,k}=(b^{(l)})^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}}-\Sigma^{\mathcal{Q}})b^{(k)}.$$

Note that $D^{(a)}_{l,k}$ is a function of $X^{\mathcal{Q}}_{A,\cdot}$, $\{X^{(l)}\}_{1\leq l\leq L}$ and $\{\epsilon^{(k)}\}_{1\leq l\leq L}$ and $D^{(b)}_{l,k}$ is a function of the sub-sample $X^{\mathcal{Q}}_{B,\cdot}$ and hence $D^{(a)}_{l,k}$ is independent of $D^{(b)}_{l,k}$. We define $\mathrm{Rem}_{l,k}$ as

$$\begin{aligned}\mathrm{Rem}_{l,k}=&-(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}^{(k)}_{init}-b^{(k)})+(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})\\&+(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(l)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(k,l)})^{\intercal}(\widehat{b}^{(k)}_{init}-b^{(k)}).\end{aligned}$$

By (59), we have $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}-\Gamma^{\mathcal{Q}}_{l,k}=D_{l,k}+\mathrm{Rem}_{l,k}$. In the following, we control the decomposition term by term.

Control of the reminder term $\mathrm{Rem}_{l,k}$. The following lemma controls the reminder term $\mathrm{Rem}_{l,k}$ in (30).

Lemma 1. *With probability larger than* $1-\min\{n,p\}^{-c}$, *we have*

$$\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\sqrt{\frac{\|b^{(l)}\|_0\|b^{(k)}\|_0(\log p)^2}{n_l n_k}}.\tag{60}$$

$$\left|(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\lesssim\|\omega^{(k)}\|_2\frac{\|b^{(l)}\|_0\log p}{n_l}+\|\widehat{b}^{(k)}_{init}\|_2\sqrt{\frac{\|b^{(l)}\|_0(\log p)^2}{n_l N_{\mathcal{Q}}}}\tag{61}$$

$$\left|(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(l)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(k,l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\|\omega^{(l)}\|_2\frac{\|b^{(k)}\|_0\log p}{n_k}+\|\widehat{b}^{(l)}_{init}\|_2\sqrt{\frac{\|b^{(k)}\|_0(\log p)^2}{n_k N_{\mathcal{Q}}}}\tag{62}$$

The proof of the above lemma is presented in Section C.1.

Distribution of $D^{(a)}$. Since $\widehat{b}^{(k)}_{init}$ is a function of $(X^{(k)}_{A_k,\cdot}, \epsilon^{(k)}_{A_k})$, the projection direction $\widehat{u}^{(l,k)}$ is a function of $X^{\mathcal{Q}}_{A,\cdot}$, $X^{(l)}_{B_l,\cdot}$ and $(X^{(l)}_{A_k,\cdot}, \epsilon^{(k)}_{A_k})$. By the Gaussian error assumption, we establish

$$\mathrm{vecl}(D^{(a)}) \mid X^{\mathcal{Q}}_{A,\cdot}, \{X^{(l)}\}_{1\leq l\leq L}, \{\epsilon^{(k)}_{A_l}\}_{1\leq l\leq L} \sim N(\mathbf{0}, \mathbf{V}^{(a)}) \tag{63}$$

where

$$\begin{aligned}\mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} &= \mathrm{Cov}(D^{(a)}_{l_1,k_1}, D^{(a)}_{l_2,k_2}) \\ &= \frac{\sigma^2_{l_1}}{|B_{l_1}|}(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1)+\widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right] \\ &+ \frac{\sigma^2_{k_1}}{|B_{k_1}|}(\widehat{u}^{(k_1,l_1)})^{\intercal}\widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=k_1)+\widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=k_1)\right]\end{aligned} \tag{64}$$

with the index mapping π defined in (3).

Distribution of $D^{(2)}$. For $i \in B$ and $(l,k) \in \mathcal{I}_L$, we define $W_{i,\cdot} \in \mathbb{R}^{L(L+1)/2}$ as

$$W_{i,\pi(l,k)} = [b^{(l)}]^{\intercal}(X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal} - \Sigma^{\mathcal{Q}})b^{(k)}$$

Then we have

$$[\mathrm{vecl}(D^{(b)})]_{\pi(l,k)} = D^{(b)}_{l,k} = \frac{1}{|B|}\sum_{i\in B} W_{i,\pi(l,k)} \quad \text{and} \quad \mathrm{vecl}(D^{(b)}) = \frac{1}{|B|}\sum_{i\in B} W_{i,\cdot}$$

where the index mapping π is defined in (3). The random variable $W_{i,\cdot} \in \mathbb{R}^{L(L+1)/2}$ is of mean zero and covariance matrix $\mathbf{C} \in \mathbb{R}^{L(L+1)/2 \times L(L+1)/2}$, defined as,

$$\begin{aligned}&\mathbf{C}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \mathbf{E}W_{i,\pi(l_1,k_1)}[W_{i,\pi(l_2,k_2)}]^{\intercal} \\ &= \mathbf{E}\left([b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (b^{(l_1)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_1)}\right)\left([b^{(l_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (b^{(l_2)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_2)}\right).\end{aligned} \tag{65}$$

Now we shall show that $[\mathrm{vecl}(D^{(b)})]$ can be approximated by a normal random variable. The following lemma is the multivariate version of Komlós, Major, and Tusnády theory, which restates Theorem 1 of Einmahl (1989) in the current paper's terminology.

Lemma 2. *Consider a sequence of i.i.d. random vectors $W_{i,\cdot}$. Let $W_{i,\cdot} : \Omega \to \mathbb{R}^d$ be a random vector with mean zero and covariance matrix $\mathbf{C}$. Suppose that $\mathbf{E}\|W_{i,\cdot}\|_2^{r_0} < \infty$ for $r_0 \geq 3$.*

Then one can construct a probability space $(\Omega_0, \mathcal{A}_0, P_0)$ *and two sequences of independent random vectors* $\{W^0_{i,\cdot}\}_{1\le i\le n}$ *and* $\{Z^0_{i,\cdot}\}_{1\le i\le n}$ *with* $W^0_{i,\cdot}$ *having the same distribution as* $W_{i,\cdot}$ *and* $Z^0_{i,\cdot} \sim \mathcal{N}(0, \mathbf{C})$ *such that*

$$\sum_{i=1}^{n} W^0_{i,\cdot} - \sum_{i=1}^{n} Z^0_{i,\cdot} = o(n^{1/r_0}) \quad a.s. \tag{66}$$

Now we verify conditions of the above lemma. For $i \in B$, $\mathbf{E}W_{i,\cdot}$ is of mean zero and covariance $\mathbf{C}$ defined in (65). For any $r_0 \ge 3$, we have

$$\begin{aligned}\mathbf{E}\|W_{i,\cdot}\|_2^{r_0} = \mathbf{E}\left(\sum_{(l,k)\in\mathcal{I}_L} W^2_{i,\pi(l,k)}\right)^{\frac{r_0}{2}} &\le (L(L+1)/2)^{\frac{r_0}{2}}\, \mathbf{E} \max_{(l,k)\in\mathcal{I}_L} |W_{i,\pi(l,k)}|^{r_0} \\ &\le (L(L+1)/2)^{\frac{r_0}{2}} \sum_{(l,k)\in\mathcal{I}_L} \mathbf{E}|W_{i,\pi(l,k)}|^{r_0} \\ &\le (L(L+1)/2)^{\frac{r_0}{2}+1} \max_{(l,k)\in\mathcal{I}_L} \mathbf{E}|W_{i,\pi(l,k)}|^{r_0}\end{aligned}$$

Since $(b^{(l)})^\intercal X^{\mathcal{Q}}_{i,\cdot}$ and $(b^{(k)})^\intercal X^{\mathcal{Q}}_{i,\cdot}$ are sub-gaussian random variables, then $[b^{(l)}]^\intercal (X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^\intercal - \Sigma) b^{(k)}$ is sub-exponential, that is, $\mathbf{E}|W_{i,\pi(l,k)}|^{r_0} \lesssim r_0^{r_0}$. Hence

$$\mathbf{E}\|W_{i,\cdot}\|_2^{r} \lesssim (L(L+1)/2)^{\frac{r_0}{2}+1} r_0^{r_0}. \tag{67}$$

For any given positive $r_0 \ge 3$, we have shown that $\mathbf{E}\|W_{i,\cdot}\|_2^{r_0} \le C_0$ for some constant $C_0 > 0$. By applying Lemma 2, we establish that there exist two sequences of independent random vectors $W^0_{i,\cdot}$ and $Z^0_{i,\cdot} \sim \mathcal{N}(0, \mathbf{C})$ for $i \in B$ such that $W^0_{i,\cdot}$ has the same distribution as $W_{i,\cdot}$ and

$$\left\|\sum_{i\in B} W^0_{i,\cdot} - \sum_{i\in B} Z^0_{i,\cdot}\right\|_2 \lesssim N_{\mathcal{Q}}^{\frac{1}{r_0}} \quad \text{a.s.} \tag{68}$$

We define

$$\text{vecl}(D^{(b),*}) = \frac{1}{|B|} \sum_{i\in B} W^0_{i,\cdot}$$

and as a consequence, the corresponding random matrix $D^{(b),*}$ has the same distribution as $D^{(b)}$. We couple the underlying probability spaces of $D^{(a)}$ and $D^{(b),*}$ and use the product measure of these two spaces as the joint probability measure of $D^{(a)}$ and $D^{(b),*}$. After coupling, the two random objects $D^{(b),*}$ and $D^{(a)}$ are independent of each other. Since $D^{(a)}$

is independent of $D^{(b)}$, we establish

$$\mathrm{vecl}(D^{(a)})+\mathrm{vecl}(D^{(b)})\overset{d}{=}\mathrm{vecl}(D^{(a)})+\mathrm{vecl}(D^{(b),*}) \tag{69}$$

We define $D^*=D^{(a)}+D^{(b),*}$. A combination of (68) and (69) leads to

$$\|\mathrm{vecl}(D^*)-S^*\|_2=o(N_{\mathcal{Q}}^{1/r_0-1})\quad\text{with}\quad S^*=\mathrm{vecl}(D^{(a)})+\frac{1}{|B|}\sum_{i\in B}Z^0_{i,\cdot}. \tag{70}$$

Conditioning on $X^{\mathcal{Q}}_{A,\cdot}$ and $\{X^{(l)},\epsilon^{(k)}_{A_l}\}_{1\le l\le L}$, the random variable $S^*\sim\mathcal{N}(0,\mathbf{V})$ with $\mathbf{V}$ defined in (29). We then establish the distribution of D in Theorem 2 since $1-1/r_0\in(2/3,1)$.

B.5 Proof of Proposition 3

We have the expression for the diagonal element of $\mathbf{V}$ as

$$\begin{aligned}\mathbf{V}_{\pi(l,k),\pi(l,k)}=&\frac{\sigma_l^2}{|B_l|}(\widehat{u}^{(l,k)})^\intercal\widehat{\Sigma}^{(l)}\left[\widehat{u}^{(l,k)}+\widehat{u}^{(k,l)}\mathbf{1}(k=l)\right]+\frac{\sigma_k^2}{|B_k|}(\widehat{u}^{(k,l)})^\intercal\widehat{\Sigma}^{(k)}\left[\widehat{u}^{(l,k)}\mathbf{1}(l=k)+\widehat{u}^{(k,l)}\right]\\&+\frac{1}{|B|}(\mathbf{E}[b^{(l)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}-(b^{(l)})^\intercal\Sigma^{\mathcal{Q}}b^{(k)}(b^{(l)})^\intercal\Sigma^{\mathcal{Q}}b^{(k)})\end{aligned}$$

We introduce the following lemma, which restates Lemma 1 of Cai et al. (2019) in the current paper's terminology.

Lemma 3. *Suppose that the condition* (A1) *holds, then with probability larger than* $1-p^{-c}$,

$$c\frac{\|\omega^{(k)}\|_2^2}{n_l}\le\frac{1}{|B_l|}(\widehat{u}^{(l,k)})^\intercal\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)}\le C\frac{\|\omega^{(k)}\|_2^2}{n_l},\quad\textit{for}\quad 1\le l,k\le L,$$

$$c\frac{\|\omega^{(l)}\|_2^2}{n_k}\le\frac{1}{|B_k|}(\widehat{u}^{(k,l)})^\intercal\widehat{\Sigma}^{(k)}\widehat{u}^{(l,k)}\le C\frac{\|\omega^{(l)}\|_2^2}{n_k},\quad\textit{for}\quad 1\le l,k\le L,$$

for some positive constants $C>c>0$.

Since

$$\begin{aligned}&\mathbf{E}[b^{(l)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}-(b^{(l)})^\intercal\Sigma^{\mathcal{Q}}b^{(k)}(b^{(l)})^\intercal\Sigma^{\mathcal{Q}}b^{(k)}\\&=\mathbf{E}\left([b^{(l)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^\intercal[b^{(k)}]-(b^{(l)})^\intercal\Sigma^{\mathcal{Q}}b^{(k)}\right)^2\ge 0,\end{aligned}$$

we apply Lemma 3 to establish the lower bound in (31). We also apply Lemma 3 to establish part of the upper bound in (31). Since $X^{\mathcal{Q}}_{i,\cdot}$ is sub-gaussian, we have

$$\left|\mathbf{E}[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}\right| \lesssim \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2 \tag{71}$$

and

$$(b^{(l_1)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_1)}(b^{(l_2)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_2)} \lesssim \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2. \tag{72}$$

Then we establish the upper bound of (31) by taking $l_1 = l_2 = l$ and $k_1 = k_2 = k$.

For the setting of known $\Sigma^{\mathcal{Q}}$, on the event $\mathcal{G}_2$ defined in (48), we establish

$$\|\omega^{(k)}\|_2 = \|\Sigma^{\mathcal{Q}}\widehat{b}^{(k)}_{init}\|_2 \lesssim \lambda_{\max}(\Sigma^{\mathcal{Q}})\|\widehat{b}^{(k)}_{init}\|_2 \lesssim \lambda_{\max}(\Sigma^{\mathcal{Q}})\left(\|b^{(k)}\|_2 + \sqrt{s\log p/n}\right).$$

To establish (32), we control $\|\omega^{(k)}\|_2$ as follows,

$$\begin{aligned}\|\omega^{(k)}\|_2 = \|\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}\|_2 &\leq \|\Sigma^{\mathcal{Q}}\widehat{b}^{(k)}_{init}\|_2 + \|(\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}})\widehat{b}^{(k)}_{init}\|_2 \\ &\leq \lambda_{\max}(\Sigma^{\mathcal{Q}})\|\widehat{b}^{(k)}_{init}\|_2 + \|\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\|_2\|\widehat{b}^{(k)}_{init}\|_2\end{aligned}$$

On the event $\mathcal{G}_4$, we establish

$$\|\omega^{(k)}\|_2 \lesssim \lambda_{\max}(\Sigma^{\mathcal{Q}})\left(1 + \sqrt{\frac{p}{N_{\mathcal{Q}}}} + \frac{p}{N_{\mathcal{Q}}}\right)\left(\|b^{(k)}\|_2 + \sqrt{s\log p/n}\right)$$

With a similar bound for $\|\omega^{(l)}\|_2$, we establish (32).

B.6 Proof of Theorem 3

In the following, we first prove

$$\mathbf{P}\left(\min_{1\leq m\leq M}\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F/\sqrt{2} \leq \mathrm{err}_n(M)/\sqrt{n}\right) \geq 1 - \alpha_0 - n^{-c_1} - p^{-c_1}, \tag{73}$$

and then prove (35).

