Electronic Journal of Statistics
Vol. 18 (2024) 4407–4458
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2304

Score function-based tests for ultrahigh-dimensional linear models

Weichao Yang[1], **Xu Guo**[2] **and Lixing Zhu**[3]

[1] *School of Statistics, Beijing Normal University,*
e-mail: yangweichao@mail.bnu.edu.cn

[2] *School of Statistics, Beijing Normal University,*
e-mail: xustat12@bnu.edu.cn

[3] *Center for Statistics and Data Science, Beijing Normal University at Zhuhai,*
e-mail: lzhu@bnu.edu.cn

Abstract: In this paper, we investigate score function-based tests to check the significance of an ultrahigh-dimensional sub-vector of the model coefficients when the nuisance parameter vector is also ultrahigh-dimensional in linear models. We first reanalyze and extend a recently proposed score function-based test to derive, under weaker conditions, its limiting distributions under the null and local alternative hypotheses. As it may fail to work when the correlation between testing covariates and nuisance covariates is high, we propose an orthogonalized score function-based test with two merits: debiasing to make the non-degenerate error term degenerate and reducing the asymptotic variance to enhance power performance. Simulations evaluate the finite-sample performances of the proposed tests, and a real data analysis illustrates its application.

MSC2020 subject classifications: Primary 62F03; secondary 62H15.
Keywords and phrases: Ultrahigh-dimensional inference, U-statistics, orthogonalization.

Received January 2024.

1. Introduction

It is important to check whether the covariates of interest contribute to the response, given the other covariates. In linear regression models, this is formulated as testing whether the parameter vector of interest is equal to zero. This paper studies inference of an ultrahigh-dimensional parameter vector of interest with an ultrahigh-dimensional nuisance parameter vector. This problem is of great importance in practice. For instance, researchers may aim to test whether a gene pathway, consisting of ultrahigh-dimensional genes for the same biological functions, is important for a certain clinical outcome, given the other ultrahigh-dimensional genes.

For this challenging problem, there are several proposals available in the literature. The coordinate-based maximum tests have been proposed recently. See for

arXiv: 2212.08446

instance [30, 39, 16, 28, 37]. These methods are computationally expensive because many penalized optimization implementations with ultrahigh-dimensional parameter vector are involved. A Wald-type test was suggested by [23], which is computationally low cost. They imposed the boundedness of the eigenvalues of the covariance matrix. To tackle the problem in the study of the asymptotic properties, brought by too small variance ([23] pointed out), a positive tuning parameter over the sample size is added to the estimated variance. As the limiting null distribution remains unknown, their test with the critical values determined by the standard normal distribution is conservative in theory (see the discussion on page 11 of [23]).

Score function-based testing procedures are also popular. When the dimension of the nuisance parameter vector is low or diverging at a relatively slow rate and the parameter vector of interest is high-dimensional, the references include [19, 40, 20] (for generalized linear models), [14], and [21]. The recent development of score function-based testing procedure is made by [8] who extended the score function-based test of [20] to handle ultrahigh-dimensional nuisance parameter vector. This test is suitable under dense alternative hypotheses and the numerical studies supported some merits of their test. However, their results require that the eigenvalues of the covariance matrix are all bounded and the dimension of the parameter vector of interest grows polynomially with the sample size to guarantee nontrivial power. Further the limiting distributions under the local alternatives are not established.

The above observations motivate us to further study the score function-based test for linear models, extend the results in the literature and propose new test to handle high correlation between nuisance and testing covariates. To be specific, we will do the following. First, we need to reanalyze, under weaker conditions, the properties of the test statistic and extend the results to the case where the testing parameter and nuisance parameter vectors are ultrahigh-dimensional simultaneously at the rates up to the exponential of the sample size. To this end, we derive the limiting distributions under the null and local alternative hypotheses. Second, when the correlation between the covariates of interest and the nuisance covariates is strong, a non-negligible bias causes the tests in [8] fail to work. Therefore, we propose an orthogonalization procedure to reduce the possible bias. Although this technique has been adopted in the recent high-dimensional inference literature [38, 33, 25, 4, 12], to the best of our knowledge, it has not been applied to constructing test statistics based on the quadratic norm of the score function for ultrahigh-dimensional testing parameter vector. Two merits shown in our investigation are as follows. The orthogonalization can debiase the error terms and convert the non-degenerate error terms to degenerate, thus relaxing the correlation assumption between the covariates of interest and nuisance covariates; it can also reduce the variance of the test statistic and thus enhance the power performance, which was not observed in the literature.

Technically, we establish the asymptotic normality of the two proposed test statistics in a different way from those used by [40, 20, 14] for the quadratic norm-based test statistics. Instead of calculating the relatively complex spectral norm of the ultrahigh-dimensional sample matrix, we derive the order of element-max norm of the ultrahigh-dimensional U-statistics with the help of maximal inequalities established in [13] and [11]. The technique developed in this paper can be useful for other high-dimensional inference problems.

The rest of the paper is organized as follows. Section 2 re-analyzes the test statistic in [8] to handle the case with higher dimensional parameter vector of interest and presents the limiting distributions under both null and local alternative hypotheses. The failure of this test statistic in the high correlation case is also discussed. Section 3 introduces the orthogonalization procedure. Section 4 contains an oracle inference procedure to illustrate the merits of the orthogonalization approach. Further, Section 5 develops an orthogonalization-based test in the general case and derives the relevant asymptotic analysis. Section 6 presents simulation studies and a real data analysis. Section 7 offers some conclusions. The Appendix includes detailed proofs of the theoretical properties, technical lemmas, and additional simulation results.

Before closing this section, we introduce some necessary notations. For a d-dimension vector $\mathbf{U}$, write $\|\mathbf{U}\|_r = (\sum_{k=1}^d |U_k|^r)^{1/r}$ and $\|\mathbf{U}\|_\infty = \max_{1\le k\le d}|U_k|$ to denote L_r and L_∞ norms of $\mathbf{U}$, where U_k is the k-th element of $\mathbf{U}$. Further define $\|\mathbf{U}\|_0 = \#\{k : U_k \neq 0\}$. A random variable X is *sub-Gaussian* if the moment generating function (MGF) of X^2 is bounded at some point, namely $\mathbb{E}\exp(X^2/K^2) \le 2$, where K is a positive constant. A random vector $\mathbf{X}$ in $\mathbb{R}^p$ is called *sub-Gaussian* if $x^\top\mathbf{X}$ are sub-gaussian random variables for all $x \in \mathbb{R}^p$. For $a, b \in \mathbb{R}$, write $a \vee b = \max\{a, b\}$. For $p_1 \times p_2$ dimensional matrix $\mathbf{A}$, write $\lambda_{\max}(\mathbf{A})$ to denote the spectral norm of $\mathbf{A}$. Further define $\|\mathbf{A}\|_0 = \#\{(i, j) : A_{ij} \neq 0\}$ and $\|\mathbf{A}\|_F = \{\mathrm{tr}(\mathbf{A}\mathbf{A}^\top)\}^{1/2}$, where A_{ij} is the (i, j)-th element of $\mathbf{A}$.

2. The score function-based test and new results

Let $Y \in \mathbb{R}$ be the response variable along with the covariates $\mathbf{X} = (X_1, \dots, X_{p_{\boldsymbol\beta}})^\top \in \mathbb{R}^{p_{\boldsymbol\beta}}$ and $\mathbf{Z} = (Z_1, \dots, Z_{p_{\boldsymbol\gamma}})^\top \in \mathbb{R}^{p_{\boldsymbol\gamma}}$. Consider the following linear model:

$$Y = \boldsymbol\beta^\top\mathbf{X} + \boldsymbol\gamma^\top\mathbf{Z} + \epsilon, \tag{1}$$

where ϵ is the random error satisfying $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon^2) = \sigma^2$. Let $\boldsymbol{V} = (\mathbf{X}^\top, \mathbf{Z}^\top)^\top$ and $\boldsymbol\Sigma$ be the covariance matrix of $\boldsymbol{V}$. Without loss of generality, assume that $\mathbb{E}(\boldsymbol{V}) = \mathbf{0}$, $\boldsymbol\Sigma$ is positive definite and ϵ is uncorrelated with $\boldsymbol{V}$. Our primary interest is to detect whether $\mathbf{X}$ contributes to the response Y or not given the other covariates, which is testing the following inference problem:

$$\mathbb{H}_0 : \boldsymbol\beta = \mathbf{0}, \qquad \text{versus} \qquad \mathbb{H}_1 : \boldsymbol\beta \neq \mathbf{0}. \tag{2}$$

To test the above hypothesis, we can construct test statistics based on score functions. An advantage of score function-based tests is that we do not need to estimate the parametric vector of interest. To be precise, consider the following L_2 loss function:

$$\mathcal{L}(\boldsymbol\beta, \boldsymbol\gamma) = \mathbb{E}(Y - \boldsymbol\beta^\top\mathbf{X} - \boldsymbol\gamma^\top\mathbf{Z})^2/2,$$

and the corresponding score function of $\boldsymbol{\beta}$:

$$\partial\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma})/\partial\boldsymbol{\beta} = \nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma}) = -\mathbb{E}\{(Y-\boldsymbol{\beta}^\top\mathbf{X}-\boldsymbol{\gamma}^\top\mathbf{Z})\mathbf{X}\}.$$

Then $\nabla_{\boldsymbol{\beta}}\mathcal{L}(\mathbf{0},\boldsymbol{\gamma}) = \mathbf{0}$ corresponds to $\mathbb{H}_0$; otherwise, to $\mathbb{H}_1$.

A test statistic can be based on the quadratic norm $\nabla_{\boldsymbol{\beta}}\mathcal{L}(\mathbf{0},\boldsymbol{\gamma})^\top\nabla_{\boldsymbol{\beta}}\mathcal{L}(\mathbf{0},\boldsymbol{\gamma})$. As the nuisance parameter $\boldsymbol{\gamma}$ is unknown, we can replace $\boldsymbol{\gamma}$ with an estimator $\hat{\boldsymbol{\gamma}}$. Now suppose $\{\mathbf{X}_i,\mathbf{Z}_i,Y_i\}_{i=1}^n$ is a random sample from the population $(\mathbf{X},\mathbf{Z},Y)$. [20] proposed the following test statistic based on the quadratic norm of the score function:

$$T_n = \frac{1}{n}\sum_{i\neq j}(Y_i-\hat{\boldsymbol{\gamma}}^\top\mathbf{Z}_i)(Y_j-\hat{\boldsymbol{\gamma}}^\top\mathbf{Z}_j)\mathbf{X}_i^\top\mathbf{X}_j, \tag{3}$$

where $\hat{\boldsymbol{\gamma}}$ is the least squares estimator. Clearly T_n can be seen as a U-statistic type estimator of $(n-1)\nabla_{\boldsymbol{\beta}}\mathcal{L}(\mathbf{0},\hat{\boldsymbol{\gamma}})^\top\nabla_{\boldsymbol{\beta}}\mathcal{L}(\mathbf{0},\hat{\boldsymbol{\gamma}})$. In the asymptotic analysis, the growth rate of the dimension of $\boldsymbol{\gamma}$ is required to be slower than $n^{1/4}$. Recently, [8] extended it to handle the ultrahigh-dimensional nuisance parameter situation. They obtained the estimator $\hat{\boldsymbol{\gamma}}$ by solving the following penalized problem,

$$\hat{\boldsymbol{\gamma}} = \operatorname*{argmin}_{\boldsymbol{\gamma}\in\mathbb{R}^{p_{\boldsymbol{\gamma}}}}\frac{1}{2n}\sum_{i=1}^n(Y_i-\boldsymbol{\gamma}^\top\mathbf{Z}_i)^2+\lambda_Y\|\boldsymbol{\gamma}\|_1, \tag{4}$$

where λ_Y is the tuning parameter and $\boldsymbol{\gamma}$ has a sparse structure. Other penalties such as SCAD and MCP, are also applicable. To deal with the case with ultrahigh-dimensional parameter vector of interest and relax some conditions, we conduct a further investigation for their test firstly.

Compared with existing high-dimensional literature in which the dimension of nuisance parameter is much less than sample size, there are some technical difficulties listed as following.

- There is no explicit formula for the penalized estimator. When the dimension of the nuisance parameter vector is low or diverging at relatively slow rate, the nuisance parameter was estimated by least square estimation or maximum likelihood estimation in [20, 21]. These estimators aforementioned have explicit formulas. The theoretical results can be proved by pluging the explicit forms of the estimators in relevant test statistics. However, when the nuisance parameter vector is ultrahigh-dimensional, these estimators are no longer feasible. Instead we use penalized estimation procedures to deal with the ultrahigh-dimensional unknown nuisance parameter vector. This brings great challenge now since the penalized estimators do not have explicit formulas.
- It is difficult to calculate the spectral norm of an ultrahigh-dimensional sample matrix. To establish the asymptotic normality, we need to derive the order of the following term

$$(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^\top\frac{1}{n}\sum_{i\neq j}\mathbf{Z}_i\mathbf{Z}_j^\top\mathbf{X}_i^\top\mathbf{X}_j(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}).$$

When the dimension of nuisance covariates $\mathbf{Z}$ is relatively low, previous literature [40, 20, 14] calculated the spectral norm of $n^{-1}\sum_{i\neq j}\mathbf{Z}_i\mathbf{Z}_j^\top\mathbf{X}_i^\top\mathbf{X}_j$. However, when the dimension of nuisance covariates $\mathbf{Z}$ is ultrahigh, it is very difficult to obtain the order of the spectral norm of the above ultrahigh-dimensional matrix. Alternatively we turn to derive the order of element-max norm of the related ultrahigh-dimensional sample matrix with the help of maximal inequalities established in [13] and [11]. This is a new technical approach for analyzing score function-based tests.

2.1. Limiting null distribution

Let $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}}$ be the covariance matrices of the covariates $\mathbf{X}$ and $\mathbf{Z}$ respectively. Denote $p_{\boldsymbol{\beta}}$, $p_{\boldsymbol{\gamma}}$ as the dimension of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and let $p = p_{\boldsymbol{\beta}} + p_{\boldsymbol{\gamma}}$. Denote $\boldsymbol{\gamma}_\phi = \boldsymbol{\Sigma}_{\mathbf{Z}}^{-1}\mathbb{E}(\mathbf{Z}Y)$, and $\boldsymbol{\gamma}_\phi = \boldsymbol{\gamma}$ under $\mathbb{H}_0$. Let s be a positive integer and represents the sparsity level of $\boldsymbol{\gamma}_\phi$. Let $\varrho^2 = \max_{1\le k\le p_{\boldsymbol{\gamma}}}\|\mathbb{E}(Z_k\mathbf{X})\|_2^2$ and $\varpi^2 = s^2\log p_{\boldsymbol{\gamma}}\varrho^2$. Here ϱ^2 describes the dependence of covariates of interest and nuisance covariates. Next, under some technical assumptions, we study the asymptotic null distribution of the test statistic T_n with $\hat{\boldsymbol{\gamma}}$ in (4).

Assumption 1. $\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}}^4) = o(\mathrm{tr}^2(\boldsymbol{\Sigma}_{\mathbf{X}}^2))$ and $\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}}^2)\to\infty$ as $(n, p_{\boldsymbol{\beta}})\to\infty$.

Assumption 2. $\boldsymbol{V}$ can be expressed as

$$\boldsymbol{V} = \boldsymbol{\Gamma}\boldsymbol{\nu},$$

where $\boldsymbol{\Gamma}$ is a $p\times m$ dimensional matrix with $p\le m$. The L_2 norms of row vectors in $\boldsymbol{\Gamma}$ are uniformly bounded. $\boldsymbol{\nu}$ is an m-dimensional sub-Gaussian random vector with mean zero and identity covariance matrix.

Assumption 3. $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_\phi\|_1 = O_p(s\sqrt{\log p_{\boldsymbol{\gamma}}/n})$.

Assumption 4. $\log p_{\boldsymbol{\gamma}} = O(n^b)$ for some constant $0 < b < 1/3$.

Assumption 5. ϵ is *sub-Gaussian* with bounded *sub-Gaussian* norm.

Assumption 1 frequently appeared in the literature [10, 20, 14] and is required in applying the martingale central limit theorem. If all the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{X}}$ are bounded, then Assumption 1 holds. In the previous works such as [28, 23, 8], they required that the eigenvalues of $\boldsymbol{\Sigma}$ are all bounded, which is stronger than our assumption 1. Assumption 2 says that $\boldsymbol{V}$ can be expressed as a linear transformation of an m-dimensional sub-Gaussian vector $\boldsymbol{\nu}$ with zero mean and unit variance. This assumption is similar to the pseudo-independence assumption, which is widely used in the literature such as [1, 9, 10, 40, 14]. The boundedness assumption of L_2 norms of row vectors in $\boldsymbol{\Gamma}$ is imposed to ensure that sub-Gaussian norms of the components of $\boldsymbol{V}$ are uniformly bounded. Assumption 3 requires the L_1 error bound of $\hat{\boldsymbol{\gamma}}$ at the order of $s(\log p_{\boldsymbol{\gamma}}/n)^{1/2}$. Many estimators, such as Lasso, SCAD, and MCP, can achieve such a rate of convergence. See for instance [27]. Assumption 4 allows the dimension of the nuisance parameter in an exponential order of the sample size. Assumption 5 is standard in the analysis for high-dimensional linear models.

Theorem 2.1. *Under $\mathbb{H}_0$ in (2) and Assumptions 1 – 5, and the following two conditions:*

$$\varpi^2 \vee (\log p_{\boldsymbol{\gamma}})^{1/2}\varpi\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}) = o(\sqrt{\boldsymbol{\Lambda}_{\mathbf{X}}}), \tag{5}$$

and

$$s(\log p_{\boldsymbol{\gamma}})^{3/2}/\sqrt{n} = o(1), \tag{6}$$

we have

$$\frac{T_n}{\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}} \to N(0,1) \quad \textit{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}, p_{\boldsymbol{\gamma}}) \to \infty$, where $\boldsymbol{\Lambda}_{\mathbf{X}} = \sigma^4 \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}}^2)$.

Note that [39] assumed $s^*(\log p)^{3/2}/\sqrt{n} = o(1)$. Here s^* represents the sparsity level of the entire parameter vector $(\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$. Under null hypothesis, $s^*(\log p)^{3/2}/\sqrt{n} = o(1)$ becomes $s(\log p)^{3/2}/\sqrt{n} = o(1)$. Since p can be much larger than $p_{\boldsymbol{\gamma}}$, the sparsity requirement in (6) is still weaker than that in [39]. Condition (5) restricts the sparsity, the dimensions of the nuisance and testing parameter, and the correlations among the covariates. Generally speaking, if the correlations are weak and the dimension of the testing parameter is high, the sparsity level and the dimension of the nuisance parameter can be very high. Therefore, when the correlation between $\mathbf{X}$ and $\mathbf{Z}$ is weak, T_n still has a tractable limiting null distribution. Particularly, when $\boldsymbol{\Sigma}$ has bounded eigenvalues, ϱ^2 and $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})$ can be bounded by a constant. Thus condition (5) can be simplified as $s^2 \log p_{\boldsymbol{\gamma}} = o(\sqrt{p_{\boldsymbol{\beta}}})$. Furthermore, if $p_{\boldsymbol{\beta}} = cn^2$ with $c > 0$, condition (5) can be further simplified as $s^2 \log p_{\boldsymbol{\gamma}} = o(n)$, which is weaker than the conditions on sparsity and dimension in previous works (for example $s = o(\sqrt{n}/(\log p)^{3/2})$ in [39]; $s = o(\sqrt{n}/(\log p)^3)$ in [28]). With larger $p_{\boldsymbol{\beta}}$, the condition $s^2 \log p_{\boldsymbol{\gamma}} = o(\sqrt{p_{\boldsymbol{\beta}}})$ is milder. This implies that T_n has tractable null distribution even when both $p_{\boldsymbol{\beta}}$ and $p_{\boldsymbol{\gamma}}$ are of exponential order of n.

To formulate the testing procedure based on Theorem 2.1, we use

$$R_{1n} = \hat{\sigma}^4 \frac{1}{2\binom{n}{4}} \sum_{i<j<k<l}^{n} (\mathbf{X}_i - \mathbf{X}_j)^\top (\mathbf{X}_k - \mathbf{X}_l)(\mathbf{X}_j - \mathbf{X}_k)^\top (\mathbf{X}_l - \mathbf{X}_i)$$

to estimate $\boldsymbol{\Lambda}_{\mathbf{X}}$, where $\hat{\sigma}^2$ is a consistent estimator of the error variance σ^2, such as the one in [32]. Under the null hypothesis, R_{1n} is a ratio consistent estimator of $\boldsymbol{\Lambda}_{\mathbf{X}}$. The consistency of R_{1n} has been discussed by many authors such as [40, 14, 21]. Combining Theorem 2.1 and Slutsky Theorem, we reject $\mathbb{H}_0$ at a significance level α if

$$T_n \geq z_\alpha \sqrt{2R_{1n}}.$$

Here z_α is the upper-α quantile of standard normal distribution.

2.2. *Power analysis*

Next, we study the limiting distribution of T_n under a class of alternative hypotheses. Let $\mathbb{E}(\mathbf{X}\mathbf{Z}^\top) =: \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Z}} = \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{X}}^\top$ be the covariance matrix between $\mathbf{X}$ and $\mathbf{Z}$ and $\boldsymbol{\eta} = \mathbf{X} - \boldsymbol{W}^\top\mathbf{Z}$, where

$$\boldsymbol{W} = \boldsymbol{\Sigma}_{\mathbf{Z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{X}}. \tag{7}$$

Let $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ be the covariance matrix of $\boldsymbol{\eta}$.

Define the following family of local alternatives:

$$\begin{aligned}\mathscr{L}_1(\boldsymbol{\beta}) = &\left\{\boldsymbol{\beta} \in \mathbb{R}^{p_{\boldsymbol{\beta}}} \middle| \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta} = o(1),\ \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta} = o\left(\frac{\boldsymbol{\Lambda}_{\mathbf{X}}}{n\varpi^2}\right)\right.\\ &\left.\text{and } \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta} = o\left(\frac{\boldsymbol{\Lambda}_{\mathbf{X}}}{n}\right)\right\}.\end{aligned}$$

Similar definitions of local alternative sets have been also considered in [40, 14, 21]. The class $\mathscr{L}_1(\boldsymbol{\beta})$ prescribes a small difference between $\boldsymbol{\beta}$ and $\mathbf{0}$. For illustration, suppose the eigenvalues of $\boldsymbol{\Sigma}$ are bounded by a constant, then $\mathscr{L}_1(\boldsymbol{\beta})$ can be simplified as

$$\mathscr{L}_1(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} \in \mathbb{R}^{p_{\boldsymbol{\beta}}} \middle| \|\boldsymbol{\beta}\|_2 = o\left(\min\left(1, \sqrt{\frac{p_{\boldsymbol{\beta}}}{ns^2\log p_{\boldsymbol{\gamma}}}}\right)\right)\right\}.$$

We have the following theorem.

Theorem 2.2. *Under the conditions in Theorem 2.1, for $\boldsymbol{\beta} \in \mathscr{L}_1(\boldsymbol{\beta})$, we have*

$$\frac{T_n - n\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}} \to N(0,1) \quad \textit{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}, p_{\boldsymbol{\gamma}}) \to \infty$.

Theorem 2.2 indicates that the asymptotic power of the test statistic T_n under the local alternatives $\mathscr{L}_1(\boldsymbol{\beta})$ is given by

$$\phi_{1n} = \Phi\left(-z_\alpha + \frac{n\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}}\right), \tag{8}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Equation (8) implies that the proposed test has nontrivial power as long as the signal-to-noise ratio $n\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta}/\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}$ does not vanish to 0 as $(n, p_{\boldsymbol{\beta}}) \to \infty$.

It is noteworthy that the power performance of T_n relies not only on the difference between $\boldsymbol{\beta}$ and $\mathbf{0}$ but also on $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. T_n gains more power when the difference between $\boldsymbol{\beta}$ and $\mathbf{0}$ is larger. Additionally, T_n gains more power when the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ are larger. Further recall that $\boldsymbol{\Lambda}_{\mathbf{X}} = \sigma^4\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}}^2)$. Thus when the variance of the error term is small, the power of T_n can also be large.

Compared with [8], we now establish the asymptotic distribution of T_n under local alternative hypotheses. Both the parameter vector of interest and nuisance parameter vector are allowed to be ultrahigh-dimensional.

2.3. The failure of T_n

As discussed, the asymptotic distribution of T_n can be established when the correlation between $\mathbf{X}$ and $\mathbf{Z}$ is weak and $p_{\boldsymbol{\beta}}$ is relatively large compared with $p_{\boldsymbol{\gamma}}$. However, in highly correlated cases, condition (5) fails as ϱ^2 can be divergent in this situation. This may cause the failure of the asymptotic normality. Practically, the situation that $\mathbf{X}$ and $\mathbf{Z}$ are correlated can be motivated by the study where $\mathbf{Z}$ can represent some confounding variables, which can affect many variables in $\mathbf{X}$. For instance, in genetic studies, the effect of certain segments of the deoxyribonucleic acid on the gene expression may be confounded by population structure and microarray expression artifacts [26]. The confounding biases can yield an inflated distribution of test statistics in genome-wide association studies [6]. To illustrate this problem intuitively, consider the following toy example.

Example. Generate the covariates according to the following model:

$$\mathbf{X} = \boldsymbol{W}^\top \mathbf{Z} + \boldsymbol{\eta}, \tag{9}$$

where $\boldsymbol{\eta}$ is a $p_{\boldsymbol{\beta}}$-dimensional random vector and $\boldsymbol{\eta}$ is independent of $\mathbf{Z}$. $\boldsymbol{W}$ is a $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\beta}}$ dimensional matrix with non-zero corners

$$\boldsymbol{W}^\top = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{Q}^\top & \mathbf{0} & \boldsymbol{Q}^\top \end{pmatrix}, \tag{10}$$

where $\boldsymbol{Q}$ is a $d_{\mathbf{Z}} \times 2d_{\mathbf{X}}$ dimensional matrix and all the elements of $\boldsymbol{Q}$ equal 0.5. In this example, we set $\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\gamma} \neq \mathbf{0}$. The null hypothesis $\mathbb{H}_0$ then holds. We derive ϱ^2 increases with the increase of $d_{\mathbf{X}}$ by some calculation. Let $d_{\mathbf{Z}} = 3$ and vary $d_{\mathbf{X}}$ from 0 to 15.

Left panel of Figure 1 plots the empirical sizes of T_n. As $d_{\mathbf{X}}$ increases, the empirical size rises rapidly, and the significance level cannot be maintained. Right panel of Figure 1 reports the empirical probability density function of T_n, which can be well approximated by the standard normal distribution when $d_{\mathbf{X}} = 0$. However, the deviation gradually increases with the increase of $d_{\mathbf{X}}$. This toy example illustrates that when $\mathbf{X}$ and $\mathbf{Z}$ are not weakly dependent, T_n is not applicable.

In the proof of Theorem 2.1, the following error term depends on the correlation between $\mathbf{X}$ and $\mathbf{Z}$:

$$\mathrm{ERROR}_1 = (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^\top \frac{1}{n} \sum_{i \neq j} \mathbf{Z}_i \mathbf{X}_i^\top \mathbf{X}_j \mathbf{Z}_j^\top (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}), \tag{11}$$

where $n^{-1} \sum_{i \neq j} \mathbf{Z}_i \mathbf{X}_i^\top \mathbf{X}_j \mathbf{Z}_j^\top$ is a $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\gamma}}$ dimensional random matrix, and the (k, l)-th element is a non-degenerate U-statistic with non-zero mean $\mathbb{E}(Z_k \mathbf{X})^\top \mathbb{E}(Z_l \mathbf{X})$. The order of ERROR_1 is $O_p(s^2 \log p_{\boldsymbol{\gamma}} \varrho^2)$. Thus condition (5) is required to reduce the impact of bias term ERROR_1 on the asymptotic behavior of T_n. Clearly the bias is no longer negligible if $s^2 \log p_{\boldsymbol{\gamma}} \varrho^2 \geq C\sqrt{\boldsymbol{\Lambda}_{\mathbf{X}}}$ for some constant C. In the following, we suggest an orthogonalization-based test to reduce the bias term.

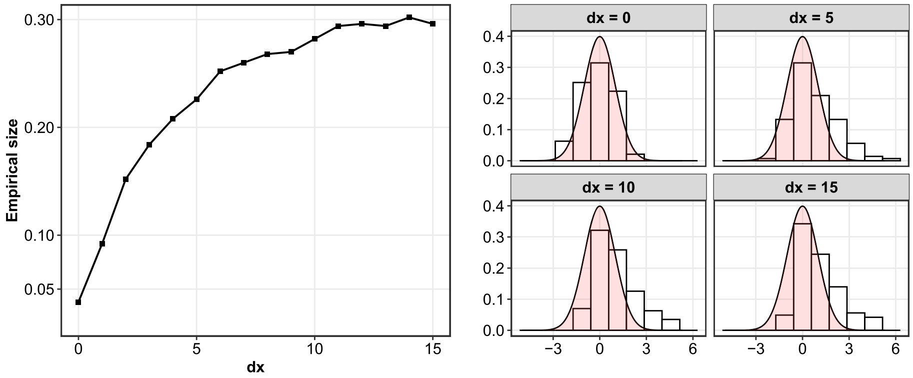

Fig 1. *Left panel presents the empirical sizes of* T_n *with different* $d_{\mathbf{X}}$*. Right panel presents the empirical probability density function of* T_n *with different* $d_{\mathbf{X}}$ *(*$d_{\mathbf{X}}$ *= 0,5,10,15). The pink shade represents the probability density function of the standard normal distribution. Set* $n = 100$*,* $p = 600$ *and* $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}}$*. We generate 500 replications and reject the null hypothesis at the significance level* $\alpha = 0.05$*. More details can be seen in Scenario 3 in Section 6.*

3. An orthogonal score function-based test

As discussed in subsection 2.3, the estimation error of $\hat{\boldsymbol{\gamma}}$ may make the test statistic T_n fail to work when the correlation of $\mathbf{X}$ and $\mathbf{Z}$ is high. To make the score function immune to the estimation error of $\hat{\boldsymbol{\gamma}}$, we consider orthogonalizing the score function of $\boldsymbol{\beta}$,

$$\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\beta}}\mathcal{L} - \boldsymbol{W}^{\top}\nabla_{\boldsymbol{\gamma}}\mathcal{L} \quad \text{with} \quad \boldsymbol{W}^{\top} = \nabla_{\boldsymbol{\beta}\boldsymbol{\gamma}}\mathcal{L}(\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L})^{-1}.$$

Here $\mathcal{L} = \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbb{E}(Y - \boldsymbol{\beta}^{\top}\mathbf{X} - \boldsymbol{\gamma}^{\top}\mathbf{Z})^2/2, \nabla_{\boldsymbol{\beta}}\mathcal{L} = \partial\mathcal{L}/\partial\boldsymbol{\beta}, \nabla_{\boldsymbol{\beta}\boldsymbol{\gamma}}\mathcal{L} = \partial\mathcal{L}^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}$. $\nabla_{\boldsymbol{\gamma}}\mathcal{L}$ and $\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L}$ are similarly defined.

The main idea of orthogonalization is to construct a statistic for the target parameter, which is locally insensitive to the nuisance parameter. The orthogonalization plays an important role in high-dimensional inference problems and has been successfully applied in the recent literature. See for example, [4], [30] and [3]. However, the current adoption of orthogonalization only focuses on low-dimensional parameters. Actually the coordinate-based maximum tests firstly consider orthogonalization for each element of testing parameter vector and then take the maximum of all individual test statistics for each elements. To the best of our knowledge, orthogonalization has not been investigated for test statistics based on the quadratic norm of the score function for ultrahigh-dimensional testing parameter vector.

In our model setting, $\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is equal to

$$\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\mathbb{E}\{(Y - \boldsymbol{\beta}^{\top}\mathbf{X} - \boldsymbol{\gamma}^{\top}\mathbf{Z})(\mathbf{X} - \boldsymbol{W}^{\top}\mathbf{Z})\}.$$

Again $\mathbb{H}_0$ corresponds to $\mathcal{S}(\mathbf{0}, \boldsymbol{\gamma}) = \mathbf{0}$. Compared with $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, we replace $\mathbf{X}$ with $\mathbf{X} - \boldsymbol{W}^{\top}\mathbf{Z}$ now. We can then construct a test statistic based on $\mathcal{S}(\mathbf{0}, \boldsymbol{\gamma})^{\top}$

$\mathcal{S}(\mathbf{0},\boldsymbol{\gamma})$. For the sample $\{\mathbf{X}_i,\mathbf{Z}_i,Y_i\}_{i=1}^n$, define:

$$M_n^* = \frac{1}{n}\sum_{i\neq j}(Y_i-\boldsymbol{\gamma}^\top\mathbf{Z}_i)(Y_j-\boldsymbol{\gamma}^\top\mathbf{Z}_j)(\mathbf{X}_i-\boldsymbol{W}^\top\mathbf{Z}_i)^\top(\mathbf{X}_j-\boldsymbol{W}^\top\mathbf{Z}_j). \tag{12}$$

We construct a final test statistic in Sections 4 and 5.