Proof of (73). Denote all data by $\mathcal{O}$, that is, $\mathcal{O} = \{X^{(l)}, Y^{(l)}\}_{1\leq l\leq L} \cup \{X^{\mathcal{Q}}\}$. Define $n = \min_{1\leq l\leq L} n_l$. We rescale the difference as $\widehat{Z} = \sqrt{n}\left(\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}}) - \mathrm{vecl}(\Gamma)\right)$ and the sampled difference as $Z^{[m]} = \sqrt{n}S^{[m]} = \sqrt{n}\left(\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}}) - \mathrm{vecl}(\Gamma^{[m]})\right)$. Note that $\widehat{Z} = \sqrt{n}\left(\mathrm{vecl}(\widehat{\Gamma}) - \mathrm{vecl}(\Gamma)\right)$ is a function of $\mathcal{O}$ and

$$\widehat{Z} - Z^{[m]} = \sqrt{n}\left(\mathrm{vecl}(\widehat{\Gamma}^{[m]}) - \mathrm{vecl}(\Gamma)\right) \tag{74}$$

The rescaled covariance matrix $\mathbf{Cov}$ and estimated covariance matrix $\widehat{\mathbf{Cov}}$ are defined as

$$\mathbf{Cov} = n\mathbf{V} \quad \text{and} \quad \widehat{\mathbf{Cov}} = n\widehat{\mathbf{V}},$$

with $\mathbf{V}$ and $\widehat{\mathbf{V}}$ defined in (29) and (21), respectively. Let $f(Z \mid \mathcal{O})$ denote the conditional density function of $Z^{[m]}$ given the data $\mathcal{O}$, that is,

$$f(Z \mid \mathcal{O}) = \frac{1}{\sqrt{2\pi \det(\widehat{\mathbf{Cov}} + d_0\mathbf{I})}} \exp\left(-\frac{1}{2} Z^{\intercal} (\widehat{\mathbf{Cov}} + d_0\mathbf{I})^{-1} Z\right).$$

We define the following function to facilitate the proof,

$$g(Z) = \frac{1}{\sqrt{2\pi \det(\mathbf{Cov} + \frac{3}{2}d_0\mathbf{I})}} \exp\left(-\frac{1}{2} Z^{\intercal} (\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1} Z\right). \tag{75}$$

We define the following event for the data $\mathcal{O}$,

$$\begin{aligned}
\mathcal{E}_1 &= \left\{\|\widehat{\mathbf{Cov}} - \mathbf{Cov}\|_2 < d_0/2\right\} \\
\mathcal{E}_2 &= \cup_{1 \le l,k \le L} \left\{\mathrm{Rem}_{l,k} \text{ satisfies (30)}\right\} \\
\mathcal{E}_3 &= \left\{g(\widehat{Z}) \cdot \mathbf{1}_{\mathcal{E}_1 \cap \mathcal{E}_2} \ge c^*(1-\alpha_0)\right\}
\end{aligned} \tag{76}$$

where $\|\widehat{\mathbf{Cov}} - \mathbf{Cov}\|_2$ denotes the spectral norm of the matrix $\widehat{\mathbf{Cov}} - \mathbf{Cov}$ and

$$c^*(1-\alpha_0) = \frac{1}{2\sqrt{2\pi}} \prod_{i=1}^{\frac{L(L+1)}{2}} \left[n \cdot \lambda_i(\mathbf{V}) + 3d_0/2\right]^{-1} \exp\left(-F^{-1}_{\chi^2_r}(1-\alpha_0)\right), \tag{77}$$

with r denoting the rank of $\mathbf{V}$ and $F^{-1}_{\chi^2_r}(1-\alpha_0)$ denoting $1-\alpha_0$ quantile of the χ^2 distribution with degree of freedom r.

The following two lemmas show that the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ holds with a high probability.

Lemma 4. *Suppose that the conditions of Theorem 3 hold, then we have*

$$\mathbf{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \ge 1 - \min\{N_{\mathcal{Q}}, n, p\}^{-c} \tag{78}$$

for some positive constant $c > 0$.

Lemma 5. *Suppose that the conditions of Theorem 3, then we have*

$$\mathbf{P}\left(\mathcal{E}_3\right) \geq \mathbf{P}(\mathcal{E}_1 \cap \mathcal{E}_2) - \alpha_0. \tag{79}$$

The proofs of Lemmas 4 and 5 are presented in Sections C.3 and C.4, respectively. We lower bound the targeted probability in (73) as

$$\begin{aligned}
&\mathbf{P}\left(\min_{1\leq m\leq M} \|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F \leq \sqrt{2}\mathrm{err}_n(M)/\sqrt{n}\right)\\
&\geq \mathbf{P}\left(\min_{1\leq m\leq M} \|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M)\right)\\
&\geq \mathbf{E}_{\mathcal{O}}\mathbf{P}\left(\min_{1\leq m\leq M} \|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}
\end{aligned} \tag{80}$$

where $\mathbf{P}(\cdot \mid \mathcal{O})$ denotes the conditional probability with respect to the observed data $\mathcal{O}$ and $\mathbf{E}_{\mathcal{O}}$ denotes the expectation taken with respect to the observed data $\mathcal{O}$. Note that

$$\begin{aligned}
&\mathbf{P}\left(\min_{1\leq m\leq M} \|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right)\\
&= 1 - \mathbf{P}\left(\min_{1\leq m\leq M} \|Z^{[m]} - \widehat{Z}\|_2 \geq \mathrm{err}_n(M) \mid \mathcal{O}\right)\\
&= 1 - \prod_{m=1}^{M}\left[1 - \mathbf{P}\left(\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right)\right]\\
&\geq 1 - \exp\left[-M \cdot \mathbf{P}\left(\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right)\right],
\end{aligned}$$

where the second equality follows from the conditional independence of $\{Z^{[m]}\}_{1\leq m\leq M}$ given the data $\mathcal{O}$ and the last inequality follows from $1 - x \leq e^{-x}$. Hence, we have

$$\begin{aligned}
&\mathbf{P}\left(\min_{1\leq m\leq M} \|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\
&\geq \left(1 - \exp\left[-M \cdot \mathbf{P}\left(\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right)\right]\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\
&= 1 - \exp\left[-M \cdot \mathbf{P}\left(\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\right]
\end{aligned} \tag{81}$$

Together with (80), it is sufficient to establish an lower bound for

$$\mathbf{P}\left(\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}. \tag{82}$$

On the event $\mathcal{O} \in \mathcal{E}_1$, we have

$$\mathbf{Cov} + \frac{3}{2} d_0 \mathbf{I} \succ \widehat{\mathbf{Cov}} + d_0 \mathbf{I} \succ \mathbf{Cov} + \frac{1}{2} d_0 \mathbf{I} \tag{83}$$

and

$$f(Z^{[m]} \mid \mathcal{O}) \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1} \geq g(Z^{[m]}) \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1}.$$

We apply the above inequality and further lower bound the targeted probability in (82) as

$$\begin{aligned}
&\mathbf{P}\left(\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} \\
=& \int f(Z^{[m]} \mid \mathcal{O}) \mathbf{1}_{\left\{\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} \\
\geq& \int g(Z^{[m]}) \mathbf{1}_{\left\{\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} \\
=& \int g(\widehat{Z}) \mathbf{1}_{\left\{\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} \\
&+ \int [g(Z^{[m]}) - g(\widehat{Z})] \mathbf{1}_{\left\{\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3}
\end{aligned} \tag{84}$$

By the definition of $\mathcal{E}_3$, we establish

$$\begin{aligned}
&\int g(\widehat{Z}) \mathbf{1}_{\left\{\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} \\
\geq& c^*(1 - \alpha_0) \cdot \int \mathbf{1}_{\left\{\|Z^{[m]} - \widehat{Z}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} \\
\geq& c^*(1 - \alpha_0) \cdot \mathrm{Vol}(L(L+1)/2) \cdot \mathrm{err}_n(M)^{L(L+1)/2} \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3}
\end{aligned} \tag{85}$$

where $\mathrm{Vol}(L(L+1)/2)$ denotes the volume of the unit ball in $\frac{L(L+1)}{2}$-dimension. Note that there exists $t \in (0, 1)$ such that

$$g(Z^{[m]}) - g(\widehat{Z}) = [\nabla g(\widehat{Z} + t(Z^{[m]} - \widehat{Z}))]^{\intercal} (Z^{[m]} - \widehat{Z})$$

with

$$\nabla g(w) = \frac{1}{\sqrt{2\pi \det(\mathbf{Cov} + \frac{3}{2} d_0 \mathbf{I})}} \exp\left(-\frac{1}{2} w^{\intercal} (\mathbf{Cov} + \frac{1}{2} d_0 \mathbf{I})^{-1} w\right) (\mathbf{Cov} + \frac{1}{2} d_0 \mathbf{I})^{-1} w.$$

Since $\lambda_{\min}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})\geq d_0/2$, then $\triangledown g$ is bounded and $\left|g(Z^{[m]})-g(\widehat{Z})\right|\leq C\|Z^{[m]}-\widehat{Z}\|_2$ for a positive constant $C>0$. Then we establish

$$\begin{aligned}&\left|\int[g(Z^{[m]})-g(\widehat{Z})]\mathbf{1}_{\left\{\|Z^{[m]}-\widehat{Z}\|_2\leq \mathrm{err}_n(M)\right\}}dZ^{[m]}\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\right|\\&\leq C\mathrm{err}_n(M)\cdot\int\mathbf{1}_{\left\{\|Z^{[m]}-\widehat{Z}\|_2\leq \mathrm{err}_n(M)\right\}}dZ^{[m]}\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\&=C\mathrm{err}_n(M)\cdot\mathrm{Vol}(L(L+1)/2)\cdot\mathrm{err}_n(M)^{L(L+1)/2}\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}.\end{aligned}\tag{86}$$

By assuming $C\mathrm{err}_n(M)\leq\frac{1}{2}c^*(1-\alpha_0)$, we combine (84), (85) and (86) and obtain

$$\begin{aligned}&\mathbf{P}\left(\|Z^{[m]}-\widehat{Z}\|_2\leq\mathrm{err}_n(M)\mid\mathcal{O}\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\&\geq\frac{1}{2}c^*(1-\alpha_0)\cdot\mathrm{Vol}(L(L+1)/2)\cdot\mathrm{err}_n(M)^{L(L+1)/2}\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\end{aligned}$$

Together with (81), we establish

$$\begin{aligned}&\mathbf{P}\left(\min_{1\leq m\leq M}\|Z^{[m]}-\widehat{Z}\|_2\leq\mathrm{err}_n(M)\mid\mathcal{O}\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\&\geq 1-\exp\left[-M\cdot\frac{1}{2}c^*(1-\alpha_0)\cdot\mathrm{Vol}(L(L+1)/2)\cdot\mathrm{err}_n(M)^{L(L+1)/2}\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\right]\\&=\left(1-\exp\left[-M\cdot\frac{1}{2}c^*(1-\alpha_0)\cdot\mathrm{Vol}(L(L+1)/2)\cdot\mathrm{err}_n(M)^{L(L+1)/2}\right]\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\end{aligned}\tag{87}$$

Together with (80), we have

$$\begin{aligned}&\mathbf{P}\left(\min_{1\leq m\leq M}\|Z^{[m]}-\widehat{Z}\|_2\leq\mathrm{err}_n(M)\right)\\&\geq\left(1-\exp\left[-M\cdot\frac{1}{2}c^*(1-\alpha_0)\cdot\mathrm{Vol}(L(L+1)/2)\cdot\mathrm{err}_n(M)^{L(L+1)/2}\right]\right)\mathbf{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\right)\end{aligned}$$

Since $C^*(L,\alpha_0)$ defined in (34) satisfies $C^*(L,\alpha_0)=\mathrm{Vol}(L(L+1)/2)c^*(1-\alpha_0)$, we choose

$$\mathrm{err}_n(M)=\left[\frac{2\log n}{C^*(L,\alpha_0)M}\right]^{\frac{2}{L(L+1)}}$$

and establish

$$\mathbf{P}\left(\min_{1\leq m\leq M}\|Z^{[m]}-\widehat{Z}\|_2\leq\mathrm{err}_n(M)\right)\geq(1-n^{-1})\cdot\mathbf{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\right).$$

We further apply Lemmas 4 and 5 and establish

$$\mathbf{P}\left(\min_{1\leq m\leq M}\|Z^{[m]}-\widehat{Z}\|_2\leq \mathrm{err}_n(M)\right)\geq (1-n^{-1})(1-\alpha_0-\exp(-cn)-p^{-c})$$
$$\geq 1-\alpha_0-n^{-c_1}-p^{-c_1}.$$

Proof of (35). We shall quantify the error of the ridge-type sampled weights $\widehat{\gamma}_{\delta}^{[m]}$ by the corresponding error of $\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_2$.

Lemma 6. *If $\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_2\leq(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)/2$, then*

$$\|\widehat{\gamma}_{\delta}^{[m]}-\gamma_{\delta}^{*}\|_2\leq\frac{2\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_F}{\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta}\|\gamma_{\delta}^{*}\|_2 \tag{88}$$

where $\widehat{\gamma}_{\delta}^{[m]}$ and $\gamma_{\delta}^{*}=\gamma_{\delta}^{*}(\mathcal{Q})$ are defined in (24) and (7), respectively.

The above lemma will be proved in Section B.6.1.

It follows from (73) that, with probability larger than $1-\alpha_0-n^{-c_1}-p^{-c_1}$, there exists $1\leq m^*\leq M$ such that $\|\widehat{\Gamma}^{[m^*]}-\Gamma^{\mathcal{Q}}\|_F\leq\sqrt{2}\mathrm{err}(M)/\sqrt{n}$. Under the assumption $\sqrt{2}\mathrm{err}(M)/\sqrt{n}\leq(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)/2$, we have $\|\widehat{\Gamma}^{[m^*]}-\Gamma^{\mathcal{Q}}\|_2\leq\|\widehat{\Gamma}^{[m^*]}-\Gamma^{\mathcal{Q}}\|_F\leq(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)/2$. Then we apply (88), together with $\|\gamma_{\delta}^{*}\|_2\leq\|\gamma_{\delta}^{*}\|_1\leq 1$, to establish (35).

For a finite L and $N_{\mathcal{Q}}\gtrsim\max\{n,p\}$, we show that $\mathrm{Vol}(L(L+1)/2)$ and $F^{-1}_{\chi^2_r}(1-\alpha_0)$ are bounded from above and $n\cdot\lambda_i(\mathbf{V})\lesssim n\cdot\|\mathbf{V}\|_\infty\lesssim d_0$. By Proposition 3, under Condition (A1), $s\log p\ll n$ and $N_{\mathcal{Q}}\gtrsim\max\{n,p\}$, we show that $\|n\mathbf{V}\|_\infty$ and d_0 are bounded from above with probability larger than $1-\min\{n,p\}^{-c}$ for some positive constant $c>0$. This implies that $C^*(L,\alpha_0)\geq c$ for a positive constant $c>0$.

B.6.1 Proof of Lemma 6

Recall that $\gamma^*=\gamma^*_{\delta=0}$ and $\widehat{\gamma}^{[m]}=\widehat{\gamma}^{[m]}_{\delta=0}$ where γ^*_{δ} is defined in (6) and $\widehat{\gamma}^{[m]}_{\delta}$ is defined in (24). We first consider the setting with $\delta=0$ and control $\|\gamma^*-\widehat{\gamma}^{[m]}\|_2$; then we extend the proof to control $\|\gamma^*_{\delta}-\widehat{\gamma}^{[m]}_{\delta}\|_2$ for any $\delta\geq 0$. By the definition of γ^* in (6) with $\delta=0$, for any $t\in(0,1)$, we have

$$(\gamma^*)^{\intercal}\Gamma^{\mathcal{Q}}\gamma^*\leq\left[\gamma^*+t(\widehat{\gamma}^{[m]}-\gamma^*)\right]^{\intercal}\Gamma^{\mathcal{Q}}\left[\gamma^*+t(\widehat{\gamma}^{[m]}-\gamma^*)\right]$$

and hence

$$0\leq 2t(\gamma^*)^{\intercal}\Gamma^{\mathcal{Q}}(\widehat{\gamma}^{[m]}-\gamma^*)+t^2(\widehat{\gamma}^{[m]}-\gamma^*)^{\intercal}\Gamma^{\mathcal{Q}}(\widehat{\gamma}^{[m]}-\gamma^*)$$

By taking $t \to 0+$, we have

$$(\gamma^*)^\intercal \Gamma^{\mathcal{Q}}(\widehat{\gamma}^{[m]} - \gamma^*) \geq 0. \tag{89}$$

By the definition of $\widehat{\gamma}^{[m]}$ in (24) with $\delta = 0$, for any $t \in (0,1)$, we have

$$(\widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}_{+}^{[m]} \widehat{\gamma}^{[m]} \leq \left[\widehat{\gamma}^{[m]} + t(\gamma^* - \widehat{\gamma}^{[m]})\right]^\intercal \widehat{\Gamma}_{+}^{[m]} \left[\widehat{\gamma}^{[m]} + t(\gamma^* - \widehat{\gamma}^{[m]})\right]$$

This gives us

$$2(\gamma^*)^\intercal \widehat{\Gamma}_{+}^{[m]}(\gamma^* - \widehat{\gamma}^{[m]}) + (t-2)(\gamma^* - \widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}_{+}^{[m]}(\gamma^* - \widehat{\gamma}^{[m]}) \geq 0 \tag{90}$$

Since $2 - t > 0$, we have

$$(\gamma^* - \widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}_{+}^{[m]}(\gamma^* - \widehat{\gamma}^{[m]}) \leq \frac{2}{2-t}(\gamma^*)^\intercal \widehat{\Gamma}_{+}^{[m]}(\gamma^* - \widehat{\gamma}^{[m]}). \tag{91}$$

It follows from (89) that

$$\begin{aligned}(\gamma^*)^\intercal \widehat{\Gamma}_{+}^{[m]}(\gamma^* - \widehat{\gamma}^{[m]}) &= (\gamma^*)^\intercal \Gamma^{\mathcal{Q}}(\gamma^* - \widehat{\gamma}^{[m]}) + (\gamma^*)^\intercal(\widehat{\Gamma}_{+}^{[m]} - \Gamma^{\mathcal{Q}})(\gamma^* - \widehat{\gamma}^{[m]}) \\ &\leq (\gamma^*)^\intercal(\widehat{\Gamma}_{+}^{[m]} - \Gamma^{\mathcal{Q}})(\gamma^* - \widehat{\gamma}^{[m]})\end{aligned}$$