4. The oracle inference

To illustrate the merits of the orthogonalization technique, we first consider the case with given $\boldsymbol{W}$. Recall $\boldsymbol{W}$ is defined in (7) and $\boldsymbol{W}=\boldsymbol{\Sigma}_{\mathbf{Z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZX}}$. A sufficient condition for known $\boldsymbol{W}$ is that the joint distribution of $(\mathbf{X}^\top,\mathbf{Z}^\top)^\top$ is known in advance. This assumption is given in the recent literature on high-dimensional statistics, such as the model-X knockoff procedure in [7]. Based on the term in (12), consider the following test statistic:

$$M_n^{\mathrm{o}} = \frac{1}{n}\sum_{i\neq j}(Y_i-\hat{\boldsymbol{\gamma}}^\top\mathbf{Z}_i)(Y_j-\hat{\boldsymbol{\gamma}}^\top\mathbf{Z}_j)(\mathbf{X}_i-\boldsymbol{W}^\top\mathbf{Z}_i)^\top(\mathbf{X}_j-\boldsymbol{W}^\top\mathbf{Z}_j). \tag{13}$$

We obtain the estimator $\hat{\boldsymbol{\gamma}}$ by solving a penalized least squares problem in (4).

4.1. Limiting null distribution

Recall $\boldsymbol{\eta}$ is defined in subsection 2.2 and $\boldsymbol{\eta}=\mathbf{X}-\boldsymbol{W}^\top\mathbf{Z}$. $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ is the covariance matrix of $\boldsymbol{\eta}$. Give the following assumption.

Assumption 6. $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^4)=o(\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2))$ and $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2)\to\infty$ as $(n,p_{\boldsymbol{\beta}})\to\infty$.

Assumption 6 is a counterpart of Assumption 1 in the case of the orthogonal score function. Theorem 4.1 states the limiting null distribution of the oracle test statistic M_n^{o} in (13).

Theorem 4.1. *Under $\mathbb{H}_0$, and Assumptions 2-5, 6 and condition (6), we have*

$$\frac{M_n^o}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}}\to N(0,1)\quad \textit{in distribution}$$

as $(n,p_{\boldsymbol{\beta}},p_{\boldsymbol{\gamma}})\to\infty$, where $\boldsymbol{\Lambda}_{\boldsymbol{\eta}}=\sigma^4\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2)$.

Notably, compared with Theorem 2.1, condition (5) is removed in Theorem 4.1. Except for Assumption 6, there are no additional restrictions on the relationship between the covariates in Theorem 4.1. The dependence requirement is greatly relaxed. The proof for this theorem is similar to that for Theorem 2.1, but the error term becomes

$$\mathrm{ERROR}^{\mathrm{o}}=(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})^\top\frac{1}{n}\sum_{i\neq j}\mathbf{Z}_i\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\mathbf{Z}_j^\top(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}),$$

where $\boldsymbol{\eta}_i = \mathbf{X}_i - \boldsymbol{W}^\top \mathbf{Z}_i$ and $n^{-1}\sum_{i\neq j}\mathbf{Z}_i\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\mathbf{Z}_j^\top$ is a $p_{\boldsymbol{\gamma}}\times p_{\boldsymbol{\gamma}}$ dimensional matrix with zero-mean degenerate U-statistic components. Benefitting from the zero-mean property of $n^{-1}\sum_{i\neq j}\mathbf{Z}_i\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\mathbf{Z}_j^\top$, the order of the bias term $\mathrm{ERROR}^{\mathrm{o}}$ is much smaller than that of ERROR_1. From the theory of U-statistics (see, e.g., [31]), the order of a typical degenerate U statistic is $O_p(n^{-1})$. Thus, by the property of degenerate U statistic, $n^{-1}\sum_{i\neq j}\mathbf{Z}_i\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\mathbf{Z}_j^\top$ is much easier to handle than $n^{-1}\sum_{i\neq j}\mathbf{Z}_i\mathbf{X}_i^\top\mathbf{X}_j\mathbf{Z}_j^\top$. In the proof, we show that $\mathrm{ERROR}^{\mathrm{o}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$. In summary, the orthogonalization technique has two merits: debiasing and converting a non-degenerate U-statistic to a degenerate one. The first is frequently observed in the literature, such as [41], [38], and [33]. We have not seen in the literature any study to show the second merit of the orthogonalization technique.

4.2. *Power analysis*

To study the power performance of M_n^{o}, consider the following local alternatives:

$$\mathscr{L}^{\mathrm{o}}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta}\in\mathbb{R}^{p_{\boldsymbol{\beta}}}\middle|\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta} = o(1) \text{ and } \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^3\boldsymbol{\beta} = o\left(\frac{\Lambda_{\boldsymbol{\eta}}}{n}\right)\right\}.$$

Compared with previous literature [40, 14], we replace $\boldsymbol{\Sigma}_{\mathbf{X}}$, $\boldsymbol{\Lambda}_{\mathbf{X}}$ with $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, $\boldsymbol{\Lambda}_{\boldsymbol{\eta}}$ in $\mathscr{L}^o(\boldsymbol{\beta})$. This modification is due to the construction of M_n. While previous test statistics are constructed based on $\mathbf{X}_i^\top\mathbf{X}_j$, the construction of M_n is based on $\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j$. Similar to $\mathscr{L}_1(\boldsymbol{\beta})$, the class $\mathscr{L}^o(\boldsymbol{\beta})$ also describes a small discrepancy between $\boldsymbol{\beta}$ and $\mathbf{0}$. Under the bounded eigenvalue assumption of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, $\mathscr{L}^o(\boldsymbol{\beta})$ can be simplified as

$$\mathscr{L}^o(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta}\in\mathbb{R}^{p_{\boldsymbol{\beta}}}\middle|\|\boldsymbol{\beta}\|_2 = o\left(\min\left(1,\sqrt{\frac{p_{\boldsymbol{\beta}}}{n}}\right)\right)\right\}.$$

We have the following theorem.

Theorem 4.2. *Under conditions in Theorem 4.1, and for $\boldsymbol{\beta}\in\mathscr{L}^o(\boldsymbol{\beta})$, we derive*

$$\frac{M_n^o - n\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \textit{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}, p_{\boldsymbol{\gamma}}) \to \infty$.

Similarly, the asymptotic power function of M_n^{o} under the local alternatives is

$$\phi_n^{\mathrm{o}} = \Phi\left(-z_\alpha + \frac{n\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}}\right). \tag{14}$$

Equation (14) implies that the proposed test has nontrivial power as long as the signal-to-noise ratio $n\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta}/\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}$ does not vanish to 0 as $(n, p_{\boldsymbol{\beta}}) \to \infty$.

Comparing with Theorem 2.2, the Pitman asymptotic relative efficiency (ARE) of M_n^{o} concerning the T_n is

$$\mathrm{ARE}(M_n^{\mathrm{o}}, T_n) = \left\{ \frac{\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}}^2)}{\mathrm{tr}(\mathbf{\Sigma}_{\boldsymbol{\eta}}^2)} \right\}^{1/2}.$$

By the definition of $\boldsymbol{\eta}$, $\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}}^2) \geq \mathrm{tr}(\mathbf{\Sigma}_{\boldsymbol{\eta}}^2)$. Thus the power of M_n^o is higher than T_n. This result is important as the orthogonalization technique can simultaneously reduce bias and variance, thus improving power performance.

5. The test with unknown $\boldsymbol{W}$

In this case, the test statistic is defined by using a plug-in estimator of $\boldsymbol{W}$:

$$M_n = \frac{1}{n}\sum_{i\neq j}(Y_i - \hat{\boldsymbol{\gamma}}^\top \mathbf{Z}_i)(Y_j - \hat{\boldsymbol{\gamma}}^\top \mathbf{Z}_j)(\mathbf{X}_i - \hat{\boldsymbol{W}}^\top \mathbf{Z}_i)^\top(\mathbf{X}_j - \hat{\boldsymbol{W}}^\top \mathbf{Z}_j). \tag{15}$$

We obtain the estimator $\hat{\boldsymbol{\gamma}}$ by solving a penalized least squares problem in (4). Suppose we use data $\{\mathbf{X}_i, \mathbf{Z}_i\}_{i=1}^{n'}$ to estimate $\boldsymbol{W}$. The k-th column of $\boldsymbol{W}$ can be estimated by

$$\hat{\boldsymbol{W}}_k = \underset{\boldsymbol{W}_k \in \mathbb{R}^{p_{\boldsymbol{\gamma}}}}{\operatorname{argmin}} \frac{1}{2n'}\sum_{i=1}^{n'}(X_{ik} - \boldsymbol{W}_k^\top \mathbf{Z}_i)^2 + \lambda_{X_k}\|\boldsymbol{W}_k\|_1, \tag{16}$$

where X_{ik} is the k-th component of $\mathbf{X}_i$, and λ_{X_k} is the tuning parameter.

5.1. *Limiting null distribution*

We let $\|\boldsymbol{W}\|_0 \leq s'$, where s' is a positive integer and represents the sparsity level of $\boldsymbol{W}$. Let $\varphi^2 = \max_{1\leq k\leq p_{\boldsymbol{\gamma}}} \|\mathbb{E}(Z_k \mathbf{Z})\|_2^2$ and $\vartheta^2 = s^2 s'(\log p_{\boldsymbol{\gamma}})^2\varphi^2/n'$. Next, under the following assumption, we study the asymptotic null distribution of the test statistic M_n.

Assumption 7. The estimator $\hat{\boldsymbol{W}}$ is independent of the data $\{\mathbf{X}_i, \mathbf{Z}_i, Y_i\}_{i=1}^n$, and $\|\hat{\boldsymbol{W}} - \boldsymbol{W}\|_F = O_p(\sqrt{s' \log p_{\boldsymbol{\gamma}}/n'})$ for some positive integer n'.

Assumption 7 requires the Frobenius norm bound of $\hat{\boldsymbol{W}}$ in the order of $(s' \log p_{\boldsymbol{\gamma}}/n')^{1/2}$, which can be satisfied by most of existing high-dimensional estimators. See section 9.7 in [36] for instance. Besides, n' represents the sample size used to estimate $\boldsymbol{W}$. We can estimate $\boldsymbol{W}$ using additional data of $\mathbf{X}$ and $\mathbf{Z}$ if we have, otherwise, applying the sample-splitting approach. See, for instance, [2] and [12]. We also note that the independence between $\hat{\boldsymbol{W}}$ and $\{\mathbf{X}_i, \mathbf{Z}_i, Y_i\}_{i=1}^n$ can make the asymptotic analysis more easily. However, the simulation study shows that the proposed test statistic M_n still works well numerically even when we estimate $\boldsymbol{W}$ based on the same data set. Therefore, we guess this independence might not be necessary, although we have not yet got rid of it in the technical deduction.

Theorem 5.1. *Under $\mathbb{H}_0$, Assumptions 2-5, 6, and 7, condition (6), and*

$$\vartheta^2 \vee (\log p_{\boldsymbol{\gamma}})^{1/2}\vartheta\left(\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right)^{1/2} \vee \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'} = o(\sqrt{\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}), \tag{16}$$

we have

$$\frac{M_n}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \textit{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}, p_{\boldsymbol{\gamma}}) \to \infty$, where $\boldsymbol{\Lambda}_{\boldsymbol{\eta}}$ is defined in Theorem 4.1.

Compared with Theorem 2.1, condition (5) is replaced by condition (16) that handles the impact caused by the estimation error of $\hat{\boldsymbol{W}}$. As discussed, the error term caused by the correlation between $\mathbf{X}$ and $\mathbf{Z}$ is reduced by the orthogonalization technique, but the estimation error is brought by $\hat{\boldsymbol{W}}$. Now we compare condition (16) with (5). By the formula of these conditions, it suffices to compare ϑ^2 with ϖ^2. Note that

$$\frac{\vartheta^2}{\varpi^2} = \frac{\varphi^2}{\varrho^2}\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}. \tag{17}$$

If ratio (17) is small, condition (16) can be weaker than (5). When the relationship between nuisance covariates is weak while the relationship between covariates of interest and nuisance covariates is high, the ratio (17) can be small. When the matrix $\boldsymbol{W}$ is sparse, or n' is large, the ratio (17) can also be small.

Similar to the test construction in Section 2.1, we estimate $\boldsymbol{\Lambda}_{\boldsymbol{\eta}}$ by

$$R_n = \hat{\sigma}^4\frac{1}{2\binom{n}{4}}\sum_{i<j<k<l}^{n}(\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}}_j)^\top(\hat{\boldsymbol{\eta}}_k - \hat{\boldsymbol{\eta}}_l)(\hat{\boldsymbol{\eta}}_j - \hat{\boldsymbol{\eta}}_k)^\top(\hat{\boldsymbol{\eta}}_l - \hat{\boldsymbol{\eta}}_i),$$

where $\hat{\boldsymbol{\eta}}_i = \mathbf{X}_i - \hat{\boldsymbol{W}}^\top\mathbf{Z}_i$. Under the null hypothesis and conditions in Theorem 5.1, R_n is a ratio consistent estimator of $\Lambda_{\boldsymbol{\eta}}$; see the details in Supplementary Material. We reject $\mathbb{H}_0$ at the significance level α if

$$M_n \geq z_\alpha\sqrt{2R_n}.$$

Here z_α is the upper-α quantile of standard normal distribution.

5.2. Power analysis

Consider the class of local alternatives as follows:

$$\mathscr{L}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} \in \mathbb{R}^{p_{\boldsymbol{\beta}}} \,\middle|\, \boldsymbol{\beta} \in \mathscr{L}^{\mathrm{o}}(\boldsymbol{\beta}) \text{ and } \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta} = o\left(\frac{\Lambda_{\boldsymbol{\eta}}}{n(\vartheta^2 \vee \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})s'\log p_{\boldsymbol{\gamma}}/n')}\right)\right\}.$$

Compared to $\mathscr{L}^o(\boldsymbol{\beta})$, class $\mathscr{L}(\boldsymbol{\beta})$ has more restrictions on $\boldsymbol{\beta}$. This is because we have to handle the extra error caused by the estimation of $\boldsymbol{W}$. Similar to $\mathscr{L}^o(\boldsymbol{\beta})$, the class $\mathscr{L}(\boldsymbol{\beta})$ also illustrates a small dissimilarity between $\boldsymbol{\beta}$ and $\mathbf{0}$. Under the bounded eigenvalue assumptions of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}}$, $\mathscr{L}(\boldsymbol{\beta})$ can be simplified as

$$\mathscr{L}(\boldsymbol{\beta}) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{p_{\boldsymbol{\beta}}} \middle| \|\boldsymbol{\beta}\|_2 = o\left(\min\left(1, \sqrt{\frac{p_{\boldsymbol{\beta}}}{n(1 \vee s^2 s' (\log p_{\boldsymbol{\gamma}})^2/n')}}\right)\right)\right\}.$$

Then the power performance is stated in the following theorem.

Theorem 5.2. *Assume conditions in Theorem 5.1, and for $\boldsymbol{\beta} \in \mathscr{L}(\boldsymbol{\beta})$, we derive*

$$\frac{M_n - n\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2 \boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \textit{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}, p_{\boldsymbol{\gamma}}) \to \infty$.

According to Theorem 5.2, M_n have the same asymptotic power function with M_n^o when $\boldsymbol{\beta} \in \mathscr{L}(\boldsymbol{\beta})$. Thus under designed conditions, the test M_n has the same power performance as the oracle test M_n^o with a known $\boldsymbol{W}$ asymptotically and is more powerful than T_n.

6. Numerical studies

6.1. Simulations

We first compare the performances among four tests: (1). the score function-based test T_n in Section 2; (2). the orthogonalied score function-based test M_n in Section 5; (3). the studentized bootstrap-assisted test ST in [39]; and (4). the Wald-type test $WALD$ in [23].

Generate data from the following ultrahigh-dimensional linear model:

$$Y_i = \boldsymbol{\beta}^\top \mathbf{X}_i + \boldsymbol{\gamma}^\top \mathbf{Z}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where the covariates $\boldsymbol{V}_i = (\mathbf{X}_i^\top, \mathbf{Z}_i^\top)^\top$ are generated from the multivariate normal distribution. The details are given later, and the regression error $\epsilon_i \sim N(0,1)$ independent of $\boldsymbol{V}_i$. Denote $s_{\boldsymbol{\beta}}$ and $s_{\boldsymbol{\gamma}}$ as the sparsity levels of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively. The regression coefficients $\boldsymbol{\beta}$ are set as: $\beta_j = b_0$ for $1 \le j \le s_{\boldsymbol{\beta}}$ and $\beta_j = 0$ otherwise. Similarly let $\gamma_j = g_0$ for $1 \le j \le s_{\boldsymbol{\gamma}}$ and $\gamma_j = 0$ otherwise. Throughout the simulation study, let $s_{\boldsymbol{\gamma}} = \lfloor 5\% p_{\boldsymbol{\gamma}} \rfloor$ and $g_0 = 0.5$. There are three settings for the values of $\boldsymbol{\beta}$:

Setting 1: Consider $b_0 = 0$ to assess the empirical Type-I error.

Setting 2: Let $s_{\boldsymbol{\beta}} = \lfloor 5\% p_{\boldsymbol{\beta}} \rfloor$ and $b_0 \neq 0$ to assess the empirical power with sparse alternative.

Setting 3: Let $s_{\boldsymbol{\beta}} = \lfloor 50\% p_{\boldsymbol{\beta}} \rfloor$ and $b_0 \neq 0$ to assess the empirical power with dense alternative.

The experiment is repeated 500 times for each simulation setting to assess the empirical type-I error and power at the significance level $\alpha = 0.05$. The tuning parameter λ_Y in (4) and λ_{X_k} in (16) are selected by 10-fold cross-validations using the R-package `glmnet` [18]. Based on these settings, we consider the following two scenarios.

Scenario 1. We aim to compare our tests with other testing methods in this scenario. The covariates are generated from the multivariate normal distribution $N_p(\mathbf{0}_p, \mathbf{\Sigma})$. Here $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ follows the Toeplitz design, that is, $\sigma_{ij} = 0.5^{|i-j|}$, $i, j = 1, \ldots, p$. The sample size $n = 100$, the covariate dimension $p = 600$ and $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}} = 300$. In the sparse alternative (setting 2) and the dense alternative (setting 3), we vary b_0 from 0 to $\sqrt{\|\boldsymbol{\gamma}\|_2^2/s_{\boldsymbol{\beta}}}$.

Scenario 2. This scenario investigates the performance of our tests when $\mathbf{X}$ and $\mathbf{Z}$ are highly correlated. The covariates are generated according to the following model:

$$\mathbf{X} = \boldsymbol{W}^\top \mathbf{Z} + \boldsymbol{\eta}, \tag{18}$$

where $\boldsymbol{\eta}$ is a $p_{\boldsymbol{\beta}}$-dimensional random vector and $\boldsymbol{\eta}$ is independent of $\mathbf{Z}$. $\boldsymbol{\eta} \sim N_{p_{\boldsymbol{\beta}}}(\mathbf{0}_{p_{\boldsymbol{\beta}}}, \mathbf{\Sigma}_{\boldsymbol{\eta}})$ and $\mathbf{Z} \sim N_{p_{\boldsymbol{\gamma}}}(\mathbf{0}_{p_{\boldsymbol{\gamma}}}, \mathbf{\Sigma}_{\mathbf{Z}})$, where $\mathbf{\Sigma}_{\boldsymbol{\eta}}$ and $\mathbf{\Sigma}_{\mathbf{Z}}$ follow the Toeplitz design with $\rho = 0.5$ respectively. $\boldsymbol{W}$ is defined in (10) in subsection 2.3. Throughout the scenario, $d_{\mathbf{Z}} = 3$ and $d_{\mathbf{X}} = 10$. The sample size $n = 100$, the predictor dimension $p = 600$ and $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}} = 300$. In the sparse alternative (setting 2) and the dense alternative (setting 3), we vary b_0 from 0 to $\sqrt{\|\boldsymbol{\gamma}\|_2^2/s_{\boldsymbol{\beta}}}$.

Figure 2 displays the empirical size-power curves of the four tests in scenario 1. It can be observed that T_n, M_n and ST tests control the size well. T_n and M_n are generally more powerful than ST under the sparse and dense alternative hypotheses. Under the dense alternative, the empirical powers of ST can be as low as the significance level. The empirical powers of T_n and M_n increase quickly as the signal strength b_0 becomes stronger. On the other hand, the $WALD$ test is very liberal to have very large empirical size when we use the tuning parameter $\tau = 1$ recommended by [23]. While the numerical studies in [23] suggest that the $WALD$ test with $\tau = 1$ can be very conservative in their setting. We have also conducted different settings with different dimensions and sample sizes, and found that with different values $\tau = 0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5$ the test can be either very liberal or very conservative. Thus selecting a proper tuning parameter is difficult in general. It is worth noticing that T_n and M_n have similar performances in this scenario. T_n performs well enough when the correlation between $\mathbf{X}$ and $\mathbf{Z}$ is relatively weak.

Figure 3 displays the empirical size-power curves of T_n and M_n in Scenario 2. We find that T_n is too liberal to maintain the significance level. On the contrary, M_n maintains the level well. As b_0 increases, although the empirical powers of T_n and M_n increase rapidly, the empirical power of T_n does not go to 1 as the increase of b_0. In contrast, the empirical power of M_n can increase to 1 quickly. The results show that M_n can also improve the power compared with T_n. This confirms the theory.

The above simulation studies conclude that our proposed tests perform well

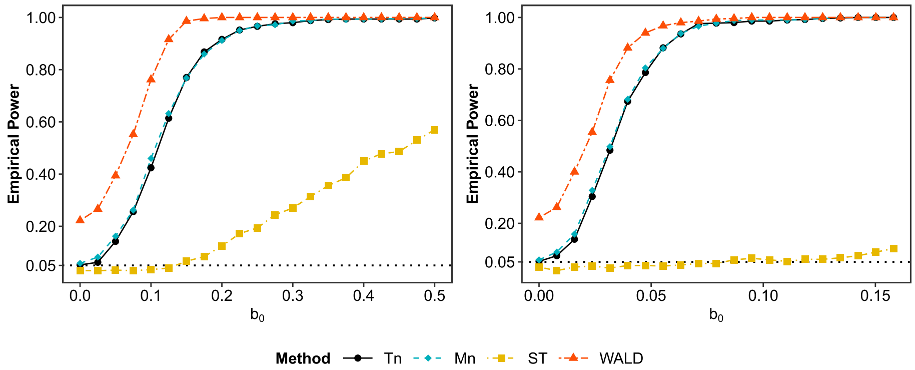

Fig 2. *The left panel represents empirical sizes and powers of the T_n, M_n, ST and WALD in the sparse alternative (setting 2). The right panel corresponds to dense alternative (setting 3). The solid line with circle points, dash line with diamond points, dot-dash line with square points and two-dash line with triangle points represent the empirical sizes and powers of T_n, M_n, ST and WALD, respectively.*

even when the testing and nuisance parameters are both ultrahigh-dimensional. When a relatively high correlation exists between the covariates of interest and the nuisance covariates, M_n can enhance the power performance over T_n. These confirm the merits of the orthogonalization technique.

6.2. Real data analysis

We have employed our tests on the identical dataset as utilized in [24], [35] and [22], for investigating the role of DNA methylation in the regulation of human stress reactivity. This dataset is accessible for download at `https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-77445` and encompasses 385882 DNA methylation loci along with various variables pertaining to 85 individuals. [24] conducted a study to explore the role of DNA methylation in cortisol stress reactivity and its relationship with childhood trauma. Besides, it is of substantial interests to study the interactions between environmental factors and DNA methylation in genome-wide DNA methylation analysis. Motivated by these observations, this article primarily focuses on the identification of interactions between DNA methylation and childhood trauma.

In this context, the variable X_e represents the childhood trauma as a one-dimensional score from a childhood trauma questionnaire, and the response variable Y denotes the increased area under the curve (iAUC) in cortisol after a stress test. Eight confounding variables are considered, which include age (Z_1), sex (Z_2), B cell proportion (Z_3), CD4 T cell proportion (Z_4), CD8 T cell proportion (Z_5), Monocytes cell proportion (Z_6), Granulocytes cell proportion (Z_7) and Natural Killer cell proportion (Z_8). Let $\mathbf{Z} = (Z_1, \dots, Z_8)^\top$ be the random vector comprising these eight confounding variables. The 385882 methylation sites are divided into 1930 sets, where 1929 sets contain 200 methylation sites

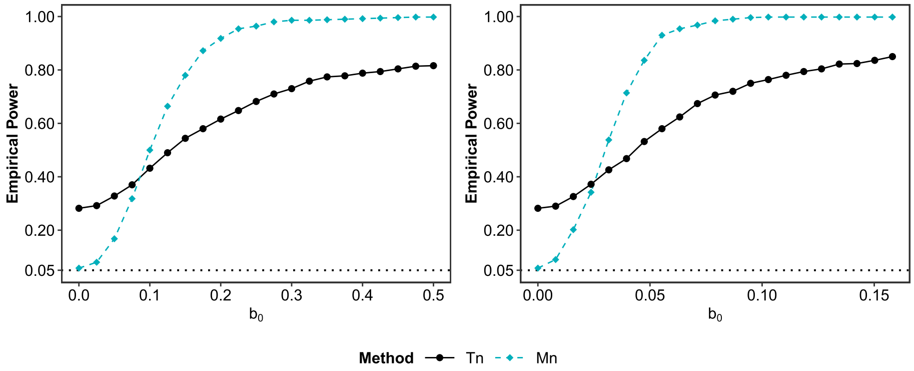

FIG 3. *The left panel represents empirical sizes and powers of the T_n and M_n in the sparse alternatives (setting 2). The right panel corresponds to dense alternatives (setting 3). The solid line with circle points and dash line with diamond points represent the empirical sizes and powers of T_n and M_n, respectively.*

each, and one set contains 82 methylation sites. Let $\mathbf{X}_{g_k}$ be the random vector for the kth DNA methylation set. To illustrate the interactions between DNA methylation and childhood trauma, we use $\mathbf{X}_{g_k} \times X_e$ to denote the interaction between the kth DNA methylation set and childhood trauma. For the kth DNA methylation set, the aim is to test whether it interacts with childhood trauma. This can be stated as:

H_{0k} : The kth DNA methylation set does not interact with childhood trauma.

Specifically, the model for the kth methylation set is represented as follows:

$$Y = \boldsymbol{\beta}_k^\top \mathbf{X}_{g_k} \times X_e + \gamma_{e_k} X_e + \boldsymbol{\gamma}_{g_k}^\top \mathbf{X}_{g_k} + \boldsymbol{\gamma}_{c_k}^\top \mathbf{Z} + \epsilon_k,$$

where ϵ_k is the error term, and $(\boldsymbol{\beta}_k^\top, \gamma_{e_k}, \boldsymbol{\gamma}_{g_k}^\top, \boldsymbol{\gamma}_{c_k}^\top)^\top$ is the coefficient vector with $k \in \{1, \ldots, 1930\}$. In the given model setting, the null hypothesis of interest is transformed as $H_{0k} : \boldsymbol{\beta}_k = \mathbf{0}$. The testing parameter is $\boldsymbol{\beta}_k$, and the nuisance parameter is $\boldsymbol{\gamma}_k = (\gamma_{e_k}, \boldsymbol{\gamma}_{g_k}^\top, \boldsymbol{\gamma}_{c_k}^\top)^\top$.

We employ T_n and M_n tests. The data is standardized, and a total of 1930 tests are conducted. Figure 4 illustrates the empirical distributions of the negative base-10 logarithm of the p-values for T_n and M_n, respectively. After Bonferroni correction, with a significant level of p-value $< \alpha_{\text{bon}} = 2.591 \times 10^{-5}$, out of the 1930 tests, we identify 126 significant DNA methylation sets using the T_n test and 149 significant sets using the M_n test. Notably, M_n identifies 18% more significant DNA methylation sets than the T_n test. The specific numbers of significant DNA methylation sets are provided in Section E in the Appendix. It is evident that M_n can detect the vast majority of sets that are significant according to the T_n test. Moreover, there are many sets that are exclusively identified by the M_n test and not by the T_n test. These results affirm the excellent performance of M_n.

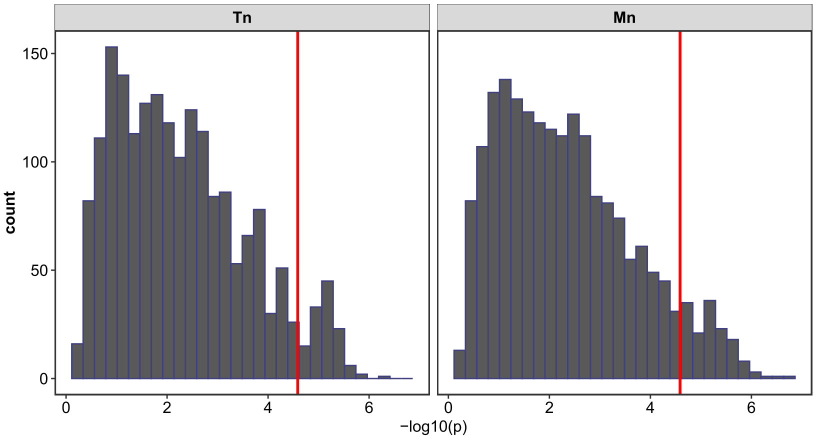

Fig 4. *The left panel illustrates the empirical distribution of the negative base-10 logarithm of the p-values for the T_n test, while the right panel corresponds to the M_n test. In both panels, the red solid line represents $-\log_{10}(\alpha_{bon})$, where $\alpha_{bon} = 2.591 \times 10^{-5}$ signifies the Bonferroni-adjusted significance level.*

7. Conclusions

This paper considers testing the significance of ultrahigh-dimensional parameter vector of interest with ultrahigh-dimensional nuisance parameter vector. We first reanalyze the score function-based test under weaker conditions to show the limiting distributions under the null and local alternative hypotheses. We construct an orthogonalized score function-based test to handle the correlation between the covariates of interest and nuisance covariates. Our investigation shows that the orthogonalization technique can debiase the error term, convert the non-degenerate error terms to degenerate, and reduce the variance to achieve higher power than the non-orthogonalized score function-based test.

Our procedure is very generic. Extensions to other regression models such as generalized linear regression models and partially linear single-index regression models [15] are possible. We would investigate these extensions in near future.

Appendix

Section A presents some notations. Section B provides the proof of theorems in the main text. Section C consists of some technical lemmas which are used in the proof of theorems in the main text. Section D presents additional simulation results. Section E gives numbers of significant DNA methylation sets in real data analysis in the main text. Section F contains an additional real data analysis.

Appendix A: Notation

For functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ to mean that $f(n) \leq cg(n)$ for some universal constant $c \in (0,\infty)$, and similarly, $f(n) \gtrsim g(n)$ when $f(n) \geq c'g(n)$ for some universal constant $c' \in (0,\infty)$. We write $f(n) \asymp g(n)$ when $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold simultaneously. The *sub-Gaussian norm* of a *sub-Gaussian* random variable X is defined as $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$. Similarly, the *sub-Exponential norm* of a *sub-Exponential* random variable Y is defined as $\|Y\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}\exp(|Y|/t) \leq 2\}$. The *sub-Gaussian norm* of a *sub-Gaussian* random vector $\mathbf{X}$ in $\mathbb{R}^p$ is defined as $\|\mathbf{X}\|_{\psi_2} := \sup_{x\in\mathbb{S}^{p-1}}\|x^\top \mathbf{X}\|_{\psi_2}$. The L_p*-norm* of a random variable X in $\mathbb{R}$ is defined as $\|X\|_p := (\mathbb{E}|X|^p)^{1/p}$. For $x \in \mathbb{R}$, $\lceil x \rceil$ represents the least integer greater than or equal to x. For $p_1 \times p_2$ dimensional matrix $\mathbf{A}$, write $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq p_1, 1\leq j\leq p_2}|A_{ij}|$ to denote the element-max norm of $\mathbf{A}$, where A_{ij} is the (i,j)-th element of $\mathbf{A}$.