Combined with (91), we have

$$\begin{aligned}(\gamma^* - \widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}_{+}^{[m]}(\gamma^* - \widehat{\gamma}^{[m]}) &\leq \frac{2}{2-t}(\gamma^*)^\intercal(\widehat{\Gamma}_{+}^{[m]} - \Gamma^{\mathcal{Q}})(\gamma^* - \widehat{\gamma}^{[m]}) \\ &\leq \frac{2\|\gamma^*\|_2}{2-t}\|\widehat{\Gamma}_{+}^{[m]} - \Gamma^{\mathcal{Q}}\|_2\|\gamma^* - \widehat{\gamma}^{[m]}\|_2 \\ &\leq \frac{2\|\gamma^*\|_2}{2-t}\|\widehat{\Gamma}_{+}^{[m]} - \Gamma^{\mathcal{Q}}\|_F\|\gamma^* - \widehat{\gamma}^{[m]}\|_2\end{aligned} \tag{92}$$

Since $\gamma^*_\delta = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \left(\Gamma^{\mathcal{Q}} + \delta \mathrm{I}\right)\gamma$ and $\widehat{\gamma}^{[m]}_\delta = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \left(\widehat{\Gamma}^{[m]} + \delta \mathrm{I}\right)\gamma$, we can apply the same argument for (92) by replacing respectively $\Gamma^{\mathcal{Q}}$ and $\widehat{\Gamma}_{+}^{[m]}$ by $\Gamma^{\mathcal{Q}} + \delta \mathrm{I}$ and $\widehat{\Gamma}_{+}^{[m]} + \delta \mathrm{I}$ and establish

$$\begin{aligned}(\widehat{\gamma}^{[m]}_\delta - \gamma^*_\delta)^\intercal \left(\widehat{\Gamma}_{+}^{[m]} + \delta \mathrm{I}\right)(\widehat{\gamma}^{[m]}_\delta - \gamma^*_\delta) &\leq \frac{2\|\gamma^*_\delta\|_2}{2-t}\|\widehat{\Gamma}_{+}^{[m]} - \Gamma^{\mathcal{Q}}\|_F\|\widehat{\gamma}^{[m]}_\delta - \gamma^*_\delta\|_2 \\ &\leq \frac{2\|\gamma^*_\delta\|_2}{2-t}\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F\|\widehat{\gamma}^{[m]}_\delta - \gamma^*_\delta\|_2\end{aligned}$$

where the last inequality follows from the definition of $\widehat{\Gamma}^{[m]}_{+}$ and that $\Gamma^{\mathcal{Q}}$ is positive semi-definite. We apply the above bound by taking $t \to 0+$ and establish

$$\|\widehat{\gamma}_{\delta}^{[m]} - \gamma_{\delta}^{*}\|_2 \leq \frac{\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F}{\max\{\lambda_{\min}(\widehat{\Gamma}^{[m]}), 0\} + \delta} \|\gamma_{\delta}^{*}\|_2$$

Since $\left|\lambda_{\min}(\widehat{\Gamma}^{[m]}) - \lambda_{\min}(\Gamma^{\mathcal{Q}})\right| \leq \|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_2$, if $\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta > 2\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_2$, then we have $\lambda_{\min}(\widehat{\Gamma}^{[m]}) + \delta > \frac{1}{2}(\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)$ and

$$\|\widehat{\gamma}_{\delta}^{[m]} - \gamma_{\delta}^{*}\|_2 \leq \frac{2\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \|\gamma_{\delta}^{*}\|_2.$$

B.7 Proof of Theorem 4

We introduce the following high-probability events to facilitate the discussion.

$$\begin{aligned}
\mathcal{E}_4 &= \left\{ \frac{1}{n_l} [\widehat{v}^{(l)}]^{\intercal} (X^{(l)})^{\intercal} X^{(l)} \widehat{v}^{(l)} \asymp \|x_{\text{new}}\|_2 \quad \text{for } 1 \leq l \leq L \right\} \\
\mathcal{E}_5 &= \left\{ \frac{\left| \sum_{l=1}^{L} \left([\widehat{\gamma}_{\delta}^{[m]}]_l - [\gamma_{\delta}^{*}]_l \right) \widehat{x_{\text{new}}^{\intercal} b^{(l)}} \right|}{\sqrt{\sum_{l=1}^{L} [\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2} [\widehat{v}^{(l)}]^{\intercal} (X^{(l)})^{\intercal} X^{(l)} \widehat{v}^{(l)}}} \lesssim \sqrt{n} \|\widehat{\gamma}_{\delta}^{[m]} - \gamma_{\delta}^{*}\|_2 \right\} \\
\mathcal{E}_6 &= \left\{ \|\widehat{\gamma}_{\delta}^{[m^*]} - \gamma_{\delta}^{*}\|_2 \leq \frac{2\sqrt{2}\text{err}_n(M)}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \cdot \frac{1}{\sqrt{n}} \quad \text{for some } 1 \leq m^* \leq M \right\}
\end{aligned} \tag{93}$$

We apply Lemma 1 of Cai et al. (2019) and establish that $\mathbf{P}(\mathcal{E}_4) \geq 1 - p^{-c}$, for some positive constant $c > 0$. On the event $\mathcal{E}_4$, we have

$$\sqrt{\sum_{l=1}^{L} [\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2} [\widehat{v}^{(l)}]^{\intercal} (X^{(l)})^{\intercal} X^{(l)} \widehat{v}^{(l)}} \asymp \frac{\|\gamma_{\delta}^{*}\|_2 \|x_{\text{new}}\|_2}{\sqrt{n}} \asymp \frac{\|x_{\text{new}}\|_2}{\sqrt{n}}. \tag{94}$$

where the last asymptotic equivalence holds since $\frac{1}{\sqrt{L}} \leq \|\gamma_{\delta}^{*}\|_2 \leq 1$. Similarly, we show that, on the event $\mathcal{E}_4$,

$$\sqrt{\sum_{l=1}^{L} [\widehat{\gamma}_{\delta}^{[m]}]_l^2 \frac{\sigma_l^2}{n_l^2} [\widehat{v}^{(l)}]^{\intercal} (X^{(l)})^{\intercal} X^{(l)} \widehat{v}^{(l)}} \asymp \frac{\|\widehat{\gamma}_{\delta}^{[m]}\|_2 \|x_{\text{new}}\|_2}{\sqrt{n}} \asymp \frac{\|x_{\text{new}}\|_2}{\sqrt{n}}. \tag{95}$$

Note that

$$\begin{aligned}\left|\sum_{l=1}^{L}\left([\widehat{\gamma}_{\delta}^{[m]}]_l-[\gamma_{\delta}^{*}]_l\right)\widehat{x_{\rm new}^{\intercal}b^{(l)}}\right| &\leq \|\widehat{\gamma}_{\delta}^{[m]}-\gamma_{\delta}^{*}\|_2\sqrt{\sum_{l=1}^{L}[\widehat{x_{\rm new}^{\intercal}b^{(l)}}]^2} \\ &\lesssim \|\widehat{\gamma}_{\delta}^{[m]}-\gamma_{\delta}^{*}\|_2\sqrt{\sum_{l=1}^{L}[\widehat{x_{\rm new}^{\intercal}b^{(l)}}-x_{\rm new}^{\intercal}b^{(l)}]^2+[x_{\rm new}^{\intercal}b^{(l)}]^2}\end{aligned} \tag{96}$$

By Theorem 2 of Cai et al. (2019) and (94), with probability larger than $1-\min\{n,p\}^{-c}$ for some positive constant $c>0$,

$$\begin{aligned}&\frac{\left|\sum_{l=1}^{L}\left([\widehat{\gamma}_{\delta}^{[m]}]_l-[\gamma_{\delta}^{*}]_l\right)\widehat{x_{\rm new}^{\intercal}b^{(l)}}\right|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2\frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \\ &\lesssim \|\widehat{\gamma}_{\delta}^{[m]}-\gamma_{\delta}^{*}\|_2\cdot\frac{\sqrt{L}\cdot(\log n\cdot\frac{\|x_{\rm new}\|_2}{\sqrt{n}}+\max_{1\leq l\leq L}|x_{\rm new}^{\intercal}b^{(l)}|)}{\frac{1}{\sqrt{L}}\cdot\frac{\|x_{\rm new}\|_2}{\sqrt{n}}}\lesssim\sqrt{n}\|\widehat{\gamma}_{\delta}^{[m]}-\gamma_{\delta}^{*}\|_2\end{aligned} \tag{97}$$

where the last inequality follows from the bounded $\|b^{(l)}\|_2$ and finite L. This implies that $\mathbf{P}(\mathcal{E}_5)\geq 1-\min\{n,p\}^{-c}$ for some positive constant $c>0$. It follows from (35) of Theorem 3 that $\mathbf{P}(\mathcal{E}_6)\geq 1-\alpha_0-n^{-c_1}-p^{-c_1}$ for some positive constant $c_1>0$.

Coverage property. It follows from Theorem 2 of Cai et al. (2019) that

$$\frac{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l[\widehat{x_{\rm new}^{\intercal}b^{(l)}}-x_{\rm new}^{\intercal}b^{(l)}]}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2\frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\xrightarrow{d}N(0,1). \tag{98}$$

We shall take m^{*} as in Theorem 3. By the CI definition in (28), we have

$$\begin{aligned}&\mathbf{P}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\not\in{\rm CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right)\right)\leq\mathbf{P}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\not\in{\rm Int}_{\alpha}^{[m^{*}]}(x_{\rm new})\right) \\ &=\mathbf{P}\left(\frac{|\widehat{x_{\rm new}^{\intercal}\beta}^{[m^{*}]}-x_{\rm new}^{\intercal}\beta_{\delta}^{*}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2\frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\geq 1.01\cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^{*}]}]_l^2\frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2\frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right)\end{aligned}$$

By applying (36) with $m = m^*$, we further upper bound the above inequality as

$$
\begin{aligned}
&\mathbf{P}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*} \notin \mathrm{CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right)\right) \\
&\leq \mathbf{P}\left(\frac{|\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l \cdot (\widehat{x_{\rm new}^{\intercal}b^{(l)}} - x_{\rm new}^{\intercal}b^{(l)})|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \geq z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right) \\
&+ \mathbf{P}\left(\frac{|\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m^*]}]_l - [\gamma_{\delta}^{*}]_l) \cdot \widehat{x_{\rm new}^{\intercal}b^{(l)}}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \geq 0.01 \cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right)
\end{aligned}
\tag{99}
$$

On the event $\mathcal{E}_5 \cap \mathcal{E}_6$, we have

$$
\frac{|\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m^*]}]_l - [\gamma_{\delta}^{*}]_l) \cdot \widehat{x_{\rm new}^{\intercal}b^{(l)}}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \leq \frac{2\sqrt{2}\mathrm{err}_n(M)}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta}.
$$

On the event $\mathcal{G}_3 \cap \mathcal{E}_4$ with $\mathcal{G}_3$ defined in (48) and $\mathcal{E}_4$ defined in (93), we apply (94) and (95) to establish

$$
0.01 \cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \geq c
$$

for some small constant $c > 0$. Since $\mathrm{err}_n(M) \ll \lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta$, we have

$$
\frac{|\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m^*]}]_l - [\gamma_{\delta}^{*}]_l) \cdot \widehat{x_{\rm new}^{\intercal}b^{(l)}}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \leq 0.01 \cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}.
$$

This implies

$$
\begin{aligned}
&\mathbf{P}\left(\frac{|\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m^*]}]_l - [\gamma_{\delta}^{*}]_l) \cdot \widehat{x_{\rm new}^{\intercal}b^{(l)}}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \leq 0.01 \cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right) \\
&\geq \mathbf{P}\left(\mathcal{G}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6\right) \geq 1 - \alpha_0 - n^{-c_1} - p^{-c_1}.
\end{aligned}
\tag{100}
$$

Since the events $\mathcal{E}_4$ and $\mathcal{E}_6$ hold with high probability, we have

$$\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2\frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2\frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\xrightarrow{p}1.$$

Together with (98), we establish

$$\mathbf{P}\left(\frac{|\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l\cdot(\widehat{x_{\text{new}}^{\intercal}b^{(l)}}-x_{\text{new}}^{\intercal}b^{(l)})|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2\frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\geq z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2\frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2\frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right)\to\alpha.$$

Combined with (99) and (100), we have

$$\lim_{n,p\to\infty}\mathbf{P}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\notin\text{CI}_{\alpha}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right)\leq\alpha+\alpha_0.$$

Confidence interval length. We control the length by the decomposition (36) for $1\leq m\leq M$. Note that with probability larger than $1-M^{-c}$, we have

$$\left|\widehat{\Gamma}_{l,k}^{[m]}-\Gamma_{l,k}^{\mathcal{Q}}\right|\lesssim\sqrt{d_0/n}\sqrt{2\log M}$$

where the high probability bound follows from Theorem 2 and Proposition 3. As a consequence, for a finite L, with probability larger than $1-M^{-c}$,

$$\max_{1\leq m\leq M}\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_2\lesssim\sqrt{L\cdot d_0/n}\cdot\sqrt{\log M}$$

By Lemma 6, if $\sqrt{L\cdot d_0/n}\sqrt{\log M}\ll(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)/2$, then

$$\max_{1\leq m\leq M}\|\widehat{\gamma}_{\delta}^{[m]}-\gamma_{\delta}^{*}\|_2\lesssim\frac{2\sqrt{L\cdot d_0/n}\cdot\sqrt{\log M}}{\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta}\|\gamma_{\delta}^{*}\|_2\tag{101}$$

Together with (96) and (97), we establish

$$\left|\sum_{l=1}^{L}\left([\widehat{\gamma}_{\delta}^{[m]}]_l-[\gamma_{\delta}^{*}]_l\right)\widehat{x_{\text{new}}^{\intercal}b^{(l)}}\right|\lesssim\frac{\sqrt{L\cdot d_0/n}\cdot\sqrt{\log M}}{2(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)}\cdot\sqrt{L}\cdot(\log n\cdot\frac{\|x_{\text{new}}\|_2}{\sqrt{n}}+\max_{1\leq l\leq L}|x_{\text{new}}^{\intercal}b^{(l)}|)\tag{102}$$

By the decomposition (36), we establish (37) by combining (98), (95) and (102).

B.8 Proof of Proposition 1

We now supply a proof of Proposition 1, which essentially follows from the same argument as that of Theorem 1 in Meinshausen and Bühlmann (2015). Let $\mathbb{H}$ denote the convex hull of the set $\mathbb{B} = \{b^{(1)}, \ldots, b^{(L)}\}$. By the linear form of b, we have

$$\begin{aligned} \beta^* &= \arg\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{B}} \left[2b^\intercal \Sigma^{\mathbb{Q}} \beta - \beta^\intercal \Sigma^{\mathbb{Q}} \beta \right] \\ &= \arg\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{H}} \left[2b^\intercal \Sigma^{\mathbb{Q}} \beta - \beta^\intercal \Sigma^{\mathbb{Q}} \beta \right] \end{aligned}$$

We decompose $\Sigma^{\mathbb{Q}} = C^\intercal C$ such that C is invertible. Define $\widetilde{\mathbb{H}} = C^{-1}\mathbb{H}$. Then we have $\beta^* = C^{-1}\xi^*$ with

$$\xi^* = \arg\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} \left[2u^\intercal \xi - \xi^\intercal \xi \right] \tag{103}$$

If we interchange min and max in the above equation, then we have

$$\xi^* = \arg\min_{\xi \in \widetilde{\mathbb{H}}} \xi^\intercal \xi \tag{104}$$

We will justify this inter-change by showing that the solution ξ^* defined in (104) is the solution to (103). For any $\nu \in [0,1]$ and $\mu \in \widetilde{\mathbb{H}}$, we use the fact $\xi^* + \nu(\mu - \xi^*) \in \widetilde{\mathbb{H}}$ and obtain

$$\|\xi^* + \nu(\mu - \xi^*)\|_2^2 \geq \|\xi^*\|_2^2$$

This leads to

$$(\xi^*)^\intercal \mu - (\xi^*)^\intercal \xi^* \geq 0$$

and hence

$$2(\xi^*)^\intercal \mu - (\xi^*)^\intercal \xi^* \geq (\xi^*)^\intercal \xi^*, \quad \text{for any} \quad \mu \in \widetilde{\mathbb{H}}$$

By taking ξ as ξ^* in the optimization problem (103), we have

$$\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} \left[2u^\intercal \xi - \xi^\intercal \xi \right] \geq \min_{u \in \widetilde{\mathbb{H}}} \left[2u^\intercal \xi^* - [\xi^*]^\intercal \xi^* \right] \geq (\xi^*)^\intercal \xi^*.$$

In (103), we take $u = \xi^*$, then we have

$$\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} \left[2u^\intercal \xi - \xi^\intercal \xi \right] \leq \max_{\xi \in \mathbb{R}^p} \left[2[\xi^*]^\intercal \xi - \xi^\intercal \xi \right] = (\xi^*)^\intercal \xi^*$$

By matching the above two bounds, ξ^* is the optimal solution to (103) and

$$\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} [2u^\intercal \xi - \xi^\intercal \xi] = [\xi^*]^\intercal \xi^*.$$

Since $\beta^* = C^{-1}\xi^*$ and $\Sigma^{\mathcal{Q}} = C^\intercal C$, we have

$$\beta^* = \arg\min_{\beta \in \mathbb{H}} \beta^\intercal \Sigma^{\mathcal{Q}} \beta \tag{105}$$

and

$$\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{H}} \left[2b^\intercal \Sigma^{\mathcal{Q}} \beta - \beta^\intercal \Sigma^{\mathcal{Q}} \beta\right] = [\beta^*]^\intercal \Sigma^{\mathcal{Q}} \beta^*.$$

We establish (6) by combining (105) and the fact that $\beta \in \mathbb{H}$ can be expressed as $\beta = \mathcal{B}\gamma$ for $\gamma \in \Delta^L$.