Appendix B: Proof of theorems in the main text

To simplify the representations of the proofs, we give nine lemmas in Section C. Therefore, we will cite them in the proofs. For brevity, assume $\|\boldsymbol{\nu}\|_{\psi_2} = 1$. To simplify the notation, let $\epsilon_i = Y_i - \boldsymbol{\beta}^\top \mathbf{X}_i - \boldsymbol{\gamma}^\top \mathbf{Z}_i$, $\hat{\epsilon}_i = Y_i - \hat{\boldsymbol{\gamma}}^\top \mathbf{Z}_i$ and $\breve{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_\phi - \hat{\boldsymbol{\gamma}}$. Under $\mathbb{H}_0$, $\epsilon_i = Y_i - \boldsymbol{\gamma}^\top \mathbf{Z}_i$. Let $\boldsymbol{\eta}_i = \mathbf{X}_i - \boldsymbol{W}^\top \mathbf{Z}_i$, $\hat{\boldsymbol{\eta}}_i = \mathbf{X}_i - \hat{\boldsymbol{W}}^\top \mathbf{Z}_i$, $\boldsymbol{\Gamma}_{\boldsymbol{\eta}} = \boldsymbol{\Gamma}_{\mathbf{X}} - \boldsymbol{W}^\top \boldsymbol{\Gamma}_{\mathbf{Z}}$, $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}} = \boldsymbol{\Gamma}_{\mathbf{X}} - \hat{\boldsymbol{W}}^\top \boldsymbol{\Gamma}_{\mathbf{Z}}$ and $\breve{\boldsymbol{W}} = \boldsymbol{W} - \hat{\boldsymbol{W}}$, where $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_{\mathbf{X}}^\top, \boldsymbol{\Gamma}_{\mathbf{Z}}^\top)^\top$ and $\boldsymbol{\Gamma}$ is defined in Assumption 2. Write $\mathbf{X}_i = \boldsymbol{\Gamma}_{\mathbf{X}}\boldsymbol{\nu}_i$, $\mathbf{Z}_i = \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\nu}_i$, $\boldsymbol{\eta}_i = \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\nu}_i$ and $\hat{\boldsymbol{\eta}}_i = \boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{\nu}_i$. Further, write $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}_{\mathbf{X}}\boldsymbol{\Gamma}_{\mathbf{X}}^\top$, $\boldsymbol{\Sigma}_{\mathbf{Z}} = \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}} = \boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}^\top$. Write $q = \lceil 4/(1-3b) \rceil$.

B.1. Proofs of Theorem 2.1

As $\boldsymbol{\gamma}_\phi = \boldsymbol{\gamma}$ and $\epsilon_i = Y_i - \boldsymbol{\gamma}^\top \mathbf{Z}_i$ under $\mathbb{H}_0$, we decompose T_n it as

$$\begin{aligned} T_n &= \frac{1}{n}\sum_{i\neq j}\hat{\epsilon}_i\hat{\epsilon}_j\mathbf{X}_i^\top\mathbf{X}_j = \frac{1}{n}\sum_{i\neq j}(\epsilon_i + \breve{\boldsymbol{\gamma}}^\top\mathbf{Z}_i)(\epsilon_j + \breve{\boldsymbol{\gamma}}^\top\mathbf{Z}_j)\mathbf{X}_i^\top\mathbf{X}_j \\ &= \frac{1}{n}\sum_{i\neq j}\epsilon_i\epsilon_j\mathbf{X}_i^\top\mathbf{X}_j + \frac{1}{n}\sum_{i\neq j}\breve{\boldsymbol{\gamma}}^\top\mathbf{Z}_i\breve{\boldsymbol{\gamma}}^\top\mathbf{Z}_j\mathbf{X}_i^\top\mathbf{X}_j \\ &\quad + \frac{1}{n}\sum_{i\neq j}(\epsilon_i\breve{\boldsymbol{\gamma}}^\top\mathbf{Z}_j + \breve{\boldsymbol{\gamma}}^\top\mathbf{Z}_i\epsilon_j)\mathbf{X}_i^\top\mathbf{X}_j \\ &=: I_1^{\mathrm{c}} + I_2^{\mathrm{c}} + I_3^{\mathrm{c}}. \end{aligned}$$

Similar to the proof of Theorem 3 in [20], we have

$$\frac{I_1^{\mathrm{c}}}{\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}} \to N(0,1) \quad \text{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}) \to \infty$. Lemma C.1 in the Supplementary Material gives the detailed proof about I_1^{c}. We now prove that I_2^{c} and I_3^{c} are $o_p(\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}})$.

Following Lemma C.10 in the Supplementary Material, we have

$$I_2^{\mathrm{c}} \leq \left\| \frac{1}{n} \sum_{i \neq j} \mathbf{Z}_i \mathbf{Z}_j^\top \mathbf{X}_i^\top \mathbf{X}_j \right\|_\infty \|\breve{\gamma}\|_1^2, \tag{19}$$

Let $\mathbf{U}_1^{\mathrm{c}}$ be a $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\gamma}}$ dimension matrix with (k_1, k_2)-th element

$$U_{1,(k_1,k_2)}^{\mathrm{c}} = \frac{1}{n} \sum_{i \neq j} \frac{1}{2} (Z_{i,k_1} Z_{j,k_2} + Z_{i,k_2} Z_{j,k_1}) \mathbf{X}_i^\top \mathbf{X}_j.$$

Note that $\mathbf{U}_1^{\mathrm{c}} = n^{-1} \sum_{i \neq j} \mathbf{Z}_i \mathbf{Z}_j^\top \mathbf{X}_i^\top \mathbf{X}_j$ and all of its elements are U-statistics. By Hoeffding decomposition, we derive

$$\frac{1}{n-1} U_{1,(k_1,k_2)}^{\mathrm{c}} = \mathbb{E}(Z_{k_1}\mathbf{X})^\top \mathbb{E}(Z_{k_2}\mathbf{X}) + 2S_{1,1,(k_1,k_2)}^{\mathrm{c}} + S_{1,2,(k_1,k_2)}^{\mathrm{c}},$$

where

$$S_{1,1,(k_1,k_2)}^{\mathrm{c}} = \frac{1}{n} \sum_{i=1}^n g_{1,1,(k_1,k_2)}^{\mathrm{c}} (\mathbf{X}_i, \mathbf{Z}_i)$$

with $g_{1,1,(k_1,k_2)}^{\mathrm{c}}(\mathbf{X}_1, \mathbf{Z}_1) = \{Z_{1,k_1}\mathbf{X}_1 - \mathbb{E}(Z_{k_1}\mathbf{X})\}^\top \mathbb{E}(Z_{k_2}\mathbf{X})/2 + \{Z_{1,k_2}\mathbf{X}_1 - \mathbb{E}(Z_{k_2}\mathbf{X})\}^\top \mathbb{E}(Z_{k_1}\mathbf{X})/2$,

$$S_{1,2,(k_1,k_2)}^{\mathrm{c}} = \frac{1}{n(n-1)} \sum_{i \neq j}^n g_{1,2,(k_1,k_2)}^{\mathrm{c}} (\mathbf{X}_i, \mathbf{Z}_i; \mathbf{X}_j, \mathbf{Z}_j)$$

with $g_{1,2,(k_1,k_2)}^{\mathrm{c}}(\mathbf{X}_1, \mathbf{Z}_1; \mathbf{X}_2, \mathbf{Z}_2) = \{Z_{1,k_1}\mathbf{X}_1 - \mathbb{E}(Z_{k_1}\mathbf{X})\}^\top \{Z_{2,k_2}\mathbf{X}_2 - \mathbb{E}(Z_{k_2}\mathbf{X})\} /2 + \{Z_{1,k_2}\mathbf{X}_1 - \mathbb{E}(Z_{k_2}\mathbf{X})\}^\top \{Z_{2,k_1}\mathbf{X}_2 - \mathbb{E}(Z_{k_1}\mathbf{X})\}/2$.

For any $1 \leq k_1, k_2 \leq p_{\boldsymbol{\gamma}}$, Z_{1,k_1} is a sub-Gaussian random variable and $\mathbb{E}(Z_{k_2}\mathbf{X})^\top \mathbf{X}$ is a sub-Gaussian random variable with norm $\|\mathbb{E}(Z_{k_2}\mathbf{X})^\top \mathbf{X}\|_{\psi_2} \lesssim \{\mathbb{E}(Z_{k_2}\mathbf{X})^\top \boldsymbol{\Sigma}_{\mathbf{X}} \mathbb{E}(Z_{k_2}\mathbf{X})\}^{1/2}$. Thus $\{Z_{1,k_1}\mathbf{X}_1 - \mathbb{E}(Z_{k_1}\mathbf{X})\}^\top \mathbb{E}(Z_{k_2}\mathbf{X})$ is sub-Exponential with norm

$$\begin{aligned} \|\{Z_{1,k_1}\mathbf{X}_1 - \mathbb{E}(Z_{k_1}\mathbf{X})\}^\top \mathbb{E}(Z_{k_2}\mathbf{X})\|_{\psi_1} &\leq \|Z_{1,k_1}\|_{\psi_2} \|\mathbb{E}(Z_{k_2}\mathbf{X})^\top \mathbf{X}\|_{\psi_2} \\ &\lesssim \{\mathbb{E}(Z_{k_2}\mathbf{X})^\top \boldsymbol{\Sigma}_{\mathbf{X}} \mathbb{E}(Z_{k_2}\mathbf{X})\}^{1/2}. \end{aligned} \tag{20}$$

Therefore we derive

$$\begin{aligned} \|\max_{1\leq i\leq n} \max_{1\leq k_1,k_2\leq p_{\boldsymbol{\gamma}}} g_{1,1,(k_1,k_2)}^{\mathrm{c}}\|_2 &\lesssim \log(np_{\boldsymbol{\gamma}}) \max_{1\leq i\leq n} \max_{1\leq k_1,k_2\leq p_{\boldsymbol{\gamma}}} \|g_{1,1,(k_1,k_2)}^{\mathrm{c}}\|_{\psi_1} \\ &\lesssim \log(np_{\boldsymbol{\gamma}}) \max_{1\leq k\leq p_{\boldsymbol{\gamma}}} \{\mathbb{E}(Z_k\mathbf{X})^\top \boldsymbol{\Sigma}_{\mathbf{X}} \mathbb{E}(Z_k\mathbf{X})\}^{1/2} \\ &\lesssim \log(np_{\boldsymbol{\gamma}}) \lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}) \varrho. \end{aligned} \tag{21}$$

The first inequality follows from Lemma C.9 in the Supplementary Material. The second inequality holds by inequality (20).

Denote $q = \lceil 4/(1-3b) \rceil$. We have

$$\begin{aligned}
&\|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} g^{\mathrm{c}}_{1,2,(k_1,k_2)}\|_4\\
&\quad\le \|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} g^{\mathrm{c}}_{1,2,(k_1,k_2)}\|_q\\
&\quad\lesssim \|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} Z_{i,k_1}Z_{j,k_2}\mathbf{X}_i^\top\mathbf{X}_j\|_q\\
&\quad\lesssim \|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} Z_{i,k_1}Z_{j,k_2}\|_{2q}\|\max_{1\le i\neq j\le n}\mathbf{X}_i^\top\mathbf{X}_j\|_{2q}\\
&\quad\lesssim \log(np_{\boldsymbol{\gamma}})\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|Z_{1,k_1}Z_{2,k_2}\|_{\psi_1} n^{1/q}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}})\\
&\quad\lesssim \log(np_{\boldsymbol{\gamma}})n^{1/q}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}}).
\end{aligned}\tag{22}$$

The first inequality holds by Liapounov inequality. The third inequality holds by Cauchy-Schwartz inequality, the fourth inequality holds by the property of Orlicz norm (Page 96 in [34]), Lemmas C.9 and C.4 in the Supplementary Material. The last inequality holds by Assumption 2. Similarly, we derive

$$\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|g^{\mathrm{c}}_{1,1,(k_1,k_2)}\|_2 \lesssim \lambda^{1/2}_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho \tag{23}$$

and

$$\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|g^{\mathrm{c}}_{1,2,(k_1,k_2)}\|_2 \le \max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|g^{\mathrm{c}}_{1,2,(k_1,k_2)}\|_4 \lesssim \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}}). \tag{24}$$

Applying Lemma C.2 in the Supplementary Material, we derive

$$\begin{aligned}
\mathbb{E}(\|\mathbf{U}^{\mathrm{c}}_1\|_\infty) \lesssim& n\varrho^2 + \{n^{1/2}(\log p_{\boldsymbol{\gamma}})^{1/2} + \log p_{\boldsymbol{\gamma}}\log(np_{\boldsymbol{\gamma}})\}\lambda^{1/2}_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho\\
&+ \{\log p_{\boldsymbol{\gamma}} + n^{-1/2+1/q}(\log p_{\boldsymbol{\gamma}})^{3/2}\log(np_{\boldsymbol{\gamma}})\}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}})\\
\lesssim& n\varrho^2 + n^{1/2}(\log p_{\boldsymbol{\gamma}})^{1/2}\lambda^{1/2}_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho + \log p_{\boldsymbol{\gamma}}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}}),
\end{aligned}\tag{25}$$

where the first inequality holds by Lemma C.2 in the Supplementary Material and the inequalities (21)–(24). The last inequality holds by the fact $\log p_{\boldsymbol{\gamma}} = O(n^b)$, where $0 < b < 1/3$. Combining equations (19) and (25), Assumption 3, we have

$$I^{\mathrm{c}}_2 = O_p\{s^2\log p_{\boldsymbol{\gamma}}\varrho^2 + n^{-1/2}s^2(\log p_{\boldsymbol{\gamma}})^{3/2}\lambda^{1/2}_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho + n^{-1}s^2(\log p_{\boldsymbol{\gamma}})^2\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}})\}. \tag{26}$$

Thus $I^{\mathrm{c}}_2 = o_p(\sqrt{\boldsymbol{\Lambda}_{\mathbf{X}}})$ when $s^2\log p_{\boldsymbol{\gamma}}\varrho^2 = o(\sqrt{\boldsymbol{\Lambda}_{\mathbf{X}}})$, $n^{-1/2}s^2(\log p_{\boldsymbol{\gamma}})^{3/2}\lambda^{1/2}_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho = o(\sqrt{\boldsymbol{\Lambda}_{\mathbf{X}}})$ and $n^{-1}s^2(\log p_{\boldsymbol{\gamma}})^2 = o(1)$ hold simultaneously.

Similar to the proof of I^{c}_2, we derive

$$I^{\mathrm{c}}_3 \le \left\|\frac{1}{n}\sum_{i\neq j}(\epsilon_i\mathbf{Z}_j + \mathbf{Z}_i\epsilon_j)\mathbf{X}_i^\top\mathbf{X}_j\right\|_\infty \|\check{\boldsymbol{\gamma}}\|_1. \tag{27}$$

Let $\mathbf{U}_2^{\mathrm{c}}$ be a $p_{\boldsymbol{\gamma}}$ dimension vector with k-th element

$$U_{2,k}^{\mathrm{c}} = \frac{1}{n}\sum_{i\neq j}(\epsilon_i Z_{j,k} + Z_{i,k}\epsilon_j)\mathbf{X}_i^\top \mathbf{X}_j.$$

Note that $\mathbf{U}_2^{\mathrm{c}} = n^{-1}\sum_{i\neq j}(\epsilon_i \mathbf{Z}_j + \mathbf{Z}_i\epsilon_j)\mathbf{X}_i^\top \mathbf{X}_j$ and all of its elements are U-statistics. By Hoeffding decomposition again, we derive

$$\frac{1}{n-1}U_{2,k}^{\mathrm{c}} = 2S_{2,1,k}^{\mathrm{c}} + S_{2,2,k}^{\mathrm{c}},$$

where

$$S_{2,1,k}^{\mathrm{c}} = \frac{1}{n}\sum_{i=1}^n g_{2,1,k}^{\mathrm{c}}(\mathbf{X}_i, \mathbf{Z}_i, \epsilon_i)$$

with $g_{2,1,k}^{\mathrm{c}}(\mathbf{X}_1, \mathbf{Z}_1, \epsilon_1) = \epsilon_1 \mathbf{X}_1^\top \mathbb{E}(Z_k \mathbf{X})$,

$$S_{2,2,k}^{\mathrm{c}} = \frac{1}{n(n-1)}\sum_{i\neq j}^n g_{2,2,k}^{\mathrm{c}}(\mathbf{X}_i, \mathbf{Z}_i, \epsilon_i; \mathbf{X}_j, \mathbf{Z}_j, \epsilon_j)$$

with $g_{2,2,k}^{\mathrm{c}}(\mathbf{X}_1, \mathbf{Z}_1, \epsilon_1; \mathbf{X}_2, \mathbf{Z}_2, \epsilon_2) = \epsilon_1 \mathbf{X}_1^\top\{Z_{2,k}\mathbf{X}_2 - \mathbb{E}(Z_k\mathbf{X})\} + \epsilon_2\mathbf{X}_2^\top\{Z_{1,k}\mathbf{X}_1 - \mathbb{E}(Z_k\mathbf{X})\}$.