B.9 Proof of Proposition 4

We can write the maximin definition in the following form

$$\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{B}} \beta^\intercal b \tag{106}$$

where $\mathbb{B} = \{b^{(1)}, \ldots, b^{(L)}\}$. Since $b^\intercal \beta$ is linear in b, we can replace $\mathbb{B}$ with its convex hull $\mathbb{H}$,

$$\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta$$

We exchange the max and min in the above equation and have

$$\min_{b \in \mathbb{H}} \max_{\|\beta\|_2 \leq 1} b^\intercal \beta = \min_{b \in \mathbb{H}} \|b\|_2$$

We define

$$\xi = \arg\min_{b \in \mathbb{H}} \|b\|_2.$$

We claim that $\xi^* = \xi/\|\xi\|$ is the optimal solution of (106). For any $\mu \in \mathbb{H}$, we have $\xi + \nu(\mu - \xi) \in \mathbb{H}$ for $\nu \in [0,1]$ and have

$$\|\xi + \nu(\mu - \xi)\|_2^2 \geq \|\xi\|_2^2 \quad \text{for any} \quad \nu \in [0,1]$$

By taking $\nu \to 0$, we have

$$\mu^\intercal \xi - \|\xi\|_2^2 \geq 0$$

By dividing both sides by $\|\xi\|_2$, we have

$$\mu^\intercal \xi^* \geq \|\xi\|_2 \quad \text{for any} \quad \mu \in \mathbb{H}. \tag{107}$$

In the definition of (106), we take $\beta = \xi^*$ and have

$$\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta \geq \min_{b \in \mathbb{H}} b^\intercal \xi^* \geq \|\xi\|_2 \tag{108}$$

where the last inequality follows from (107) . Additionally, we take $b = \xi$ in the definition of (106) and have

$$\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta \leq \max_{\|\beta\|_2 \leq 1} \xi^\intercal \beta = \|\xi\|_2$$

Combined with (108), we have shown that

$$\xi^* = \arg \max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta$$

that is, $\beta^{*,\mathrm{MP}} = \xi^*$.

B.10 Proof of Proposition 5

The difference between the proposed estimator $\widehat{\Gamma}_{l,k}$ and $\Gamma_{l,k}$ is

$$\widehat{\Gamma}_{l,k} - \Gamma_{l,k} = D_{l,k}^{(1)} + D_{l,k}^{(2)} + \mathrm{Rem}_{l,k},$$

where $D_{l,k}^{(1)} = \frac{1}{n_k}(b^{(l)})^\intercal [X^{(k)}]^\intercal \epsilon^{(k)} + \frac{1}{n_l}(b^{(k)})^\intercal [X^{(l)}]^\intercal \epsilon^{(l)}$, $D_{l,k}^{(2)} = (b^{(l)})^\intercal (\widehat{\Sigma} - \Sigma) b^{(k)}$, and

$$\begin{aligned} \mathrm{Rem}_{l,k} = & \frac{1}{n_l}(\widehat{b}_{init}^{(k)} - b^{(k)})^\intercal [X^{(l)}]^\intercal \epsilon^{(l)} + \frac{1}{n_k}(\widehat{b}_{init}^{(l)} - b^{(l)})^\intercal [X^{(k)}]^\intercal \epsilon^{(k)} \\ & - (\widehat{b}_{init}^{(l)} - b^{(l)})^\intercal \widehat{\Sigma} (\widehat{b}_{init}^{(k)} - b^{(k)}) + [\widehat{b}_{init}^{(l)}]^\intercal (\widehat{\Sigma} - \widetilde{\Sigma}^{(k)})(\widehat{b}_{init}^{(k)} - b^{(k)}) + [\widehat{b}_{init}^{(k)}]^\intercal (\widehat{\Sigma} - \widetilde{\Sigma}^{(l)})(\widehat{b}_{init}^{(l)} - b^{(l)}) \end{aligned} \tag{109}$$

with $\widetilde{\Sigma}^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} X_{i,\cdot}^{(l)} [X_{i,\cdot}^{(l)}]^\intercal$ for $1 \leq l \leq L$ and

$$\widehat{\Sigma} = \frac{1}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}} \left(\sum_{l=1}^{L} \sum_{i=1}^{n_l} X_{i,\cdot}^{(l)} [X_{i,\cdot}^{(l)}]^\intercal + \sum_{i=1}^{N_{\mathcal{Q}}} X_{i,\cdot}^{(l)} [X_{i,\cdot}^{(l)}]^\intercal \right).$$

In the following, we shall control $D^{(1)}, D^{(2)}$ and $\mathrm{Rem}_{l,k}$ separately.

Control of $D^{(1)}$ and $D^{(2)}$. By the Gaussian error assumption, we show that $\mathrm{vecl}(D^{(1)}) \mid \{X^{(l)}\}_{1\leq l\leq L} \sim \mathcal{N}(0, \mathbf{V}^{(1)})$ with $\mathbf{V}^{(1)}$ defined in (39).

For $(l,k) \in \mathcal{I}_L$, we define $W_{i,\cdot} \in \mathbb{R}^{L(L+1)/2}$ as

$$W^{(j)}_{i,\pi(l,k)} = [b^{(l)}]^\intercal (X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^\intercal - \Sigma^{\mathcal{Q}})b^{(k)} \quad \text{for} \quad 1\leq j\leq L,\ 1\leq i\leq n_j$$

and

$$W^{\mathcal{Q}}_{i,\pi(l,k)} = [b^{(l)}]^\intercal (X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^\intercal - \Sigma^{\mathcal{Q}})b^{(k)} \quad \text{for} \quad 1\leq i\leq N_{\mathcal{Q}}$$

where the index mapping π is defined in (3). Then we have

$$\{W^{(j)}_{i,\pi(l,k)}\}_{1\leq j\leq L, 1\leq i\leq n_j} \quad \text{and} \quad \{W^{\mathcal{Q}}_{i,\pi(l,k)}\}_{1\leq i\leq N_{\mathcal{Q}}} \quad \text{are i.i.d. random variables.}$$

We can express the elements of $D^{(2)}$ as

$$[\mathrm{vecl}(D^{(2)})]_{\pi(l,k)} = D^{(2)}_{l,k} = \frac{1}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}}\left(\sum_{j=1}^{L}\sum_{i=1}^{n_j} W^{(j)}_{i,\pi(l,k)} + \sum_{i=1}^{N_{\mathcal{Q}}} W^{\mathcal{Q}}_{i,\pi(l,k)}\right).$$

The random variable $W^{(1)}_{1,\cdot}$ is of mean zero and covariance matrix $\mathbf{C}$ defined in (65). By (67), we have shown that $\mathbf{E}\|W^{(1)}_{1,\cdot}\|_2^{r_0} \leq C_0$ for some positive constants $r_0 \geq 3$ and $C_0 > 0$. By applying Lemma 2, we establish that there exist two sequences of independent random vectors $\{W^{(j),0}_{i,\cdot}\}_{1\leq j\leq L, 1\leq i\leq n_j}$ and $\{W^{\mathcal{Q},0}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}$ and $\{Z^{(j),0}_{i,\cdot}\}_{1\leq j\leq L, 1\leq i\leq n_j}$ and $\{Z^{\mathcal{Q},0}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}$ such that

$$Z^{(j),0}_{i,\cdot} \sim \mathcal{N}(0,\mathbf{C}) \quad \text{and} \quad Z^{\mathcal{Q},0}_{i,\cdot} \sim \mathcal{N}(0,\mathbf{C}), \text{ for } 1\leq j\leq L,$$
$$\{W^{(j),0}_{i,\cdot}\}_{1\leq i\leq n_j} \overset{d}{=} \{W^{(j)}_{i,\cdot}\}_{1\leq i\leq n_j} \quad \text{and} \quad \{W^{\mathcal{Q},0}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}} \overset{d}{=} \{W^{\mathcal{Q}}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}} \tag{110}$$

and

$$\left\|\sum_{j=1}^{L}\sum_{i=1}^{n_j} W^{(j),0}_{i,\pi(l,k)} + \sum_{i=1}^{N_{\mathcal{Q}}} W^{\mathcal{Q},0}_{i,\pi(l,k)} - \sum_{j=1}^{L}\sum_{i=1}^{n_j} Z^{(j),0}_{i,\pi(l,k)} - \sum_{i=1}^{N_{\mathcal{Q}}} Z^{\mathcal{Q},0}_{i,\pi(l,k)}\right\| \lesssim (\sum_{l=1}^{L} n_l + N_{\mathcal{Q}})^{\frac{1}{r_0}} \quad \text{a.s.} \tag{111}$$

We define

$$\mathrm{vecl}(D^{(2),*}) = \frac{1}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}}\left(\sum_{j=1}^{L}\sum_{i=1}^{n_j} W^{(j),0}_{i,\cdot} + \sum_{i=1}^{N_{\mathcal{Q}}} W^{\mathcal{Q},0}_{i,\cdot}\right)$$

and

$$T^* = \frac{1}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}} \left(\sum_{j=1}^{L} \sum_{i=1}^{n_j} Z_{i,\cdot}^{(j),0} + \sum_{i=1}^{N_{\mathcal{Q}}} Z_{i,\cdot}^{\mathcal{Q},0} \right) \sim \mathcal{N}(0, \mathbf{V}^{(2)})$$

with $\mathbf{V}^{(2)}$ defined in (40). As a consequence, the corresponding random matrix $D^{(2),*}$ has the same distribution with $D^{(2)}$ and $\|\text{vecl}(D^{(2),*}) - T^*\| \ll (\sum_{l=1}^{L} n_l + N_{\mathcal{Q}})^{1-\frac{1}{r_0}} \leq (\sum_{l=1}^{L} n_l + N_{\mathcal{Q}})^{-2/3}$ almost surely.

Control of $\text{Rem}_{l,k}$. We control $\text{Rem}_{l,k}$ by the definition of $\text{Rem}_{l,k}$ in (109) and the following Lemma.

Lemma 7. *With probability larger than* $1 - \min\{n, p\}^{-c}$, *we have*

$$\begin{aligned}
\left| \frac{1}{n_l} (\widehat{b}_{init}^{(k)} - b^{(k)})^\intercal [X^{(l)}]^\intercal \epsilon^{(l)} \right| &\lesssim \frac{s \log p}{n}; \\
\left| \frac{1}{n_k} (\widehat{b}_{init}^{(l)} - b^{(l)})^\intercal [X^{(k)}]^\intercal \epsilon^{(k)} \right| &\lesssim \frac{s \log p}{n}; \\
\left| (\widehat{b}_{init}^{(l)} - b^{(l)})^\intercal \widehat{\Sigma} (\widehat{b}_{init}^{(k)} - b^{(k)}) \right| &\lesssim \frac{s \log p}{n}. \\
\left| [\widehat{b}_{init}^{(l)}]^\intercal (\widehat{\Sigma} - \widetilde{\Sigma}^{(k)}) (\widehat{b}_{init}^{(k)} - b^{(k)}) \right| &\lesssim \left(\|b^{(l)}\|_2 + 1 \right) \frac{s \log p}{n} \\
\left| [\widehat{b}_{init}^{(k)}]^\intercal (\widehat{\Sigma} - \widetilde{\Sigma}^{(l)}) (\widehat{b}_{init}^{(l)} - b^{(l)}) \right| &\lesssim \left(\|b^{(k)}\|_2 + 1 \right) \frac{s \log p}{n}
\end{aligned}$$

Proof of (42). The control of $\mathbf{V}^{(1)}_{\pi(l,k),\pi(l,k)}$ follows from the definition (39) and the concentration result, for any $1 \leq l \leq L$,

$$\mathbf{P} \left([b^{(l)}]^\intercal \left(\frac{1}{n_l} (X^{(l)})^\intercal X^{(l)} - \Sigma^{(l)} \right) b^{(l)} \leq C \sqrt{\frac{\log p}{n_l}} \|b^{(l)}\|_2^2 \right) \geq 1 - p^{-c}$$

for some positive constant $C > 0$ and $c > 0$. The control of $\mathbf{V}^{(2)}_{\pi(l,k),\pi(l,k)}$ follows from (71) and (72).

B.11 Proof of Proposition 6

The proof of Proposition 6 is similar to that of Theorem 3 in Section B.6. We define similar events as those in (76) by modifying the definition of $\mathcal{E}_2$ as

$$\mathcal{E}_2 = \cup_{1 \leq l,k \leq L} \left\{ \text{Rem}_{l,k} \text{ satisfies } (41) \right\}$$

and re-defining $\mathbf{Cov}$ and $\widehat{\mathbf{Cov}}$ as

$$\mathbf{Cov} = n(\mathbf{V}^{(1)} + \mathbf{V}^{(2)}) \quad \text{and} \quad \widehat{\mathbf{Cov}} = n(\widehat{\mathbf{V}}^{(1)} + \widehat{\mathbf{V}}^{(2)}).$$

where $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$ are defined in (39) and (40), respectively and $\widehat{\mathbf{V}}^{(1)}$ and $\widehat{\mathbf{V}}^{(2)}$ are defined in (43) and (44), respectively. The key is to establish Lemmas 4 and 5 for the modified events $\mathcal{E}_2$ and $\mathcal{E}_3$. The proof of 4 follows the same argument as that of the general covariate shift setting in Section C.3 and is omitted here. We shall present a separate proof of Lemma 5 for the no covariate shift setting in Section C.4.

It follows from (42) in Proposition 5 and Condition (A2) that $n \cdot (\mathbf{V}^{(1)} + \mathbf{V}^{(2)})$ and d_0 are bounded from above with probability larger than $1 - \min\{n, p\}^{-c}$. Then for a finite L, we show that $C_1^*(L, \alpha_0) \geq c'$ for a small positive constant $c' > 0$ with probability larger than $1 - \min\{n, p\}^{-c}$.

C Additional Proofs

C.1 Proof of Lemma 1

On the event $\mathcal{G}_1 \cap \mathcal{G}_6(\widehat{b}^{(l)}_{init} - b^{(l)}, \widehat{b}^{(l)}_{init} - b^{(l)}, \sqrt{\log p})$, we have

$$\frac{1}{|B|}\sum_{i \in B}[(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)})]^2 \lesssim \frac{\|b^{(l)}\|_0 \log p}{n_l}\sigma_l^2.$$

Then we have

$$\begin{aligned}\left|(\widehat{b}^{(l)}_{init} - b^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}^{(k)}_{init} - b^{(k)})\right| &\leq \frac{1}{|B|}\|X^{\mathcal{Q}}_{B,\cdot}(\widehat{b}^{(l)}_{init} - b^{(l)})\|_2\|X^{\mathcal{Q}}_{B,\cdot}(\widehat{b}^{(k)}_{init} - b^{(k)})\|_2 \\ &\lesssim \sqrt{\frac{\|b^{(l)}\|_0\|b^{(k)}\|_0(\log p)^2}{n_l n_k}}\end{aligned}$$

and establish (60). We decompose

$$\begin{aligned}&(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)}) \\ &= (\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)}) + [\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}} - \widetilde{\Sigma}^{\mathcal{Q}})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)}).\end{aligned} \tag{112}$$

Regarding the first term of (112), we apply Hölder's inequality and establish

$$\left|(\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\leq\|\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)}\|_{\infty}\|\widehat{b}^{(l)}_{init}-b^{(l)}\|_1$$

By the optimization constraint (16), on the event $\mathcal{G}_2$, we have

$$\left|(\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}-\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\lesssim\|\omega^{(k)}\|_2\sqrt{\frac{\log p}{n_l}}\cdot\|b^{(l)}\|_0\sqrt{\frac{\log p}{n_l}}\tag{113}$$

Regarding the second term of (112), conditioning on $\widehat{b}^{(k)}_{init}$ and $\widehat{b}^{(l)}_{init}$, on the event $\mathcal{G}_6(\widehat{b}^{(k)}_{init},\widehat{b}^{(l)}_{init}-b^{(l)},\sqrt{\log p})$,

$$\left|[\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}}-\widetilde{\Sigma}^{\mathcal{Q}})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\lesssim\frac{\sqrt{\log p}}{\sqrt{N_{\mathcal{Q}}}}\|\widehat{b}^{(k)}_{init}\|_2\|\widehat{b}^{(l)}_{init}-b^{(l)}\|_2.$$

On the event $\mathcal{G}_1$, we further have

$$\left|[\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}}-\widetilde{\Sigma}^{\mathcal{Q}})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\lesssim\|\widehat{b}^{(k)}_{init}\|_2\sqrt{\frac{\|b^{(l)}\|_0(\log p)^2}{n_l N_{\mathcal{Q}}}}$$

Combined with (113), we establish (61). We establish (62) through applying the similar argument for (61) by exchanging the role of l and k. Together with (49), (51) and (52) with $t=\sqrt{\log p}$, we establish the lemma.