Similar to the proof of inequality (21), we derive

$$\begin{aligned}\|\max_{1\le i\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} g_{2,1,k}^{\mathrm{c}}\|_2 &\lesssim \|\max_{1\le i\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} \epsilon_i \mathbf{X}_i^\top \mathbb{E}(Z_k\mathbf{X})\|_2 \\ &\lesssim \log(np_{\boldsymbol{\gamma}})\max_{1\le k\le p_{\boldsymbol{\gamma}}} \lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho. \qquad (28)\end{aligned}$$

The last inequality is derived by Assumption 5, Lemma C.9 in the Supplementary Material, and the technique used in the last inequality in the proof for the inequality (21). Similar to the proof of (22), we derive

$$\begin{aligned}\|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} g_{2,2,k}^{\mathrm{c}}\|_4 &\le \|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} g_{2,2,k}^{\mathrm{c}}\|_q \\ &\lesssim \|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} \epsilon_1 Z_{2,k}\mathbf{X}_1^\top \mathbf{X}_2\|_q \\ &\lesssim \log(np_{\boldsymbol{\gamma}})n^{1/q}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}^2). \qquad (29)\end{aligned}$$

Similarly,

$$\max_{1\le k\le p_{\boldsymbol{\gamma}}} \|g_{2,1,k}^{\mathrm{c}}\|_2 \lesssim \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho^2 \qquad (30)$$

and

$$\max_{1\le k\le p_{\boldsymbol{\gamma}}} \|g_{2,2,k}^{\mathrm{c}}\|_2 \le \max_{1\le k\le p_{\boldsymbol{\gamma}}} \|g_{2,2,k}^{\mathrm{c}}\|_4 \lesssim \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}^2). \qquad (31)$$

The argument for proving the inequality (25) yields

$$\mathbb{E}(\|\mathbf{U}_2^{\mathrm{c}}\|_\infty) \lesssim n^{1/2}(\log p_{\boldsymbol{\gamma}})^{1/2}\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho + \log p_{\boldsymbol{\gamma}}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}^2). \qquad (32)$$

Combining equations (27) and (32), Assumption 3, we have

$$I_3^{\mathrm{c}} = O_p\big\{s\log p_{\boldsymbol{\gamma}}\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho + n^{-1/2}s(\log p_{\boldsymbol{\gamma}})^{3/2}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}^2)\big\}. \tag{33}$$

Thus $I_3^{\mathrm{c}} = o_p(\sqrt{\boldsymbol{\Lambda}_{\mathbf{X}}})$ when $s\log p_{\boldsymbol{\gamma}}\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho = o(\sqrt{\boldsymbol{\Lambda}_{\mathbf{X}}})$ and $n^{-1/2}s(\log p_{\boldsymbol{\gamma}})^{3/2} = o(1)$ hold simultaneously. The proof is concluded.

B.2. The proof of Theorem 2.2

Under the local alternatives, we decompose T_n as:

$$\begin{aligned}
T_n =& \frac{1}{n}\sum_{i\neq j}\hat{\epsilon}_i\hat{\epsilon}_j\mathbf{X}_i^\top\mathbf{X}_j \\
=& \frac{1}{n}\sum_{i\neq j}(\boldsymbol{\beta}^\top\boldsymbol{\eta}_i + \epsilon_i + \breve{\gamma}^\top\mathbf{Z}_i)(\boldsymbol{\beta}^\top\boldsymbol{\eta}_j + \epsilon_j + \breve{\gamma}^\top\mathbf{Z}_j)\mathbf{X}_i^\top\mathbf{X}_j \\
=& \frac{1}{n}\sum_{i\neq j}(\boldsymbol{\beta}^\top\boldsymbol{\eta}_i + \epsilon_i)(\boldsymbol{\beta}^\top\boldsymbol{\eta}_j + \epsilon_j)\mathbf{X}_i^\top\mathbf{X}_j \\
&+ \frac{1}{n}\sum_{i\neq j}\breve{\gamma}^\top\mathbf{Z}_i\breve{\gamma}^\top\mathbf{Z}_j\mathbf{X}_i^\top\mathbf{X}_j \\
&+ \frac{1}{n}\sum_{i\neq j}(\epsilon_i\breve{\gamma}^\top\mathbf{Z}_j + \breve{\gamma}^\top\mathbf{Z}_i\epsilon_j)\mathbf{X}_i^\top\mathbf{X}_j \\
&+ \frac{1}{n}\sum_{i\neq j}(\boldsymbol{\beta}^\top\boldsymbol{\eta}_i\breve{\gamma}^\top\mathbf{Z}_j + \breve{\gamma}^\top\mathbf{Z}_i\boldsymbol{\beta}^\top\boldsymbol{\eta}_j)\mathbf{X}_i^\top\mathbf{X}_j \\
=:& A_1^{\mathrm{c}} + A_2^{\mathrm{c}} + A_3^{\mathrm{c}} + A_4^{\mathrm{c}},
\end{aligned}$$

where the second equality holds by the fact that

$$\begin{aligned}
\hat{\epsilon}_i &= \epsilon_i + \boldsymbol{\gamma}^\top\mathbf{Z}_i - \boldsymbol{\gamma}_\phi^\top\mathbf{Z}_i + \breve{\gamma}^\top\mathbf{Z}_i + \boldsymbol{\beta}^\top\mathbf{X}_i \\
&= \epsilon_i + \big[\boldsymbol{\gamma} - \boldsymbol{\Sigma}_{\mathbf{Z}}^{-1}\mathbb{E}\{\mathbf{Z}(\mathbf{Z}^\top\boldsymbol{\gamma} + \mathbf{X}_i^\top\boldsymbol{\beta} + \epsilon_i)\}\big]^\top\mathbf{Z}_i + \breve{\gamma}^\top\mathbf{Z}_i + \boldsymbol{\beta}^\top\mathbf{X}_i \\
&= \epsilon_i - \boldsymbol{\beta}^\top\boldsymbol{W}^\top\mathbf{Z}_i + \breve{\gamma}^\top\mathbf{Z}_i + \boldsymbol{\beta}^\top\mathbf{X}_i \\
&= \boldsymbol{\beta}^\top\boldsymbol{\eta}_i + \epsilon_i + \breve{\gamma}^\top\mathbf{Z}_i.
\end{aligned} \tag{34}$$

The first term A_1^{c} can be rewritten as

$$\begin{aligned}
A_1^{\mathrm{c}} =& \frac{1}{n}\sum_{i\neq j}\epsilon_i\epsilon_j\mathbf{X}_i^\top\mathbf{X}_j + \frac{1}{n}\sum_{i\neq j}\boldsymbol{\beta}^\top\boldsymbol{\eta}_i\mathbf{X}_i^\top\mathbf{X}_j\boldsymbol{\eta}_j^\top\boldsymbol{\beta} \\
&+ \frac{1}{n}\sum_{i\neq j}(\boldsymbol{\beta}^\top\boldsymbol{\eta}_i\epsilon_j + \epsilon_i\boldsymbol{\beta}^\top\boldsymbol{\eta}_j)\mathbf{X}_i^\top\mathbf{X}_j \\
=:& A_{11}^{\mathrm{c}} + A_{12}^{\mathrm{c}} + A_{13}^{\mathrm{c}}.
\end{aligned}$$

Following Lemma C.1 below, we have

$$\frac{A^{\mathrm{c}}_{11}}{\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}} \to N(0,1) \quad \text{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}) \to \infty$. Noticing $\mathbb{E}(A^{\mathrm{c}}_{12}) = (n-1)\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^2_{\boldsymbol{\eta}} \boldsymbol{\beta}$ and applying the Hoeffding decomposition we derive

$$\frac{1}{n-1} A^{\mathrm{c}}_{12} = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^2_{\boldsymbol{\eta}} \boldsymbol{\beta} + 2A^{\mathrm{c}}_{121} + A^{\mathrm{c}}_{122},$$

where

$$A^{\mathrm{c}}_{121} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} (\mathbf{X}_i \boldsymbol{\eta}_i^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}),$$

and

$$A^{\mathrm{c}}_{122} = \frac{1}{n(n-1)} \sum_{i \neq j} (\mathbf{X}_i \boldsymbol{\eta}_i^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^\top (\mathbf{X}_j \boldsymbol{\eta}_j^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}).$$

By the variance formula of U-statistics, we derive the variance of A^{c}_{12}

$$\begin{aligned}
\operatorname{Var}(A^{\mathrm{c}}_{12}) &\lesssim n^2 \mathbb{E}(A^{\mathrm{c}}_{121})^2 + n^2 \mathbb{E}(A^{\mathrm{c}}_{122})^2 \\
&\lesssim n \mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta})^2 + \mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \mathbf{X}_2 \boldsymbol{\eta}_2^\top \boldsymbol{\beta})^2 \\
&= o(\boldsymbol{\Lambda}_{\mathbf{X}}).
\end{aligned}$$

The last equality holds by Lemma C.6 below and the model structure under the local alternatives. Similarly, we derive $\mathbb{E}(A^{\mathrm{c}}_{13}) = 0$ and

$$\frac{1}{n-1} A^{\mathrm{c}}_{13} = 2A^{\mathrm{c}}_{131} + A^{\mathrm{c}}_{132},$$

where

$$A^{\mathrm{c}}_{131} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \epsilon_i \mathbf{X}_i$$

and

$$A^{\mathrm{c}}_{132} = \frac{1}{n(n-1)} \sum_{i \neq j} \big\{ (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \mathbf{X}_i^\top - \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) \epsilon_j \mathbf{X}_j + \epsilon_i \mathbf{X}_i^\top (\mathbf{X}_j \boldsymbol{\eta}_j^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}) \big\}.$$

Similar to the derivation of $\operatorname{Var}(A^{\mathrm{c}}_{12})$, we derive

$$\begin{aligned}
\operatorname{Var}(A^{\mathrm{c}}_{13}) &\lesssim n^2 \mathbb{E}(A^{\mathrm{c}}_{131})^2 + n^2 \mathbb{E}(A^{\mathrm{c}}_{132})^2 \\
&\lesssim n \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \mathbb{E}(\epsilon_1^2 \mathbf{X}_1 \mathbf{X}_1^\top) \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} + \mathbb{E}(\epsilon_2 \boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \mathbf{X}_2)^2 \\
&\lesssim n \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} + \operatorname{tr}(\boldsymbol{\Sigma}^2_{\mathbf{X}}) \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} \\
&= o(\boldsymbol{\Lambda}_{\mathbf{X}}).
\end{aligned}$$

Where the equality holds by the model structure under the local alternatives, and the third inequality holds by using Lemma C.6, Assumption 5 and the fact that ϵ is independent of $\mathbf{X}$.

In summary, we derive.

$$\frac{A_1^{\mathrm{c}} - n\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2 \boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}} \to N(0,1) \quad \text{in distribution.}$$

To prove this theorem, it suffices to prove $A_i^{\mathrm{c}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}})$, $i = 2,3,4$. By the same argument as getting I_2^{c} and I_3^{c} in Theorem 2.1, we can show

$$A_2^{\mathrm{c}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}) \quad \text{and} \quad A_3^{\mathrm{c}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}}).$$

Turn to prove $A_4^{\mathrm{c}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\mathbf{X}}})$. Recall that

$$A_4^{\mathrm{c}} = \frac{1}{n}\sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \check{\gamma}^\top \mathbf{Z}_j + \check{\gamma}^\top \mathbf{Z}_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j) \mathbf{X}_i^\top \mathbf{X}_j.$$

By the Hölder inequality, we derive

$$|A_4^{\mathrm{c}}| \leq \left\| \frac{1}{n}\sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \mathbf{Z}_j + \mathbf{Z}_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j) \mathbf{X}_i^\top \mathbf{X}_j \right\|_\infty \|\check{\gamma}\|_1. \tag{35}$$

Let $\mathbf{U}_4^{\mathrm{c}}$ be a $p_{\boldsymbol{\gamma}}$ dimension vector with k-th element

$$U_{4,k}^{\mathrm{c}} = \frac{1}{n}\sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i Z_{j,k} + Z_{i,k} \boldsymbol{\beta}^\top \boldsymbol{\eta}_j) \mathbf{X}_i^\top \mathbf{X}_j.$$

Note that $\mathbf{U}_4^{\mathrm{c}} = n^{-1}\sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \mathbf{Z}_j + \mathbf{Z}_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j) \mathbf{X}_i^\top \mathbf{X}_j$ and all of its elements are U-statistics. By the Hoeffding decomposition, we derive

$$\frac{1}{n-1} U_{4,k}^{\mathrm{c}} = 2\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \mathbb{E}(Z_k \mathbf{X}) + 2S_{4,1,k}^{\mathrm{c}} + S_{4,2,k}^{\mathrm{c}},$$

where

$$S_{4,1,k}^{\mathrm{c}} = \frac{1}{n}\sum_{i=1}^{n} g_{4,1,k}^{\mathrm{c}}(\mathbf{X}_i, \mathbf{Z}_i)$$

with $g_{4,1,k}^{\mathrm{c}}(\mathbf{X}_1, \mathbf{Z}_1) = \mathbb{E}(Z_k \mathbf{X})^\top (\mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}) + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \{Z_{1,k} \mathbf{X}_1 - \mathbb{E}(Z_k \mathbf{X})\}$,

$$S_{4,2,k}^{\mathrm{c}} = \frac{1}{n(n-1)}\sum_{i \neq j}^{n} g_{4,2,k}^{\mathrm{c}}(\mathbf{X}_i, \mathbf{Z}_i; \mathbf{X}_j, \mathbf{Z}_j)$$

with $g_{4,2,k}^{\mathrm{c}}(\mathbf{X}_1, \mathbf{Z}_1; \mathbf{X}_2, \mathbf{Z}_2) = (\mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^\top \{Z_{2,k} \mathbf{X}_2 - \mathbb{E}(Z_k X)\} + \{Z_{1,k} \mathbf{X}_1 - \mathbb{E}(Z_k \mathbf{X})\}^\top (\mathbf{X}_2 \boldsymbol{\eta}_2^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})$.

For any $1 \leq k \leq p_{\boldsymbol{\gamma}}$, $Z_{1,k}$ is a sub-Gaussian random variable with bounded sub-Gaussian norm, $\mathbb{E}(Z_k \mathbf{X})^\top \mathbf{X}_1$ is a sub-Gaussian random variable with norm $\{\mathbb{E}(Z_k \mathbf{X})^\top \boldsymbol{\Sigma}_{\mathbf{X}} \mathbb{E}(Z_k \mathbf{X})\}^{1/2}$. Also $\boldsymbol{\beta}^\top \boldsymbol{\eta}_1$ is sub-Gaussian with norm $(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2}$, $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \mathbf{X}_1$ is sub-Gaussian with norm $(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2}$. Thus $g_{4,1,k}^{\mathrm{c}}$ is sub-Exponential with norm

$$\|g_{4,1,k}^{\mathrm{c}}\|_{\psi_1} \leq \|\mathbb{E}(Z_k \mathbf{X})^\top (\mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})\|_{\psi_1} + \|\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \{Z_{1,k} \mathbf{X}_1 - \mathbb{E}(Z_k \mathbf{X})\}\|_{\psi_1}$$

$$\begin{aligned}
&\leq \|\mathbb{E}(Z_k\mathbf{X})^\top\mathbf{X}_1\|_{\psi_2}\|\boldsymbol{\beta}^\top\boldsymbol{\eta}_1\|_{\psi_2} + \|Z_{1,k}\|_{\psi_2}\|\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\mathbf{X}_1\|_{\psi_2}\\
&\lesssim \left\{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\mathbb{E}(Z_k\mathbf{X})^\top\boldsymbol{\Sigma}_{\mathbf{X}}\mathbb{E}(Z_k\mathbf{X}) + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\right\}^{1/2}. \qquad (36)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\max_{1\leq i\leq n}\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} g^{\mathrm{c}}_{4,1,k}\|_2\\
&\quad\lesssim \log(np_{\boldsymbol{\gamma}})\max_{1\leq i\leq n}\max_{1\leq k\leq p_{\boldsymbol{\gamma}}}\|g^{\mathrm{c}}_{4,1,k}\|_{\psi_1}\\
&\quad\lesssim \log(np_{\boldsymbol{\gamma}})\max_{1\leq k\leq p_{\boldsymbol{\gamma}}}\left\{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\mathbb{E}(Z_k\mathbf{X})^\top\boldsymbol{\Sigma}_{\mathbf{X}}\mathbb{E}(Z_k\mathbf{X}) + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\right\}^{1/2}\\
&\quad\leq \log(np_{\boldsymbol{\gamma}})\left\{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho^2 + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\right\}^{1/2}. \qquad (37)
\end{aligned}$$

The first inequality follows from Lemma C.9, and the second inequality is derived by the inequality (36).