C.2 Proof of Lemma 7

On the event $\mathcal{G}_0\cap\mathcal{G}_2$, we apply the Hölder's inequality and establish

$$\left|\frac{1}{n_l}(\widehat{b}^{(k)}_{init}-b^{(k)})^{\intercal}[X^{(l)}]^{\intercal}\epsilon^{(l)}\right|\leq\|\widehat{b}^{(k)}_{init}-b^{(k)}\|_1\left\|\frac{1}{n_l}[X^{(l)}]^{\intercal}\epsilon^{(l)}\right\|_{\infty}\lesssim\frac{s\log p}{n}.$$

Similarly, we establish $\left|\frac{1}{n_k}(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}[X^{(k)}]^{\intercal}\epsilon^{(k)}\right|\lesssim s\log p/n$. We define the event

$$\mathcal{G}_5'=\left\{\max_{\substack{\mathcal{S}\subset[p],|\mathcal{S}|\leq s\\ \mathcal{T}\subset[p],|\mathcal{T}|\leq s}}\ \max_{\substack{\|w_{\mathcal{S}^c}\|_1\leq C\|w_{\mathcal{S}}\|_1\\ \|v_{\mathcal{T}^c}\|_1\leq C\|v_{\mathcal{T}}\|_1}}\frac{v^{\intercal}\left(\frac{1}{N}\sum_{i=1}^{N}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}\right)w}{\sqrt{v^{\intercal}E(X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}])v}\sqrt{w^{\intercal}E(X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}])w}}\leq 1+C\sqrt{\frac{s\log p}{N}}\right\}$$

It follows from Theorem 1.6 of Zhou (2009) that the event $\mathcal{G}_5'$ holds with probability larger than $1-p^{-c}$ for some positive constant $c>0$. On the event $\mathcal{G}_1\cap\mathcal{G}_5'$ with $N=\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}$, we have

$$\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\widehat{\Sigma}(\widehat{b}^{(k)}_{init}-b^{(k)})\right| \lesssim \frac{s\log p}{n}\cdot\left(1+\sqrt{\frac{s\log p}{\sum_{l=1}^{L} n_l+N_{\mathcal{Q}}}}\right). \tag{114}$$

Note that

$$\begin{aligned}
&[\widehat{b}^{(l)}_{init}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})\\
&=[\widehat{b}^{(l)}_{init}-b^{(l)}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})+[b^{(l)}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})\\
&=[\widehat{b}^{(l)}_{init}-b^{(l)}]^{\intercal}\widehat{\Sigma}(\widehat{b}^{(k)}_{init}-b^{(k)})+[b^{(l)}]^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\\
&-[\widehat{b}^{(l)}_{init}-b^{(l)}]^{\intercal}\widetilde{\Sigma}^{(k)}(\widehat{b}^{(k)}_{init}-b^{(k)})-[b^{(l)}]^{\intercal}(\widetilde{\Sigma}^{(k)}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})
\end{aligned} \tag{115}$$

With a similar proof for (114), we show that, on the event $\mathcal{G}_1\cap\mathcal{G}_5'$,

$$\left|[\widehat{b}^{(l)}_{init}-b^{(l)}]^{\intercal}\widetilde{\Sigma}^{(k)}(\widehat{b}^{(k)}_{init}-b^{(k)})\right| \lesssim \frac{s\log p}{n}\cdot\left(1+\sqrt{\frac{s\log p}{n}}\right). \tag{116}$$

By Hölder's inequality, we have

$$\begin{aligned}
\left|[b^{(l)}]^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right| &\leq \|(\widehat{\Sigma}-\Sigma)b^{(l)}\|_\infty\|\widehat{b}^{(k)}_{init}-b^{(k)}\|_1\\
\left|[b^{(l)}]^{\intercal}(\widetilde{\Sigma}^{(k)}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right| &\leq \|(\widetilde{\Sigma}^{(k)}-\Sigma)b^{(l)}\|_\infty\|\widehat{b}^{(k)}_{init}-b^{(k)}\|_1
\end{aligned}$$

We define

$$\mathcal{G}_6'(w,v,t)=\left\{\left|w^{\intercal}\left(\widehat{\Sigma}-\Sigma\right)v\right|\lesssim t\frac{\|\Sigma^{1/2}w\|_2\|\Sigma^{1/2}v\|_2}{\sqrt{\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}}},\ \left|w^{\intercal}\left(\widetilde{\Sigma}^{(k)}-\Sigma\right)v\right|\lesssim t\frac{\|\Sigma^{1/2}w\|_2\|\Sigma^{1/2}v\|_2}{\sqrt{n_k}}\right\}$$

and it follows from Lemma 10 in the supplement of Cai and Guo (2020) that

$$\mathbf{P}(\mathcal{G}_6'(w,v,t))\geq 1-\exp(-t^2).$$

On the event $\cap_{j=1}^{p}\mathcal{G}_6'(b^{(l)},e_j,\sqrt{\log p})\cap\mathcal{G}_1$, we have

$$\max\left\{\|(\widehat{\Sigma}-\Sigma)b^{(l)}\|_\infty,\|(\widetilde{\Sigma}^{(k)}-\Sigma)b^{(l)}\|_\infty\right\}\lesssim\sqrt{\frac{\log p}{n}}\|b^{(l)}\|_2$$

and

$$\max\left\{\left|[b^{(l)}]^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right|,\left|[b^{(l)}]^{\intercal}(\widetilde{\Sigma}^{(k)}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\right\}\lesssim\|b^{(l)}\|_2\cdot\frac{s\log p}{n}.$$

Together with (114), (115), (116), (49) and (51), we establish

$$\mathbf{P}\left(\left|[\widehat{b}^{(l)}_{init}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})\right| \lesssim \left(\|b^{(l)}\|_2+1\right)\frac{s\log p}{n}\right) \geq 1-\min\{n,p\}^{-c}.$$

We apply a similar argument to show that

$$\mathbf{P}\left(\left|[\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(l)})(\widehat{b}^{(l)}_{init}-b^{(l)})\right| \lesssim \left(\|b^{(k)}\|_2+1\right)\frac{s\log p}{n}\right) \geq 1-\min\{n,p\}^{-c}.$$

C.3 Proof of Lemma 4

The control of the event $\mathcal{E}_2$ follows from the proof of (30). In the following, we shall control the probability of the event $\mathcal{E}_1$.

For $\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined in (29), we express it as

$$\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} + \mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} \tag{117}$$

where $\mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined in (64) and

$$\mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \frac{1}{|B|}(\mathbf{E}[b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}-(b^{(l_1)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_1)}(b^{(l_2)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_2)}) \tag{118}$$

For $\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined in (21), we express it as

$$\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} + \widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} \tag{119}$$

with

$$\begin{aligned}\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} =& \frac{\widehat{\sigma}^2_{l_1}}{|B_{l_1}|}(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1)+\widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right] \\ &+\frac{\widehat{\sigma}^2_{k_1}}{|B_{k_1}|}(\widehat{u}^{(k_1,l_1)})^{\intercal}\widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=k_1)+\widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=k_1)\right]\end{aligned} \tag{120}$$

$$\widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \frac{\sum_{i=1}^{N_{\mathcal{Q}}}\left((\widehat{b}^{(l_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}-(\widehat{b}^{(l_1)}_{init})^{\intercal}\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^{\intercal}\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_2)}_{init}\right)}{|B|N_{\mathcal{Q}}} \tag{121}$$

where $\bar{\Sigma}^{\mathcal{Q}} = \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} X^{\mathcal{Q}}_{i,\cdot}(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}$ and $\widehat{\sigma}^2_l = \|Y^{(l)}-X^{(l)}\widehat{b}^{(l)}\|_2^2/n_l$ for $1 \leq l \leq L$.

The control of the event $\mathcal{E}_1$ follows from the following to high probability inequalities: with probability larger than $1-\exp(-cn)-\min\{N_{\mathcal{Q}},p\}^{-c}$ for some positive constant $c>0$,

$$n \cdot \left|\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \le C d_0 \left(\frac{s\log p}{n} + \sqrt{\frac{\log p}{n}}\right) \le \frac{d_0}{4}. \tag{122}$$

$$N_{\mathcal{Q}} \cdot \left|\widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \lesssim \log\max\{N_{\mathcal{Q}},p\}\sqrt{\frac{s\log p\log N_{\mathcal{Q}}}{n}} + \frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}. \tag{123}$$

The proofs of (122) and (123) are presented in Sections C.3.1 and C.3.2, respectively.

We combine (122) and (123) and establish

$$\begin{aligned}
\|\widehat{\mathbf{Cov}} - \mathbf{Cov}\|_2 &\lesssim \max_{(l_1,k_1),(l_2,k_2)\in\mathcal{I}_L} \left|\widehat{\mathbf{Cov}}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{Cov}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \\
&\le n \cdot \max_{(l_1,k_1),(l_2,k_2)\in\mathcal{I}_L} \left|\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \\
&\quad + n \cdot \max_{(l_1,k_1),(l_2,k_2)\in\mathcal{I}_L} \left|\widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \\
&\le \frac{d_0}{4} + \frac{\sqrt{n\cdot s[\log\max\{N_{\mathcal{Q}},p\}]^3}}{N_{\mathcal{Q}}} + \frac{n\cdot(\log N_{\mathcal{Q}})^{5/2}}{N_{\mathcal{Q}}^{3/2}} \le d_0/2,
\end{aligned}$$

where the first inequality holds for a finite L and the last inequality follows from $n < N_{\mathcal{Q}}^{4/3}$ and $\sqrt{s[\log\max\{N_{\mathcal{Q}},p\}]^3} \le cN_{\mathcal{Q}}/\sqrt{n}$.

C.3.1 Proof of (122)

$$\begin{aligned}
& n \cdot \left|\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \\
& \lesssim \left|\widehat{\sigma}^2_{l_1} - \sigma^2_{l_1}\right| (\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)} \left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right] \\
& + \left|\widehat{\sigma}^2_{k_1} - \sigma^2_{k_1}\right| (\widehat{u}^{(k_1,l_1)})^{\intercal}\widehat{\Sigma}^{(k_1)} \left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=k_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=k_1)\right]
\end{aligned} \tag{124}$$

Since

$$\begin{aligned}
& \left|(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)} \left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right]\right| \\
& \le \sqrt{(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,k_1)} \cdot (\widehat{u}^{(l_1,k_2)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,k_2)}} \\
& + \sqrt{(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,k_1)} \cdot (\widehat{u}^{(l_1,l_2)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,l_2)}}
\end{aligned} \tag{125}$$

we have

$$\left|(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)} \left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right]\right| \lesssim n \max_{(l,k)\in\mathcal{I}_L} \mathbf{V}^{(a)}_{\pi(l,k),\pi(l,k)} \lesssim d_0$$

Similarly, we have $\left|(\widehat{u}^{(k_1,l_1)})^\intercal\widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=k_1)+\widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=k_1)\right]\right|\lesssim d_0$. Hence, on the event $\mathcal{G}_3$, we establish (122).

C.3.2 Proof of (123)

Define

$$W_{i,1}=[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot},\ W_{i,2}=[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot},\ W_{i,3}=[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot},\ W_{i,4}=[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot};$$

and

$$\widehat{W}_{i,1}=(\widehat{b}^{(l_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot},\ \widehat{W}_{i,2}=(\widehat{b}^{(k_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot},\ \widehat{W}_{i,3}=(\widehat{b}^{(l_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot},\ \widehat{W}_{i,4}=(\widehat{b}^{(k_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}.$$

With the above definitions, we have

$$\begin{aligned}&\mathbf{E}[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}-(b^{(l_1)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_1)}(b^{(l_2)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_2)}\\&=\mathbf{E}\prod_{t=1}^{4}W_{i,t}-\mathbf{E}W_{i,1}W_{i,2}\cdot\mathbf{E}W_{i,3}W_{i,4}\end{aligned}\tag{126}$$

and

$$\begin{aligned}&\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left((\widehat{b}^{(l_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}-(\widehat{b}^{(l_1)}_{init})^\intercal\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^\intercal\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_2)}_{init}\right)\\&=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}\cdot\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,3}\widehat{W}_{i,4}\end{aligned}\tag{127}$$

Hence, it is sufficient to control the following terms.

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\mathbf{E}\prod_{t=1}^{4}W_{i,t}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}-\mathbf{E}\prod_{t=1}^{4}W_{i,t}$$

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}-\mathbf{E}W_{i,1}W_{i,2}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,1}W_{i,2}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,1}W_{i,2}-\mathbf{E}W_{i,1}W_{i,2}$$

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,3}\widehat{W}_{i,4}-\mathbf{E}W_{i,3}W_{i,4}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,3}\widehat{W}_{i,4}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,3}W_{i,4}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,3}W_{i,4}-\mathbf{E}W_{i,3}W_{i,4}$$

Specifically, we will show that, with probability larger than $1-\min\{N_{\mathcal{Q}},p\}^{-c}$,

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} W_{i,1}W_{i,2}-\mathbf{E}W_{i,1}W_{i,2}\right| \lesssim \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}}, \tag{128}$$

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} W_{i,3}W_{i,4}-\mathbf{E}W_{i,3}W_{i,4}\right| \lesssim \|b^{(l_2)}\|_2\|b^{(k_2)}\|_2\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}}, \tag{129}$$

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(\prod_{t=1}^{4} W_{i,t}-\mathbf{E}\prod_{t=1}^{4} W_{i,t}\right) \lesssim \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}, \tag{130}$$

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,1}W_{i,2}\right| \lesssim \sqrt{\frac{s\log p}{n}}\left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_1)}\|_2+\|b^{(k_1)}\|_2)+\sqrt{\frac{s\log p}{n}}\right), \tag{131}$$

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,3}\widehat{W}_{i,4}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,3}W_{i,4}\right| \lesssim \sqrt{\frac{s\log p}{n}}\left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_2)}\|_2+\|b^{(k_2)}\|_2)+\sqrt{\frac{s\log p}{n}}\right). \tag{132}$$

If we further assume that $\|b^{(l)}\|_2\le C$ for $1\le l\le L$ and $s^2(\log p)^2/n\le c$ for some positive constants $C>0$ and $c>0$, then with probability larger than $1-\min\{N_{\mathcal{Q}},p\}^{-c}$,

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}\right| \lesssim \log\max\{N_{\mathcal{Q}},p\}\sqrt{\frac{s\log p\log N_{\mathcal{Q}}}{n}}. \tag{133}$$

By the expression (126) and (127), we establish (123) by applying (128), (129), (130), (131), (132), (133). In the following, we prove (128), (129) and (130). Then we will present the proofs of (131), (132), (133).

Proofs of (128), (129) and (130). We shall apply the following lemma to control the above terms, which re-states the Lemma 1 in Cai and Liu (2011).

Lemma 8. *Let $\xi_1,\cdots,\xi_n$ be independent random variables with mean 0. Suppose that there exists some $c>0$ and U_n such that $\sum_{i=1}^n\mathbf{E}\xi_i^2\exp(c|\xi_i|)\le U_n^2$. Then for $0<t\le U_n$,*

$$\mathbf{P}\left(\sum_{i=1}^n\xi_i\ge CU_nt\right)\le\exp(-t^2), \tag{134}$$

where $C=c+c^{-1}$.

Define

$$W_{i,1}^{0}=\frac{[b^{(l_1)}]^{\intercal}X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(l_1)}]^{\intercal}\Sigma^{\mathcal{Q}}b^{(l_1)}}},\ W_{i,2}^{0}=\frac{[b^{(k_1)}]^{\intercal}X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(k_1)}]^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_1)}}}$$

and

$$W_{i,3}^{0}=\frac{[b^{(l_2)}]^{\intercal}X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(l_2)}]^{\intercal}\Sigma^{\mathcal{Q}}b^{(l_2)}}}\ W_{i,4}^{0}=\frac{[b^{(k_2)}]^{\intercal}X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(k_2)}]^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_2)}}}$$

Since $X_{i,\cdot}^{\mathcal{Q}}$ is sub-gaussian, $W_{i,t}^{0}$ is sub-gaussian and both $W_{i,1}^{0}W_{i,2}^{0}$ and $W_{i,3}^{0}W_{i,4}^{0}$ are sub-exponetial random variables, which follows from Remark 5.18 in Vershynin (2012). By Corollary 5.17 in Vershynin (2012), we have

$$\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(W_{i,1}^{0}W_{i,2}^{0}-\mathbf{E}W_{i,1}^{0}W_{i,2}^{0}\right)\right|\geq C\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}}\right)\leq 2N_{\mathcal{Q}}^{-c}$$

and

$$\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(W_{i,3}^{0}W_{i,4}^{0}-\mathbf{E}W_{i,3}^{0}W_{i,4}^{0}\right)\right|\geq C\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}}\right)\leq 2N_{\mathcal{Q}}^{-c}$$

where c and C are positive constants. The above inequalities imply (128) and (129) after rescaling.

For $1\leq t\leq 4$, since $W_{i,t}^{0}$ is a sub-gaussian random variable, there exist positive constants $C_1>0$ and $c>2$ such that the following concentration inequality holds,

$$\sum_{i=1}^{N_{\mathcal{Q}}}\mathbf{P}\left(\max_{1\leq t\leq 4}|W_{i,t}^{0}|\geq C_1\sqrt{\log N_{\mathcal{Q}}}\right)\leq N_{\mathcal{Q}}\max_{1\leq i\leq N_{\mathcal{Q}}}\mathbf{P}\left(\max_{1\leq t\leq 4}|W_{i,t}^{0}|\geq C_1\sqrt{\log N_{\mathcal{Q}}}\right)\lesssim N_{\mathcal{Q}}^{-c} \tag{135}$$

Define

$$H_{i,a}=\prod_{t=1}^{4}W_{i,t}^{0}\cdot\mathbf{1}\left(\max_{1\leq t\leq 4}|W_{i,t}^{0}|\leq C_1\sqrt{\log N_{\mathcal{Q}}}\right)\quad\text{for}\quad 1\leq t\leq 4,$$

and

$$H_{i,b}=\prod_{t=1}^{4}W_{i,t}^{0}\cdot\mathbf{1}\left(\max_{1\leq t\leq 4}|W_{i,t}^{0}|\geq C_1\sqrt{\log N_{\mathcal{Q}}}\right)\quad\text{for}\quad 1\leq t\leq 4.$$

Then we have

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}^{0}-\mathbf{E}\prod_{t=1}^{4}W_{i,t}^{0}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}(H_{i,a}-\mathbf{E}H_{i,a})+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}(H_{i,b}-\mathbf{E}H_{i,b}) \tag{136}$$

By applying the Cauchy-Schwarz inequality, we bound $\mathbf{E}H_{i,b}$ as

$$
\begin{aligned}
|\mathbf{E}H_{i,b}| &\leq \sqrt{\mathbf{E}\left(\prod_{t=1}^{4} W_{i,t}^{0}\right)^{2} \mathbf{P}\left(\max_{1\leq t\leq 4} |W_{i,t}^{0}| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right)} \\
&\lesssim \mathbf{P}\left(|W_{i,t}^{0}| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right)^{1/2} \lesssim N_{\mathcal{Q}}^{-1/2},
\end{aligned}
\tag{137}
$$

where the second and the last inequalities follow from the fact that $W_{i,t}^{0}$ is a sub-gaussian random variable. Now we apply Lemma 8 to bound $\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}(H_{i,a} - \mathbf{E}H_{i,a})$. By taking $c = c_1/(C_1^2 \log N_{\mathcal{Q}})^2$ for some small positive constant $c_1 > 0$, we have

$$
\sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{E}\left(H_{i,a} - \mathbf{E}H_{i,a}\right)^2 \exp\left(c\left|H_{i,a} - \mathbf{E}H_{i,a}\right|\right) \leq C\sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{E}\left(H_{i,a} - \mathbf{E}H_{i,a}\right)^2 \leq C_2 N_{\mathcal{Q}}.
$$

By applying Lemma 8 with $U_n = \sqrt{C_2 N_{\mathcal{Q}}}$, $c = c_1/(C_1^2 \log N_{\mathcal{Q}})^2$ and $t = \sqrt{\log N_{\mathcal{Q}}}$, then we have

$$
\mathbf{P}\left(\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}(H_{i,a} - \mathbf{E}H_{i,a}) \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \lesssim N_{\mathcal{Q}}^{-c}.
\tag{138}
$$

Note that

$$
\begin{aligned}
\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} H_{i,b}\right| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) &\leq \mathbf{P}\left(\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} |H_{i,b}| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \\
&\leq \sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{P}\left(|H_{i,b}| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \\
&\leq \sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{P}\left(\max_{1\leq t\leq 4} |W_{i,t}^{0}| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right) \lesssim N_{\mathcal{Q}}^{-c}
\end{aligned}
\tag{139}
$$

where the last inequality follows from (135).

By the decomposition (136), we have

$$
\begin{aligned}
&\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(\prod_{t=1}^{4}W_{i,t}^{0}-\mathbf{E}\prod_{t=1}^{4}W_{i,t}^{0}\right)\right|\geq 3C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right)\\
&\leq \mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}(H_{i,a}-\mathbf{E}H_{i,a})\right|\geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right)\\
&+\mathbf{P}\left(|\mathbf{E}H_{i,b}|\geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right)+\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}H_{i,b}\right|\geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right)\lesssim N_{\mathcal{Q}}^{-c}.
\end{aligned}
$$

where the final upper bound follows from (137), (138) and (139). Hence, we establish that (130) holds with probability larger than $1-N_{\mathcal{Q}}^{-c}$.

Proofs (131), (132) and (133). It follows from the definitions of $\widehat{W}_{i,t}$ and $W_{i,t}$ that

$$
\begin{aligned}
&\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,1}W_{i,2}=[\widehat{b}_{init}^{(l_1)}-b^{(l_1)}]^{\intercal}\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}[\widehat{b}_{init}^{(k_1)}-b^{(k_1)}]\\
&+[b^{(l_1)}]^{\intercal}\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}[\widehat{b}_{init}^{(k_1)}-b^{(k_1)}]+[\widehat{b}_{init}^{(l_1)}-b^{(l_1)}]^{\intercal}\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}b^{(k_1)}
\end{aligned}
\tag{140}
$$

On the event $\mathcal{G}_2\cap\mathcal{G}_5$ with $\mathcal{G}_2$ defined in (48) and $\mathcal{G}_5$ defined in (50), we establish (131). By a similar argument, we establish (132). Furthermore, we define the event

$$
\begin{aligned}
\mathcal{G}_7&=\left\{\max_{1\leq l\leq L}\max_{1\leq i\leq N_{\mathcal{Q}}}\left|X_{i,\cdot}^{\mathcal{Q}}b^{(l)}\right|\lesssim(\sqrt{C_0}+\sqrt{\log N_{\mathcal{Q}}})\|b^{(l)}\|_2\right\}\\
\mathcal{G}_8&=\left\{\max_{1\leq i\leq N_{\mathcal{Q}}}\|X_{i,\cdot}^{\mathcal{Q}}\|_{\infty}\lesssim(\sqrt{C_0}+\sqrt{\log N_{\mathcal{Q}}+\log p})\right\}
\end{aligned}
\tag{141}
$$

It follows from the assumption (A1) that $\mathbf{P}(\mathcal{G}_7)\geq 1-N_{\mathcal{Q}}^{-c}$ and $\mathbf{P}(\mathcal{G}_8)\geq 1-\min\{N_{\mathcal{Q}},p\}^{-c}$ for some positive constant $c>0$. Note that

$$
\begin{aligned}
&\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|\widehat{W}_{i,1}\widehat{W}_{i,2}-W_{i,1}W_{i,2}\right|\leq\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[\widehat{b}_{init}^{(l_1)}-b^{(l_1)}]^{\intercal}X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}[\widehat{b}_{init}^{(k_1)}-b^{(k_1)}]\right|\\
&+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[b^{(l_1)}]^{\intercal}X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}[\widehat{b}_{init}^{(k_1)}-b^{(k_1)}]\right|+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[\widehat{b}_{init}^{(l_1)}-b^{(l_1)}]^{\intercal}X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}b^{(k_1)}\right|
\end{aligned}
\tag{142}
$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[\widehat{b}^{(l_1)}_{init}-b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init}-b^{(k_1)}]\right| \\
&\leq \frac{1}{N_{\mathcal{Q}}}\sqrt{\sum_{i=1}^{N_{\mathcal{Q}}}\left([\widehat{b}^{(l_1)}_{init}-b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}\right)^2\sum_{i=1}^{N_{\mathcal{Q}}}\left([X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init}-b^{(k_1)}]\right)^2}
\end{aligned}$$

Hence, on the event $\mathcal{G}_1\cap\mathcal{G}_6(\widehat{b}^{(k_1)}_{init}-b^{(k_1)},\widehat{b}^{(k_1)}_{init}-b^{(k_1)},\sqrt{\log p})\cap\mathcal{G}_6(\widehat{b}^{(l_1)}_{init}-b^{(l_1)},\widehat{b}^{(l_1)}_{init}-b^{(l_1)},\sqrt{\log p})$,

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[\widehat{b}^{(l_1)}_{init}-b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init}-b^{(k_1)}]\right| \lesssim \frac{s\log p}{n} \tag{143}$$

On the event $\mathcal{G}_7$, we have

$$\begin{aligned}
&\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init}-b^{(k_1)}]\right| \\
&\lesssim (\sqrt{C_0}+\sqrt{\log N_{\mathcal{Q}}})\|b^{(l_1)}\|_2\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init}-b^{(k_1)}]\right| \\
&\leq (\sqrt{C_0}+\sqrt{\log N_{\mathcal{Q}}})\|b^{(l_1)}\|_2\frac{1}{\sqrt{N_{\mathcal{Q}}}}\sqrt{\sum_{i=1}^{N_{\mathcal{Q}}}\left([X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init}-b^{(k_1)}]\right)^2}
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Hence, on the event $\mathcal{G}_1\cap\mathcal{G}_7\cap\mathcal{G}_6(\widehat{b}^{(k_1)}_{init}-b^{(k_1)},\widehat{b}^{(k_1)}_{init}-b^{(k_1)},\sqrt{\log p})$, we establish

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init}-b^{(k_1)}]\right| \lesssim (\sqrt{C_0}+\sqrt{\log N_{\mathcal{Q}}})\|b^{(l_1)}\|_2\sqrt{\frac{s\log p}{n}}. \tag{144}$$

Similarly, we establish

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|[\widehat{b}^{(l_1)}_{init}-b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}b^{(k_1)}\right| \lesssim (\sqrt{C_0}+\sqrt{\log N_{\mathcal{Q}}})\|b^{(k_1)}\|_2\sqrt{\frac{s\log p}{n}}.$$

Combined with (142), (143) and (144), we establish

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|\widehat{W}_{i,1}\widehat{W}_{i,2}-W_{i,1}W_{i,2}\right| \lesssim \left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_1)}\|_2+\|b^{(k_1)}\|_2)+\sqrt{\frac{s\log p}{n}}\right)\sqrt{\frac{s\log p}{n}} \tag{145}$$

Similarly, we establish

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left|\widehat{W}_{i,3}\widehat{W}_{i,4}-W_{i,3}W_{i,4}\right| \lesssim \left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_2)}\|_2+\|b^{(k_2)}\|_2)+\sqrt{\frac{s\log p}{n}}\right)\sqrt{\frac{s\log p}{n}} \tag{146}$$

Define $H_{i,1}=W_{i,1}W_{i,2}, H_{i,2}=W_{i,3}W_{i,4}, \widehat{H}_{i,1}=\widehat{W}_{i,1}\widehat{W}_{i,2}$ and $\widehat{H}_{i,2}=\widehat{W}_{i,3}\widehat{W}_{i,4}$. Then we have

$$\begin{aligned}&\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{H}_{i,1}\widehat{H}_{i,2}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}H_{i,1}H_{i,2}\\&=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(\widehat{H}_{i,1}-H_{i,1}\right)H_{i,2}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(\widehat{H}_{i,2}-H_{i,2}\right)H_{i,1}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(\widehat{H}_{i,1}-H_{i,1}\right)\left(\widehat{H}_{i,2}-H_{i,2}\right)\end{aligned} \tag{147}$$

On the event $\mathcal{G}_7$, we have

$$|H_{i,1}|\lesssim(C_0+\log N_{\mathcal{Q}})\|b^{(l_1)}\|_2\|b^{(k_1)}\|_2 \quad \text{and} \quad |H_{i,2}|\lesssim(C_0+\log N_{\mathcal{Q}})\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2 \tag{148}$$

On the event $\mathcal{G}_7\cap\mathcal{G}_8$, we have

$$\begin{aligned}&\left|\widehat{H}_{i,2}-H_{i,2}\right|\leq\left|[\widehat{b}^{(l_2)}_{init}-b^{(l_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_2)}_{init}-b^{(k_2)}]\right|\\&+\left|[b^{(l_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_2)}_{init}-b^{(k_2)}]\right|+\left|[\widehat{b}^{(l_2)}_{init}-b^{(l_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}b^{(k_2)}\right|\\&\lesssim(C_0+\log N_{\mathcal{Q}}+\log p)\left(s^2\frac{\log p}{n}+s\sqrt{\frac{\log p}{n}}\|b^{(k_2)}\|_2+s\sqrt{\frac{\log p}{n}}\|b^{(l_2)}\|_2\right)\end{aligned} \tag{149}$$

By the decomposition (147), we combine (148), (149), (145) and (146) and establish

$$
\begin{aligned}
&\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}\right|\\
&\leq (C_0+\log N_{\mathcal{Q}})\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2\left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_1)}\|_2+\|b^{(k_1)}\|_2)+\sqrt{\frac{s\log p}{n}}\right)\sqrt{\frac{s\log p}{n}}\\
&+(C_0+\log N_{\mathcal{Q}})\|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_2)}\|_2+\|b^{(k_2)}\|_2)+\sqrt{\frac{s\log p}{n}}\right)\sqrt{\frac{s\log p}{n}}\\
&+\left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_1)}\|_2+\|b^{(k_1)}\|_2)+\sqrt{\frac{s\log p}{n}}\right)\sqrt{\frac{s\log p}{n}}\\
&\cdot(C_0+\log N_{\mathcal{Q}}+\log p)\left(s^2\frac{\log p}{n}+s\sqrt{\frac{\log p}{n}}\|b^{(k_2)}\|_2+s\sqrt{\frac{\log p}{n}}\|b^{(l_2)}\|_2+\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2\right)
\end{aligned}
\tag{150}
$$

If we further assume that $\|b^{(l)}\|_2 \leq C$ for $1 \leq l \leq L$ and $s^2(\log p)^2/n \leq c$ for some positive constants $C > 0$ and $c > 0$, then we establish (133).

C.4 Proof of Lemma 5

We shall divide the proof into two parts based on whether there is possible covariate shift or not. The proofs are slightly different but the main idea remains the same.

C.4.1 The setting with possible covariate shift.

Note that, for any vectors u_1 and u_2 and any positive constant $c > 0$, we have

$$
(u_1+u_2)^{\intercal}W(u_1+u_2)\leq(1+c)u_1^{\intercal}Wu_1+\left(1+\frac{1}{c}\right)u_2^{\intercal}Wu_2 \tag{151}
$$

where W is a positive semi-definite matrix. We take $W = (\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}$, $u_1 = \sqrt{n}\text{vecl}(D)$ and $u_2 = \sqrt{n}\text{vecl}(\text{Rem})$ with D and Rem defined in Theorem 2. Then have $u_1 + u_2 = \widehat{Z} = \sqrt{n}(\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}})$ and the following lower bound,

$$
\begin{aligned}
&\exp\left(-\frac{1}{2}\widehat{Z}^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\geq\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D)]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(D)]\right)\\
&\cdot\exp\left(-\left(1+\frac{1}{c}\right)\frac{1}{2}n[\text{vecl}(\text{Rem})]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(\text{Rem})]\right),
\end{aligned}
\tag{152}
$$

where $c > 0$ is a small positive constant. Note that the approximation error in (152) can be controlled as

$$\begin{aligned} & n[\mathrm{vecl}(\mathrm{Rem})]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(\mathrm{Rem})]\mathbf{1}_{\mathcal{O}\in\mathcal{E}_2} \\ & \lesssim \frac{n}{d_0}\|\mathrm{Rem}\|_F^2 \lesssim L^2 \cdot \left(\frac{s(\log p)^2}{N_{\mathcal{Q}}} + \frac{(s\log p)^2}{n}\right) \end{aligned} \tag{153}$$

where the last inequality follows from (30), Proposition 3 and the definition of d_0 in (22).