Denote $q = \lceil 4/(1-3b)\rceil$. We have

$$\begin{aligned}
\|\max_{1\leq i\neq j\leq n}\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} g^{\mathrm{c}}_{4,2,k}\|_4 &\leq \|\max_{1\leq i\neq j\leq n}\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} g^{\mathrm{c}}_{4,2,k}\|_q\\
&\lesssim \|\max_{1\leq i\neq j\leq n}\max_{1\leq k_1,k_2\leq p_{\boldsymbol{\gamma}}} Z_{i,k}\boldsymbol{\beta}^\top\boldsymbol{\eta}_j\mathbf{X}_i^\top\mathbf{X}_j\|_q\\
&\leq \|\max_{1\leq i\neq j\leq n}\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} Z_{i,k}\boldsymbol{\beta}^\top\boldsymbol{\eta}_j\|_{2q}\|\max_{1\leq i\neq j\leq n}\mathbf{X}_i^\top\mathbf{X}_j\|_{2q}\\
&\lesssim \log(np_{\boldsymbol{\gamma}})\max_{1\leq k\leq p_{\boldsymbol{\gamma}}}\|Z_{1,k}\boldsymbol{\beta}^\top\boldsymbol{\eta}_2\|_{\psi_1}n^{1/q}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}})\\
&\lesssim \log(np_{\boldsymbol{\gamma}})n^{1/q}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta})^{1/2}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}}). \qquad (38)
\end{aligned}$$

The first inequality is from the Liapounov inequality, the third inequality is due to the Cauchy-Schwartz inequality, and the fourth inequality holds by using the property of Orlicz norm(Page 96 in [34]), Lemmas C.9 and C.4. The last inequality is based on the sub-Gaussian property of $Z_{1,k}$ and $\boldsymbol{\beta}^\top\boldsymbol{\eta}_2$. Similarly, we derive

$$\max_{1\leq k\leq p_{\boldsymbol{\gamma}}}\|g^{\mathrm{c}}_{4,1,k}\|_2 \lesssim \left\{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho^2 + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\right\}^{1/2} \qquad (39)$$

and

$$\max_{1\leq k\leq p_{\boldsymbol{\gamma}}}\|g^{\mathrm{c}}_{4,2,k}\|_2 \leq \max_{1\leq k\leq p_{\boldsymbol{\gamma}}}\|g^{\mathrm{c}}_{4,2,k}\|_4 \lesssim (\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta})^{1/2}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}}). \qquad (40)$$

Similar to the proof of the inequality (25), we derive

$$\begin{aligned}
\mathbb{E}(\|\mathbf{U}^{\mathrm{c}}_4\|_\infty) \lesssim\, & n(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}^2_{\boldsymbol{\eta}}\boldsymbol{\beta})^{1/2}\varrho\\
&+ \{n^{1/2}(\log p_{\boldsymbol{\gamma}})^{1/2} + \log p_{\boldsymbol{\gamma}}\log(np_{\boldsymbol{\gamma}})\}\\
&\quad\times\left\{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\varrho^2 + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\right\}^{1/2}\\
&+ \{\log p_{\boldsymbol{\gamma}} + n^{-1/2+1/q}(\log p_{\boldsymbol{\gamma}})^{3/2}\log(np_{\boldsymbol{\gamma}})\}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta})^{1/2}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\mathbf{X}})
\end{aligned}$$

$$\begin{aligned}
&\lesssim n(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2 \boldsymbol{\beta})^{1/2} \varrho + n^{1/2} (\log p_{\boldsymbol{\gamma}})^{1/2} \big\{ \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) \varrho^2 \\
&\qquad + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} \big\}^{1/2} \qquad (41)\\
&\quad + \log p_{\boldsymbol{\gamma}} (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} \operatorname{tr}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}^2),
\end{aligned}$$

where the first inequality holds by Lemma C.2 and inequalities (37)–(40). The last inequality is due to the fact $\log p_{\boldsymbol{\gamma}} = O(n^b)$, where $0 < b < 1/3$. Combining equations (35), (41) and Assumption 3, we have

$$\begin{aligned}
A_4^{\mathrm{c}} = O_p \Big[& n^{1/2} s (\log p_{\boldsymbol{\gamma}} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2 \boldsymbol{\beta})^{1/2} \varrho + s \log p_{\boldsymbol{\gamma}} \big\{ \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) \varrho^2 \\
& + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} \big\}^{1/2} \\
& + n^{-1/2} s (\log p_{\boldsymbol{\gamma}})^{3/2} (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} \operatorname{tr}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{X}}^2) \Big].
\end{aligned}$$

Thus $A_4^{\mathrm{c}} = o_p(\sqrt{2\Lambda_{\mathbf{X}}})$ by the conditions in Theorem 2.1 and the model structure under the local alternatives. In summary, this theorem is proved.

B.3. The proof of Theorem 4.1

As $\boldsymbol{\gamma}_\phi = \boldsymbol{\gamma}$ under $\mathbb{H}_0$, we can decompose M_n^o as

$$\begin{aligned}
M_n^o &= \frac{1}{n} \sum_{i \neq j} \hat{\epsilon}_i \hat{\epsilon}_j \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
&= \frac{1}{n} \sum_{i \neq j} (\epsilon_i + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i)(\epsilon_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
&= \frac{1}{n} \sum_{i \neq j} (\epsilon_i \epsilon_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j + \epsilon_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \epsilon_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
&= \frac{1}{n} \sum_{i \neq j} \epsilon_i \epsilon_j \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j + \frac{1}{n} \sum_{i \neq j} \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j + \frac{2}{n} \sum_{i \neq j} \epsilon_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
&=: I_1^o + I_2^o + I_3^o.
\end{aligned}$$

Similar to the arguments of proving Theorem 3 in [20], we can see

$$\frac{I_1^o}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \text{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}) \to \infty$. See Lemma C.1 below for more details. Thus to prove this theorem, it suffices to I_2^o and I_3^o are $o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$.

Following Lemma C.10 below, we derive that

$$|I_2^o| \leq \left\| \frac{1}{n} \sum_{i \neq j} \mathbf{Z}_i \mathbf{Z}_j^\top \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \right\|_\infty \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_1^2,$$

where $\|n^{-1}\sum_{i\neq j}\mathbf{Z}_i\mathbf{Z}_j^\top\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_\infty = \max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}|n^{-1}\sum_{i\neq j}Z_{i,k_1}Z_{j,k_2}\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j|$. Denote $q=\lceil 4/(1-3b)\rceil$. Similar to the proof of inequality (22) in the main text, it can be shown that

$$\begin{aligned}
&\|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} Z_{i,k_1}Z_{j,k_2}\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_4\\
&\quad\le \|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} Z_{i,k_1}Z_{j,k_2}\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_q\\
&\quad\le \|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} Z_{i,k_1}Z_{j,k_2}\|_{2q}\|\max_{1\le i\neq j\le n}\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_{2q}\\
&\quad\lesssim \log(np_{\boldsymbol{\gamma}})\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|Z_{1,k_1}Z_{2,k_2}\|_{\psi_1}n^{1/q}\|\boldsymbol{\eta}_1^\top\boldsymbol{\eta}_2\|_{2q}\\
&\quad\lesssim \log(np_{\boldsymbol{\gamma}})n^{1/q}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2).
\end{aligned}$$

Similar to proving the inequality (24) in the main text and Theorem 5.1 in [11], we derive

$$\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|Z_{1,k_1}Z_{2,k_2}\boldsymbol{\eta}_1^\top\boldsymbol{\eta}_2\|_2\le\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|Z_{1,k_1}Z_{2,k_2}\boldsymbol{\eta}_1^\top\boldsymbol{\eta}_2\|_4\lesssim\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2),$$

and

$$\left\|\frac{1}{n}\sum_{i\neq j}\mathbf{Z}_i\mathbf{Z}_j^\top\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\right\|_\infty = O_p\Big[\big\{\log p_{\boldsymbol{\gamma}}+n^{1/q-1/2}(\log p)^{3/2}\log(np_{\boldsymbol{\gamma}})\big\}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2)\Big].$$

It follows that $\|n^{-1}\sum_{i\neq j}\mathbf{Z}_i\mathbf{Z}_j^\top\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_\infty = O_p\big(\log p_{\boldsymbol{\gamma}}\sqrt{\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}\big)$ as $\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2)\lesssim\sqrt{\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}$ and $\log p_{\boldsymbol{\gamma}}=O(n^b)$ for $0<b<1/3$. Thus $|I_2^o|=o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$ is proved when $s\log p_{\boldsymbol{\gamma}}/\sqrt{n}=o(1)$.

Similarly, the argument of proving I_3^c yields

$$|I_3^o|\le\left\|\frac{1}{n}\sum_{i\neq j}(\epsilon_i\mathbf{Z}_j+\mathbf{Z}_i\epsilon_j)\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\right\|_\infty\|\breve{\gamma}\|_1.$$

Note that

$$\begin{aligned}
&\|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}}(\epsilon_iZ_{j,k}+Z_{i,k}\epsilon_j)\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_4\\
&\quad\le\|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}}(\epsilon_iZ_{j,k}+Z_{i,k}\epsilon_j)\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_q\\
&\quad\le\|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}}\epsilon_iZ_{j,k}\|_{2q}\|\max_{1\le i\neq j\le n}\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_{2q}\\
&\quad\lesssim\log(np_{\boldsymbol{\gamma}})n^{1/q}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2),
\end{aligned}$$

and

$$\max_{1\le k\le p_{\boldsymbol{\gamma}}}\|(\epsilon_iZ_{j,k}+Z_{i,k}\epsilon_j)\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_2\le\max_{1\le k\le p_{\boldsymbol{\gamma}}}\|(\epsilon_iZ_{j,k}+Z_{i,k}\epsilon_j)\boldsymbol{\eta}_i^\top\boldsymbol{\eta}_j\|_4\lesssim\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2).$$

Theorem 5.1 in [11] again shows

$$\left\| \frac{1}{n} \sum_{i \neq j} (\epsilon_i \mathbf{Z}_j + \mathbf{Z}_i \epsilon_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \right\|_\infty$$
$$= O_p \left[\left\{ \log p_{\boldsymbol{\gamma}} + n^{1/q-1/2} (\log p_{\boldsymbol{\gamma}})^{3/2} \log(n p_{\boldsymbol{\gamma}}) \right\} \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2) \right].$$

It follows that $\|n^{-1} \sum_{i \neq j} (\epsilon_i \mathbf{Z}_j + \mathbf{Z}_i \epsilon_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j\|_\infty = O_p(\log p_{\boldsymbol{\gamma}} \sqrt{\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$. Thus $|I_3^o| = o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$ is proved when Assumption 3 and $s(\log p_{\boldsymbol{\gamma}})^{3/2}/\sqrt{n} = o(1)$ hold. In summary, the proof is finished.

B.4. The proof of Theorem 4.2

The proof of this theorem is similar to the proof of Theorem 2.2. Under the local alternatives, we decompose M_n^{o} as:

$$\begin{aligned}
M_n^{\mathrm{o}} =& \frac{1}{n} \sum_{i \neq j} \hat{\epsilon}_i \hat{\epsilon}_j \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
=& \frac{1}{n} \sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i + \epsilon_i + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i)(\boldsymbol{\beta}^\top \boldsymbol{\eta}_j + \epsilon_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
=& \frac{1}{n} \sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i + \epsilon_i)(\boldsymbol{\beta}^\top \boldsymbol{\eta}_j + \epsilon_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
&+ \frac{1}{n} \sum_{i \neq j} \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
&+ \frac{1}{n} \sum_{i \neq j} (\epsilon_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \epsilon_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
&+ \frac{1}{n} \sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
=:& A_1^{\mathrm{o}} + A_2^{\mathrm{o}} + A_3^{\mathrm{o}} + A_4^{\mathrm{o}}.
\end{aligned}$$

Rewrite the first term A_1^{o} as

$$\begin{aligned}
A_1^{\mathrm{o}} =& \frac{1}{n} \sum_{i \neq j} \epsilon_i \epsilon_j \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j + \frac{1}{n} \sum_{i \neq j} \boldsymbol{\beta}^\top \boldsymbol{\eta}_i \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \boldsymbol{\eta}_j^\top \boldsymbol{\beta} \\
&+ \frac{1}{n} \sum_{i \neq j} (\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \epsilon_j + \epsilon_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j) \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \\
=:& A_{11}^{\mathrm{o}} + A_{12}^{\mathrm{o}} + A_{13}^{\mathrm{o}}.
\end{aligned}$$

Following Lemma C.1 below, it can be seen

$$\frac{A_{11}^{\mathrm{o}}}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \text{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}) \to \infty$. Note that $\mathbb{E}(A_{12}^{\circ}) = (n-1)\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{2}\boldsymbol{\beta}$. The Hoeffding decomposition yields

$$\frac{1}{n-1}A_{12}^{\circ} = \boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{2}\boldsymbol{\beta} + 2A_{121}^{\circ} + A_{122}^{\circ},$$

where

$$A_{121}^{\circ} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{\top}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}),$$

and

$$A_{122}^{\circ} = \frac{1}{n(n-1)}\sum_{i\neq j}(\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{\top}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta})^{\top}(\boldsymbol{\eta}_j\boldsymbol{\eta}_j^{\top}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}).$$

The variance of A_{12}° can be bounded by

$$\begin{aligned}\mathrm{Var}(A_{12}^{\circ}) &\lesssim n^2\mathbb{E}(A_{121}^{\circ})^2 + n^2\mathbb{E}(A_{122}^{\circ})^2 \\ &\lesssim n\mathbb{E}(\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^{\top}\boldsymbol{\beta})^2 + \mathbb{E}(\boldsymbol{\beta}^{\top}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^{\top}\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\boldsymbol{\beta})^2 \\ &= o(\boldsymbol{\Lambda}_{\boldsymbol{\eta}}).\end{aligned}$$

The last equality holds by equations (61) and (62) in Lemma C.7, the fact that $(\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{2}\boldsymbol{\beta})^2 \le \boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{3}\boldsymbol{\beta}$. Similarly, we derive $\mathbb{E}(A_{13}^{\circ}) = 0$ and

$$\frac{1}{n-1}A_{13}^{\circ} = 2A_{131}^{\circ} + A_{132}^{\circ},$$

where

$$A_{131}^{\circ} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\epsilon_i\boldsymbol{\eta}_i$$

and

$$A_{132}^{\circ} = \frac{1}{n(n-1)}\sum_{i\neq j}\big\{(\boldsymbol{\beta}^{\top}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{\top} - \boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}})\epsilon_j\boldsymbol{\eta}_j + \epsilon_i\boldsymbol{\eta}_i^{\top}(\boldsymbol{\eta}_j\boldsymbol{\eta}_j^{\top}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta})\big\}.$$

Similar to the derivation of $\mathrm{Var}(A_{12}^{\circ})$, we derive

$$\begin{aligned}\mathrm{Var}(A_{13}^{\circ}) &\lesssim n^2\mathbb{E}(A_{131}^{\circ})^2 + n^2\mathbb{E}(A_{132}^{\circ})^2 \\ &\lesssim n\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\mathbb{E}(\epsilon_1^2\boldsymbol{\eta}_1\boldsymbol{\eta}_1^{\top})\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta} + \mathbb{E}(\epsilon_2\boldsymbol{\beta}^{\top}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^{\top}\boldsymbol{\eta}_2)^2 \\ &\lesssim n\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{3}\boldsymbol{\beta} + \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{2})\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta} \\ &= o(\boldsymbol{\Lambda}_{\boldsymbol{\eta}}),\end{aligned}$$

where the equality holds by the model structure under the local alternatives, and the third inequality holds by (63) in Lemma C.7, Assumption 5 and the fact that ϵ is independent of $\boldsymbol{\nu}$.