Additionally, we take $\mathrm{vecl}(D^*)$ and S^* as in Theorem 2 and then have

$$\begin{aligned} & \exp\left(-(1+c)\frac{1}{2}n[\mathrm{vecl}(D^*)]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^*)\right) \\ & \geq \exp\left(-(1+c)^2\frac{1}{2}n[S^*]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[S^*]\right) \\ & \cdot \exp\left(-(1+c)\left(1+\frac{1}{c}\right)\frac{1}{2}n[\mathrm{vecl}(D)^* - S^*]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}(\mathrm{vecl}(D^*) - S^*)\right) \end{aligned} \tag{154}$$

where the last inequality follows from (151). Note that the approximation error in (154) can be controlled as

$$n[\mathrm{vecl}(D^*) - S^*]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}(\mathrm{vecl}(D^*) - S^*) \lesssim \frac{n}{N_{\mathcal{Q}}^{4/3}} \quad \text{a.s.} \tag{155}$$

With r denoting the rank of $\mathbf{Cov}$, we conduct the eigen-decomposition

$$\mathbf{Cov} = \begin{pmatrix} U & U_c \end{pmatrix} \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} U & U_c \end{pmatrix}^{\intercal}$$

where $\Lambda = \mathrm{diag}(\Lambda_{1,1}, \cdots, \Lambda_{r,r})$ with $\Lambda_{1,1} \geq \cdots \geq \Lambda_{r,r} > 0$ and $U \in \mathbb{R}^{L(L+1)/2\times r}$ denotes the matrix of eigenvectors corresponding to $\Lambda_{1,1} \geq \cdots \geq \Lambda_{r,r} > 0$, and $U_c \in \mathbb{R}^{L(L+1)/2\times(L(L+1)/2-r)}$ denotes the matrix of eigenvectors corresponding to zero eigenvalues. Hence, it follows from Theorem 2 that, conditioning on $X^{\mathcal{Q}}_{A,\cdot}, \{X^{(l)}, \epsilon^{(l)}_{A_l}\}_{1\leq l\leq L}$, the transformed random vector $\sqrt{n}\Lambda^{-1/2}U^{\intercal}S^* \in \mathbb{R}^m$ is Gaussian with zero mean and the diagonal covariance matrix $\mathrm{I} \in \mathbb{R}^{r\times r}$ and $\sqrt{n}U_c^{\intercal}S^*$ is Gaussian with zero mean and covariance. As a consequence, their marginal distributions remain the same, that is, $\sqrt{n}\Lambda^{-1/2}U^{\intercal}S^* \in \mathbb{R}^m$ is Gaussian with zero mean and the diagonal covariance matrix $\mathrm{I} \in \mathbb{R}^{r\times r}$ and $\sqrt{n}U_c^{\intercal}S^*$ is zero almost surely. For any given $0 < \alpha_0 < 1/2$, we have

$$\mathbf{P}\left(-n\frac{1}{2}[U^{\intercal}S^*]^{\intercal}\Lambda^{-1}U^{\intercal}S^* \geq -\frac{1}{2}F^{-1}_{\chi^2_r}(1-\alpha_0)\right) = 1 - \alpha_0, \tag{156}$$

where $F^{-1}_{\chi^2_r}(1-\alpha_0)$ denotes the $1-\alpha_0$ quantile of the χ^2 distribution with degree of freedom r. We define the positive constant

$$c_{\alpha_0} = \exp\left(-(1+c)^2\frac{1}{2}F^{-1}_{\chi^2_r}\left(1-\alpha_0\right)\right). \tag{157}$$

We now apply (152), (153) and the complexity condition (A2) and establish that, there exists a small positive constant $c_1 > 0$,

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq (1+c_1)\cdot\frac{c_{\alpha_0}}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(D)]\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_{\alpha_0}}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_{\alpha_0}\right)\\
&-\mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(D)]\right)\mathbf{1}_{\mathcal{O}\notin\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_{\alpha_0}}{2}\right)
\end{aligned}$$

where the second inequality follows from the union bound. By the above inequality and

$$\mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(D)]\right)\mathbf{1}_{\mathcal{O}\notin\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_{\alpha_0}}{2}\right) \leq \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c),$$

we establish

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq (1+c_1)\cdot\frac{c_{\alpha_0}}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_{\alpha_0}\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)\\
&= \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D^*)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\text{vecl}(D^*)\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_{\alpha_0}\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)^2\frac{1}{2}n[S^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}S^*\right) \geq c_{\alpha_0}\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)
\end{aligned} \tag{158}$$

where the equality follows from $[\text{vecl}(D)] \overset{d}{=} \text{vecl}(D^*)$ and the last inequality follows from (154), (155) and the complexity condition (33). Hence, it is sufficient to control the lower bound in (158). Note that

$$
\begin{aligned}
&\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[S^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[S^*]\right)\geq c_{\alpha_0}\right)\\
&=\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[S^*]^\intercal\left(\begin{pmatrix}U & U_c\end{pmatrix}\begin{pmatrix}\Lambda & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}U & U_c\end{pmatrix}^\intercal+\frac{1}{2}d_0\mathbf{I}\right)^{-1}S^*\right)\geq c_{\alpha_0}\right)\\
&=\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[U^\intercal S^*]^\intercal(\Lambda+\frac{1}{2}d_0\mathbf{I})^{-1}U^\intercal S^*\right)\geq c_{\alpha_0}\right)\\
&\geq\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[U^\intercal S^*]^\intercal\Lambda^{-1}U^\intercal S^*\right)\geq c_{\alpha_0}\right)=1-\alpha_0,
\end{aligned}
\tag{159}
$$

where the first equality follows from the eigen-decomposition of $\mathbf{Cov}$, the second equality follows from the fact that $\sqrt{n}U_c^\intercal S^*$ is zero almost surely and the last equality follows from the definition of c_{α_0} in (157). We establish (79) by combining (158), (157) and (159) with $c=\sqrt{2}-1$.

C.4.2 The setting with no covariate shift.

The proof is similar to the setting with possible covariate shift. We mainly highlight the differences in the proof. We recall $\mathbf{V}=\mathbf{V}^{(1)}+\mathbf{V}^{(2)}$ and define $\mathbf{Cov}=n\cdot\mathbf{V}$ and

$$
\mathbf{Cov}^{(1)}=n\cdot\mathbf{V}^{(1)}\quad\text{and}\quad\mathbf{Cov}^{(2)}=n\cdot\mathbf{V}^{(2)}.
$$

Similarly to (152), we establish

$$
\begin{aligned}
&\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\\
&\geq\exp\left(-(1+c)n\frac{1}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\\
&\cdot\exp\left(-\left(1+\frac{1}{c}\right)\frac{1}{2}n[\mathrm{vecl}(\mathrm{Rem})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(\mathrm{Rem})]\right),
\end{aligned}
\tag{160}
$$

where $c>0$ is a small positive constant. Similarly to (153), we control the last term on the right hand side of (160) by

$$
n[\mathrm{vecl}(\mathrm{Rem})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(\mathrm{Rem})]\mathbf{1}_{\mathcal{O}\in\mathcal{E}_2}\lesssim\frac{n}{d_0}\|\mathrm{Rem}\|_F^2\lesssim L^2\cdot\frac{(s\log p)^2}{n}
\tag{161}
$$

where the last inequality follows from Proposition 5. We now apply (160), (161) and the complexity condition (A2) and establish that, for any constant $c_2 > 0$, there exists a small positive constant $c_1 > 0$,

$$
\begin{aligned}
&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq (1+c_1)\cdot\frac{c_2}{2}\right) \\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{n}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_2}{2}\right) \\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{n}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_2\right) \\
&- \mathbf{P}\left(\exp\left(-(1+c)\frac{n}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \mathbf{1}_{\mathcal{O}\notin\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_2}{2}\right) \\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{n}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_2\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)
\end{aligned}
\tag{162}
$$

where the second inequality follows from the union bound.

We now control the term $\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)$ by the following inequality,

$$
\begin{aligned}
&\exp\left(-(1+c)n\frac{1}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \\
&\geq \exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D^{(1)})]\right) \\
&\cdot \exp\left(-(1+c)n[\mathrm{vecl}(D^{(2)})]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D^{(2)})]\right)
\end{aligned}
\tag{163}
$$

Additionally, we take $\mathrm{vecl}(D^{(2),*})$ and T^* as in Proposition 5 and then have

$$
\begin{aligned}
&\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2)})]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D^{(2)})]\right) \\
&\overset{d}{=} \exp\left(-(1+c)n[\mathrm{vecl}(D^{(2),*})]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(2),*})\right) \\
&\geq \exp\left(-(1+c)^2 n[T^*]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[T^*]\right) \\
&\cdot \exp\left(-(1+c)\left(1+\frac{1}{c}\right)n[\mathrm{vecl}(D^{(2),*}) - T^*]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}(\mathrm{vecl}(D^{(2),*}) - T^*)\right)
\end{aligned}
\tag{164}
$$

where the last inequality follows from (151). Note that the approximation error in (164) can be controlled as

$$n[\mathrm{vecl}(D^{(2),*})-T^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}(\mathrm{vecl}(D^{(2),*})-T^*)\lesssim\frac{n}{(\sum_{l=1}^{n}n_l+N_{\mathcal{Q}})^{4/3}}\quad\text{a.s.}\tag{165}$$

For $j=1,2$, with r_j denoting the rank of $\mathbf{Cov}^{(j)}$, we conduct the eigen-decomposition

$$\mathbf{Cov}^{(j)}=\begin{pmatrix}U^{(j)} & U_c^{(j)}\end{pmatrix}\begin{pmatrix}\Lambda^{(j)} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}U^{(j)} & U_c^{(j)}\end{pmatrix}^\intercal$$

where $\Lambda^{(j)}=\mathrm{diag}(\Lambda^{(j)}_{1,1},\cdots,\Lambda^{(j)}_{r_j,r_j})$ with $\Lambda^{(j)}_{1,1}\geq\cdots\geq\Lambda^{(j)}_{r_j,r_j}>0$ and $U^{(j)}\in\mathbb{R}^{L(L+1)/2\times r_j}$ denotes the matrix of eigenvectors corresponding to $\Lambda^{(j)}_{1,1}\geq\cdots\geq\Lambda^{(j)}_{r_j,r_j}>0$, and $U^{(j)}_c\in\mathbb{R}^{L(L+1)/2\times(L(L+1)/2-r_j)}$ denotes the matrix of eigenvectors corresponding to zero eigenvalues. We apply the same argument for (156) and establish that

$$\mathbf{P}\left(-n[\mathrm{vecl}(D^{(1)})]^\intercal U^{(1)}[\Lambda^{(1)}]^{-1}[U^{(1)}]^\intercal\mathrm{vecl}(D^{(1)})\leq-F^{-1}_{\chi^2_{r_1}}\left(1-\frac{\alpha_0}{2}\right)\right)\leq\frac{\alpha_0}{2},\tag{166}$$

$$\mathbf{P}\left(-n(T^*)^\intercal U^{(2)}[\Lambda^{(2)}]^{-1}[U^{(2)}]^\intercal T^*\leq-F^{-1}_{\chi^2_{r_2}}\left(1-\frac{\alpha_0}{2}\right)\right)\leq\frac{\alpha_0}{2},\tag{167}$$

where $F^{-1}_{\chi^2_r}\left(1-\frac{\alpha_0}{2}\right)$ denote the $1-\frac{\alpha_0}{2}$ quantile of the χ^2 distribution with degree of freedom r. Furthermore, $[U_c^{(1)}]^\intercal\mathrm{vecl}(D^{(1)})$ and $[U_c^{(2)}]^\intercal T^*$ are zero almost surely.

We define the positive constant

$$c^{(1)}_{\alpha_0}=\exp\left(-(1+c)F^{-1}_{\chi^2_{r_1}}\left(1-\frac{\alpha_0}{2}\right)\right)\quad\text{and}\quad c^{(2)}_{\alpha_0}=\exp\left(-(1+c)^2F^{-1}_{\chi^2_{r_2}}\left(1-\frac{\alpha_0}{2}\right)\right).\tag{168}$$

By applying (162) with $c_2=c^{(1)}_{\alpha_0}c^{(2)}_{\alpha_0}$, we have

$$\begin{aligned}&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2}\geq(1+c_1)\cdot\frac{c^{(1)}_{\alpha_0}c^{(2)}_{\alpha_0}}{2}\right)\\ &\geq\mathbf{P}\left(\exp\left(-(1+c)\frac{n}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\geq\left(1+\frac{c_1}{2}\right)\cdot c^{(1)}_{\alpha_0}c^{(2)}_{\alpha_0}\right)-\mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)\end{aligned}\tag{169}$$

By (163), we apply the union bound and establish

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-(1+c)\frac{n}{2}[\mathrm{vecl}(D)]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\geq\left(1+\frac{c_1}{2}\right)\cdot c^{(1)}_{\alpha_0}c^{(2)}_{\alpha_0}\right)\\
&\geq 1-\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(1)})\right)\leq c^{(1)}_{\alpha_0}\right)\\
&-\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2),*})]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(2),*})\right)\leq\left(1+\frac{c_1}{2}\right)\cdot c^{(2)}_{\alpha_0}\right)
\end{aligned}\tag{170}$$

Note that

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(1)})\right)\leq c^{(1)}_{\alpha_0}\right)\\
&\leq\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^{\intercal}(\mathbf{Cov}^{(1)}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(1)})\right)\leq c^{(1)}_{\alpha_0}\right)\\
&=\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^{\intercal}\left(\begin{pmatrix}U^{(1)} & U^{(1)}_c\end{pmatrix}\begin{pmatrix}\Lambda^{(1)} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}U^{(1)} & U^{(1)}_c\end{pmatrix}^{\intercal}+\frac{1}{2}d_0\mathbf{I}\right)^{-1}\mathrm{vecl}(D^{(1)})\right)\leq c^{(1)}_{\alpha_0}\right)\\
&\leq\mathbf{P}\left(-n[\mathrm{vecl}(D^{(1)})]^{\intercal}U^{(1)}[\Lambda^{(1)}]^{-1}[U^{(1)}]^{\intercal}\mathrm{vecl}(D^{(1)})\leq -F^{-1}_{\chi^2_{r_1}}\left(1-\frac{\alpha_0}{2}\right)\right)\leq\frac{\alpha_0}{2}
\end{aligned}\tag{171}$$

where the first inequality follows from the fact that $\mathbf{Cov}-\mathbf{Cov}^{(1)}$ is positive definite, the second inequality follows from $d_0>0$ and $[U^{(1)}_c]^{\intercal}\mathrm{vecl}(D^{(1)})$ is zero almost surely and the final upper bound follows from (166). Similarly, we can show that

$$\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2),*})]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(2),*})\right)\leq\left(1+\frac{c_1}{2}\right)\cdot c^{(2)}_{\alpha_0}\right)\leq\frac{\alpha_0}{2}.$$

Combined with (169),(170) and (171), we establish the lemma for the setting with no covariate shift.

D Additional Simulation

D.1 Bias-variance tradeoff in high dimensions

We consider the no covariate shift setting and compare the proposed estimator in (21) with the plug-in estimators in (10) with $\widetilde{b}^{(l)}$ taken as Lasso estimator Tibshirani (1996) or the debiased Lasso estimator Javanmard and Montanari (2014). We set $L=2$, generate $b^{(1)}\in\mathbb{R}^p$ as $b^{(1)}_j=j/40$ for $1\leq j\leq 10$, $b^{(1)}_j=(10-j)/40$ for $11\leq j\leq 20$ and $b^{(1)}_j=0$ for $21\leq j\leq 500$ and generate $b^{(2)}\in\mathbb{R}^p$ as $b^{(2)}_j=b^{(2)}_j+0.3$ for $1\leq j\leq 10$, $b^{(2)}_j=0.3$ for $11\leq j\leq 20$ and $b^{(2)}_j=0$ for $21\leq j\leq 500$. In Figure 8, We report average absolute bias, average standard error and average proportion of variance out of the total mean squared error. Since our goal is to estimate the lower triangular part $(\Gamma_{1,1}, \Gamma_{2,1}, \Gamma_{2,2})$ of the matrix Γ, we average the corresponding accuracy measures of estimating these three entries. The plug-in Lasso estimator has a larger bias than our proposed estimator while the plug-in debiased Lasso estimator has a large bias and variance. The variance proportion of our proposed estimator is much higher than those for the plug-in estimators, which indicates the success of bias correction and the reliable inference performance by quantifying the uncertainty of the variance component.

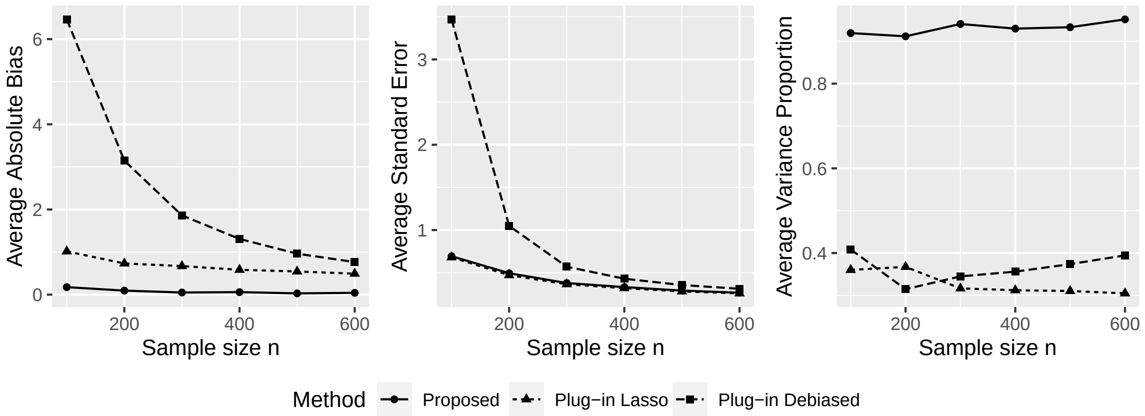
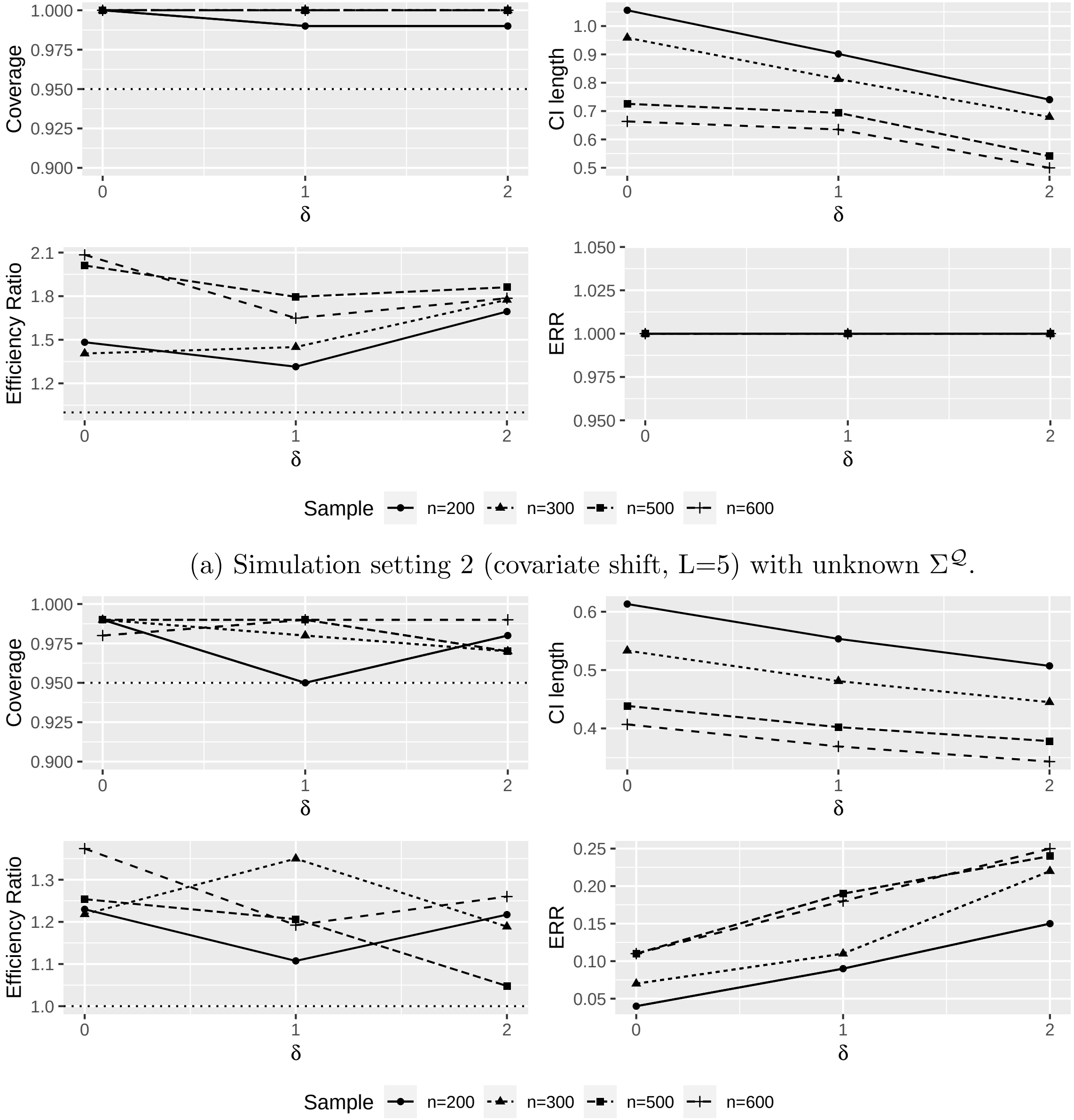


Figure 8: Comparison of our proposed estimator $\widehat{\Gamma}^{\mathcal{Q}}$ of Γ, the plug-in Lasso estimator and the Plug-in Debiased Lasso estimator.

D.2 Additional simulation for covariate shift

We report the simulation results corresponding to settings 2 and 3 in Section 7. We vary δ across $\{0, 1, 2\}$, where the corresponding values $x_{\rm new}^{\intercal}\beta_{\delta}^{*}$ for setting 2 are $\{0.977, 1.124, 1.287\}$ and for setting 3 are $\{-0.115, -0.136, -0.147\}$. The main observations are similar to those in setting 1, where the CI lengths decrease with a larger value of δ and a larger sample size. We shall point out some differences here. First, from Figure 3, the recommended choices of δ are 2 and 0 for settings 2 and 3, respectively. For setting 2, with increasing δ from 0 to 2, the corresponding confidence intervals drop by 25%. Second, in settings 2 and 3, the constructed CIs are over-coverage; for $n = 500$, the relative efficiency is around 1.8 for setting 2 and around 1.2 for setting 3. As a remark, the normal interval with an oracle SE does not necessarily achieve the desired coverage level as the asymptotic limiting distribution of the maximin effect estimator can be non-normal.



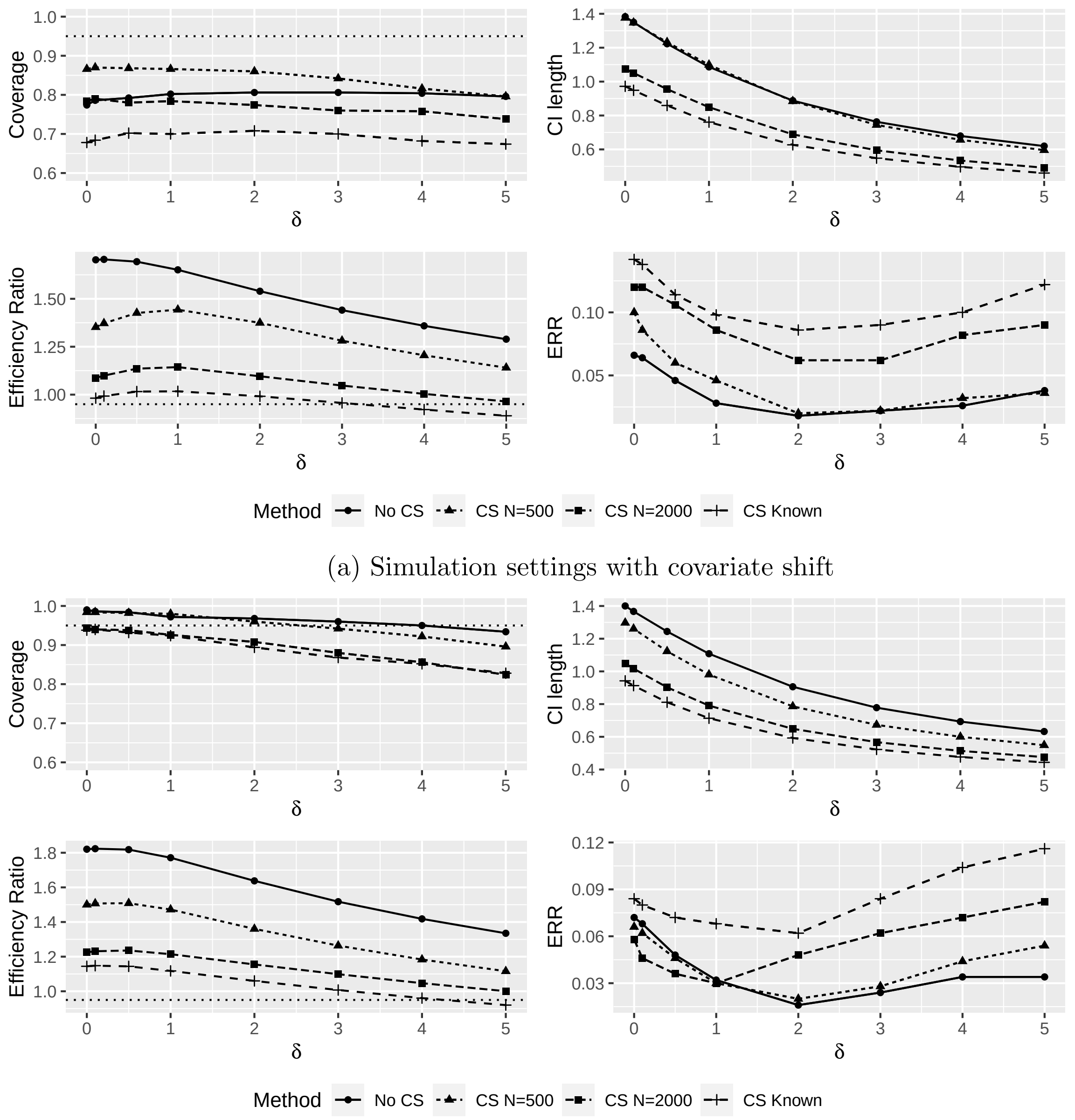
(a) Simulation setting 2 (covariate shift, L=5) with unknown $\Sigma^{\mathcal{Q}}$.

(b) Simulation setting 3 (covariate shift, L=5) with unknown $\Sigma^{\mathcal{Q}}$.

Figure 9: Dependence on δ and n. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to the normal interval with an oracle SE; "ERR" represents the empirical rejection rate out of 500 simulations.

D.3 Additional method comparison

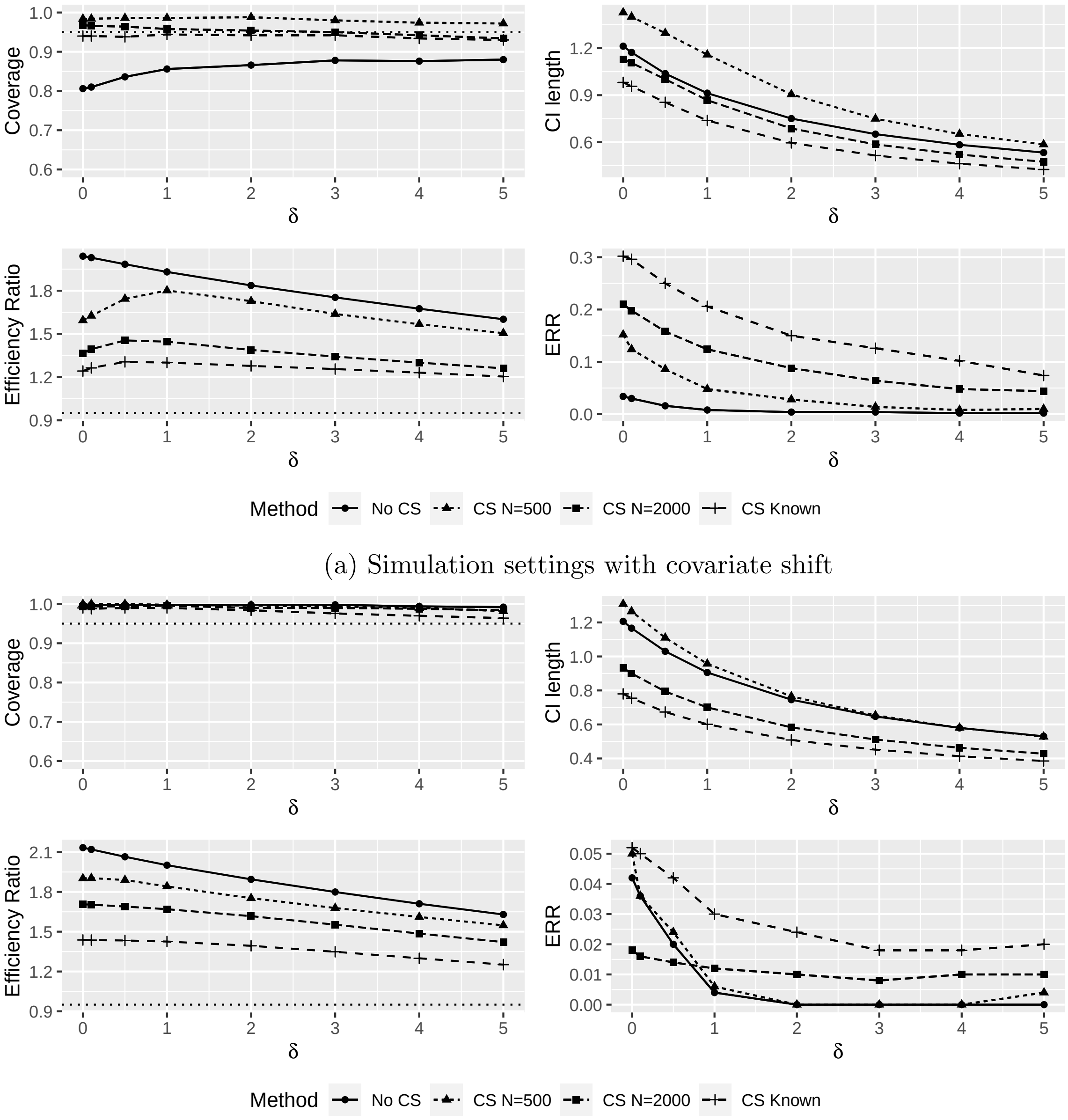
We present additional simulation results for the simulation settings considered in Figure 6. We simply replace the sample size $n = 500$ with $n \in \{100, 200, 300\}$ and present the corresponding results in Figures 10, 11 and 12, respectively. The results are similar to that in Figure 6. For $n = 100$, the coverage levels are below the 95% level while the desired 95% level is achieved for $n = 200$ and $n = 300$.



(a) Simulation settings with covariate shift

(b) Simulation settings with no covariate shift

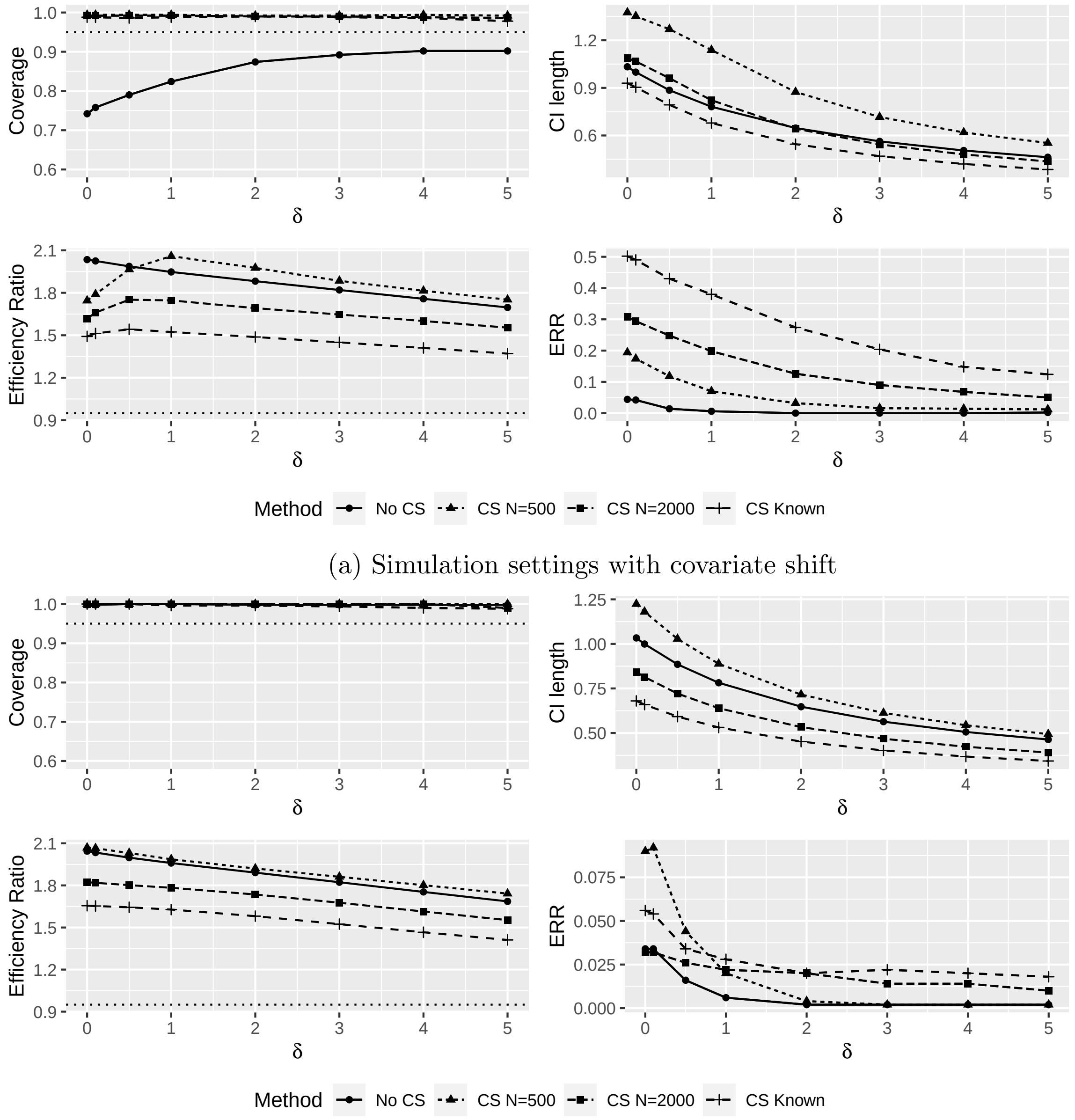
Figure 10: Comparison of covariate shift and no covariate shift algorithms with $n = 100$. The methods “No CS”, “CS $N = 500$”, “CS $N = 2000$”, “CS Known” represent algorithms assuming no covariate shift, Algorithm 1 with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. “Coverage” and “CI Length” stand for the empirical coverage and the average length of our proposed CI; “Efficiency Ratio” represents the ratio of the length of CI in (28) to the normal interval with an oracle SE; “ERR” represents the empirical rejection rate out of 500 simulations.



(a) Simulation settings with covariate shift

(b) Simulation settings with no covariate shift

Figure 11: Comparison of covariate shift and no covariate shift algorithms with $n = 200$. The methods "No CS", "CS $N = 500$", "CS $N = 2000$", "CS Known" represent algorithms assuming no covariate shift, Algorithm 1 with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to the normal interval with an oracle SE; "ERR" represents the empirical rejection rate out of 500 simulations.



(a) Simulation settings with covariate shift

(b) Simulation settings with no covariate shift

Figure 12: Comparison of covariate shift and no covariate shift algorithms with $n = 300$. The methods "No CS", "CS $N = 500$", "CS $N = 2000$", "CS Known" represent algorithms assuming no covariate shift, Algorithm 1 with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to the normal interval with an oracle SE; "ERR" represents the empirical rejection rate out of 500 simulations.