Altogether, we have

$$\frac{A_1^{\circ} - n\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{2}\boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \text{in distribution.}$$

To conclude the proof, we now prove $A_i^{\mathrm{o}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$, $i = 2, 3, 4$. By the same arguments for handling I_2^{o} and I_3^{o} in Theorem 4.1, we can show

$$A_2^{\mathrm{o}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}) \quad \text{and} \quad A_3^{\mathrm{o}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}).$$

For $A_4^{\mathrm{o}} = o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$, we recall that

$$A_4^{\mathrm{o}} = \frac{1}{n}\sum_{i\neq j}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \check{\boldsymbol{\gamma}}^\top \mathbf{Z}_j + \check{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j)\boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j.$$

By the Hölder inequality, we derive

$$|A_4^{\mathrm{o}}| \leq \left\| \frac{1}{n}\sum_{i\neq j}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \mathbf{Z}_j + \mathbf{Z}_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j)\boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j \right\|_\infty \|\check{\boldsymbol{\gamma}}\|_1. \tag{42}$$

Let $\mathbf{U}_4^{\mathrm{o}}$ be a $p_{\boldsymbol{\gamma}}$ dimension vector with k-th element

$$U_{4,k}^{\mathrm{o}} = \frac{1}{n}\sum_{i\neq j}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_i Z_{j,k} + Z_{i,k} \boldsymbol{\beta}^\top \boldsymbol{\eta}_j)\boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j.$$

Note that $\mathbf{U}_4^{\mathrm{o}} = n^{-1}\sum_{i\neq j}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_i \mathbf{Z}_j + \mathbf{Z}_i \boldsymbol{\beta}^\top \boldsymbol{\eta}_j)\boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j$ and all of its elements are U-statistics. The Hoeffding decomposition yields

$$\frac{1}{n-1}U_{4,k}^{\mathrm{o}} = 2S_{4,1,k}^{\mathrm{o}} + S_{4,2,k}^{\mathrm{o}},$$

where

$$S_{4,1,k}^{\mathrm{o}} = \frac{1}{n}\sum_{i=1}^{n} g_{4,1,k}^{\mathrm{o}}(\boldsymbol{\nu}_i)$$

with $g_{4,1,k}^{\mathrm{o}}(\boldsymbol{\nu}_1) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} Z_{1,k} \boldsymbol{\eta}_1$,

$$S_{4,2,k}^{\mathrm{o}} = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} g_{4,2,k}^{\mathrm{o}}(\boldsymbol{\nu}_i; \boldsymbol{\nu}_j)$$

with $g_{4,2,k}^{\mathrm{o}}(\boldsymbol{\nu}_1; \boldsymbol{\nu}_2) = (\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta})^\top Z_{2,k}\boldsymbol{\eta}_2 + Z_{1,k}\boldsymbol{\eta}_1^\top(\boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta})$.

For any $1 \leq k \leq p_{\boldsymbol{\gamma}}$, according to Assumption 2, $Z_{1,k}$ is sub-Gaussian with bounded norm, and $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\eta}_1$ is sub-Gaussian with norm $\|\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\eta}_1\|_{\psi_2} \leq (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^3 \boldsymbol{\beta})^{1/2}$. Then $g_{4,1,k}^{\mathrm{o}}(\boldsymbol{\nu}_1)$ is sub-Exponential with norm

$$\|g_{4,1,k}^{\mathrm{o}}\|_{\psi_1} \leq \|Z_{1,k}\|_{\psi_2} \|\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\eta}_1\|_{\psi_2} \lesssim (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^3 \boldsymbol{\beta})^{1/2}.$$

Therefore

$$\begin{aligned}\|\max_{1\leq i\leq n}\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} g_{4,1,k}^{\mathrm{o}}\|_2 &\lesssim \log(np_{\boldsymbol{\gamma}}) \max_{1\leq i\leq n}\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} \|g_{4,1,k}^{\mathrm{o}}\|_{\psi_1} \\ &\lesssim \log(np_{\boldsymbol{\gamma}})(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^3 \boldsymbol{\beta})^{1/2}.\end{aligned}$$

Similarly, we derive $\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 Z_{2,k}$ is sub-Exponential with norm $\|\boldsymbol{\beta}^\top \mathbf{X}_1 Z_{2,k}\|_{\psi_1} \lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}$. Write $q = \lceil 4/(1-3b) \rceil$. We have

$$\begin{aligned}
\|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} g^{\mathrm{o}}_{4,2,k}\|_4 &\le \|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} g^{\mathrm{o}}_{4,2,k}\|_q \\
&\le \|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} \boldsymbol{\beta}^\top \boldsymbol{\eta}_i Z_{j,k} \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j\|_q \\
&\le \|\max_{1\le i\neq j\le n}\max_{1\le k\le p_{\boldsymbol{\gamma}}} \boldsymbol{\beta}^\top \boldsymbol{\eta}_i Z_{j,k}\|_{2q} \|\max_{1\le i\neq j\le n} \boldsymbol{\eta}_i^\top \boldsymbol{\eta}_j\|_{2q} \\
&\lesssim \log(np_{\boldsymbol{\gamma}})(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} n^{1/q} \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\boldsymbol{\eta}}).
\end{aligned}$$

Similarly,

$$\max_{1\le k\le p_{\boldsymbol{\gamma}}} \|g^{\mathrm{o}}_{4,1,k}\|_2 \lesssim (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^3_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2}$$

and

$$\max_{1\le k\le p_{\boldsymbol{\gamma}}} \|g^{\mathrm{o}}_{4,2,k}\|_2 \le \max_{1\le k\le p_{\boldsymbol{\gamma}}} \|g^{\mathrm{o}}_{4,2,k}\|_4 \lesssim (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\boldsymbol{\eta}}).$$

Lemma C.2 and Assumption 4 imply

$$\mathbb{E}(\|\mathbf{U}^{\mathrm{o}}_4\|_\infty) \lesssim n^{1/2}(\log p_{\boldsymbol{\gamma}})^{1/2}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^3_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} + \log p_{\boldsymbol{\gamma}} (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\boldsymbol{\eta}}). \quad (43)$$

Combining equations (42) and (43), Assumption 3, we have

$$A^{\mathrm{o}}_4 = O_p\left(s \log p_{\boldsymbol{\gamma}} (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^3_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} + n^{-1/2} s (\log p_{\boldsymbol{\gamma}})^{3/2} (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta})^{1/2} \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^2_{\boldsymbol{\eta}}) \right).$$

Thus $A^{\mathrm{o}}_4 = o_p(\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}})$ holds by the conditions in Theorem 4.1 and the model structure under the local alternatives. Altogether, the proof is done.

B.5. The proof of Theorem 5.1

Denote $\mathcal{A} = \{\|\breve{\boldsymbol{W}}\|_F^2 \lesssim s' \log p_{\boldsymbol{\gamma}}/n'\}$. According to Assumption 7, $\Pr(\mathcal{A}^c) = o(1)$. Thus it suffices to prove

$$\frac{M_n}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \text{in distribution}$$

as $(n, p_{\boldsymbol{\beta}}, p_{\boldsymbol{\gamma}}) \to \infty$ conditional on $\mathcal{A}$. For brevity, in the rest of proof we abbreviate $\mathbb{E}(\cdot|\mathcal{A})$, $\Pr(\cdot|\mathcal{A})$ as $\mathbb{E}(\cdot)$, $\Pr(\cdot)$ respectively. As $\boldsymbol{\gamma}_\phi = \boldsymbol{\gamma}$ under $\mathbb{H}_0$, we can decompose M_n as

$$\begin{aligned}
M_n &= \frac{1}{n}\sum_{i\neq j} \hat{\epsilon}_i \hat{\epsilon}_j \hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j \\
&= \frac{1}{n}\sum_{i\neq j} (\epsilon_i + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i)(\epsilon_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j) \hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j \\
&= \frac{1}{n}\sum_{i\neq j} \epsilon_i \epsilon_j \hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j + \frac{1}{n}\sum_{i\neq j} \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j \hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{n}\sum_{i\neq j}(\epsilon_i \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_j + \breve{\boldsymbol{\gamma}}^\top \mathbf{Z}_i \epsilon_j)\hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j \\
=:\ & I_1 + I_2 + I_3.
\end{aligned}
$$

Similar to the proof of Theorem 3 in [20], it can be seen

$$
\frac{I_1}{\sqrt{2\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}}} \to N(0,1) \quad \text{in distribution}
$$

as $(n, p_{\boldsymbol{\beta}}) \to \infty$, where $\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}} = \sigma^4 \mathrm{tr}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2)$. See Lemma C.1 for more details. As the inequality (64) in Lemma C.8 implies $\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}/\boldsymbol{\Lambda}_{\boldsymbol{\eta}} \to 1$ in probability, we can then use Sluskty theorem to derive

$$
\frac{I_1}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}} \to N(0,1) \quad \text{in distribution}
$$

as $(n, p_{\boldsymbol{\beta}}, p_{\boldsymbol{\gamma}}) \to \infty$. Further, similarly, we prove I_2 and I_3 are $o_p(\sqrt{2\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}})$ to finish the proof. As the proof is very similar, we give a sketch below.

Lemma C.10 below yields

$$
I_2 \le \left\| \frac{1}{n}\sum_{i\neq j} \mathbf{Z}_i \mathbf{Z}_j^\top \hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j \right\|_\infty \|\breve{\boldsymbol{\gamma}}\|_1^2. \tag{44}
$$

Let $\mathbf{U}_1 = n^{-1}\sum_{i\neq j} \mathbf{Z}_i \mathbf{Z}_j^\top \hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j$ and all of its elements are U-statistics: for (k_1, k_2)-th element,

$$
U_{1,(k_1,k_2)} = \frac{1}{n}\sum_{i\neq j}\frac{1}{2}(Z_{i,k_1}Z_{j,k_2} + Z_{i,k_2}Z_{j,k_1})\hat{\boldsymbol{\eta}}_i^\top \hat{\boldsymbol{\eta}}_j.
$$

The Hoeffding decomposition implies

$$
\frac{1}{n-1}U_{1,(k_1,k_2)} = \mathbb{E}(Z_{k_1}\hat{\boldsymbol{\eta}})^\top \mathbb{E}(Z_{k_2}\hat{\boldsymbol{\eta}}) + 2S_{1,1,(k_1,k_2)} + S_{1,2,(k_1,k_2)},
$$

where

$$
S_{1,1,(k_1,k_2)} = \frac{1}{n}\sum_{i=1}^n g_{1,1,(k_1,k_2)}(\boldsymbol{\nu}_i)
$$

with $g_{1,1,(k_1,k_2)}(\boldsymbol{\nu}_1) = \{Z_{1,k_1}\hat{\boldsymbol{\eta}}_1 - \mathbb{E}(Z_{k_1}\hat{\boldsymbol{\eta}})\}^\top \mathbb{E}(Z_{k_2}\hat{\boldsymbol{\eta}})/2 + \{Z_{1,k_2}\hat{\boldsymbol{\eta}}_1 - \mathbb{E}(Z_{k_2}\hat{\boldsymbol{\eta}})\}^\top \mathbb{E}(Z_{k_1}\hat{\boldsymbol{\eta}})/2$,

$$
\boldsymbol{S}_{1,2,(k_1,k_2)} = \frac{1}{n(n-1)}\sum_{i\neq j}^n g_{1,2,(k_1,k_2)}(\boldsymbol{\nu}_i;\boldsymbol{\nu}_j)
$$

with $g_{1,2,(k_1,k_2)}(\boldsymbol{\nu}_1;\boldsymbol{\nu}_2) = \{Z_{1,k_1}\hat{\boldsymbol{\eta}}_1 - \mathbb{E}(Z_{k_1}\hat{\boldsymbol{\eta}})\}^\top\{Z_{2,k_2}\hat{\boldsymbol{\eta}}_2 - \mathbb{E}(Z_{k_2}\hat{\boldsymbol{\eta}})\}/2 + \{Z_{1,k_2}\hat{\boldsymbol{\eta}}_1 - \mathbb{E}(Z_{k_2}\hat{\boldsymbol{\eta}})\}^\top\{Z_{2,k_1}\hat{\boldsymbol{\eta}}_2 - \mathbb{E}(Z_{k_1}\hat{\boldsymbol{\eta}})\}/2$.

Similar to the derivation of I_2^c in the proof of Theorem 2.1, we derive

$$
\|\max_{1\le i\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} g_{1,1,(k_1,k_2)}\|_2 \lesssim \log(np_{\boldsymbol{\gamma}})\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}})\{\max_{1\le k\le p_{\boldsymbol{\gamma}}}\|\mathbb{E}(Z_k\hat{\boldsymbol{\eta}})\|_2^2\}^{1/2}
$$

$$\begin{aligned}
&\|\max_{1\le i\neq j\le n}\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}} g_{1,2,(k_1,k_2)}\|_4 \lesssim \log(np_{\boldsymbol{\gamma}})n^{1/q}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2)\\
&\qquad\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|g_{1,1,(k_1,k_2)}\|_2 \lesssim \lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}})\big\{\max_{1\le k\le p_{\boldsymbol{\gamma}}}\|\mathbb{E}(Z_k\hat{\boldsymbol{\eta}})\|_2^2\big\}^{1/2}\\
&\qquad\max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|g_{1,2,(k_1,k_2)}\|_2 \le \max_{1\le k_1,k_2\le p_{\boldsymbol{\gamma}}}\|g_{1,2,(k_1,k_2)}\|_4 \lesssim \mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2).
\end{aligned}$$

Lemma C.2, C.8 and Assumption 4 yield

$$\begin{aligned}
\mathbb{E}(\|\mathbf{U}_1\|_\infty) &\lesssim n\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\varphi^2\\
&\quad + n^{1/2}(\log p_{\boldsymbol{\gamma}})^{1/2}\left\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}})\left(\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right)^{1/2}\vee\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right\}\varphi\\
&\quad + \log p_{\boldsymbol{\gamma}}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2). \qquad (45)
\end{aligned}$$

Together equations (44) with (45) and Assumption 3, we have

$$\begin{aligned}
I_2 = O_p\Bigg[&s^2\log p_{\boldsymbol{\gamma}}\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\varphi^2\\
&+n^{-1/2}s^2(\log p_{\boldsymbol{\gamma}})^{3/2}\left\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}})\left(\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right)^{1/2}\vee\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right\}\varphi\\
&+n^{-1}s^2(\log p_{\boldsymbol{\gamma}})^2\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2)\Bigg]. \qquad (46)
\end{aligned}$$

Thus $I_2 = o_p(\sqrt{\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}})$ when $s^2\log p_{\boldsymbol{\gamma}}\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\varphi^2 = o(\sqrt{\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}})$,

$$n^{-1/2}s^2(\log p_{\boldsymbol{\gamma}})^{3/2}\left\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}})\left(\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right)^{1/2}\vee\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right\}\varphi = o(\sqrt{\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}})$$

and $n^{-1}s^2(\log p_{\boldsymbol{\gamma}})^2 = o(1)$ hold simultaneously.

Similar to the proof of I_2,

$$I_3 \le \left\|\frac{1}{n}\sum_{i\neq j}(\epsilon_i\mathbf{Z}_j + \mathbf{Z}_i\epsilon_j)\hat{\boldsymbol{\eta}}_i^\top\hat{\boldsymbol{\eta}}_j\right\|_\infty\|\check{\gamma}\|_1 =: \|\mathbf{U}_2\|_\infty\|\check{\gamma}\|_1. \qquad (47)$$

Applying the Hoeffding decomposition to every element of $\mathbf{U}_2$, we have

$$\frac{1}{n-1}U_{2,k} = 2S_{2,1,k} + S_{2,2,k},$$

where

$$S_{2,1,k} = \frac{1}{n}\sum_{i=1}^n g_{2,1,k}(\boldsymbol{\nu}_i)$$

with $g_{2,1,k}(\boldsymbol{\nu}_1) = \epsilon_1\hat{\boldsymbol{\eta}}_1^\top\mathbb{E}(Z_k\hat{\boldsymbol{\eta}})$,

$$S_{2,2,k} = \frac{1}{n(n-1)}\sum_{i\neq j}^n g_{2,2,k}(\boldsymbol{\nu}_i,\boldsymbol{\nu}_j)$$

with $g_{2,2,k}(\boldsymbol{\nu}_1;\boldsymbol{\nu}_2)=\epsilon_1\hat{\boldsymbol{\eta}}_1^\top\{Z_{2,k}\hat{\boldsymbol{\eta}}_2-\mathbb{E}(Z_k\hat{\boldsymbol{\eta}})\}+\epsilon_2\hat{\boldsymbol{\eta}}_2^\top\{Z_{1,k}\hat{\boldsymbol{\eta}}_1-\mathbb{E}(Z_k\hat{\boldsymbol{\eta}})\}$.

Similar to the proof for I_3^c in the proof of Theorem 2.1, and for the inequality (45), we can then have

$$\mathbb{E}(\|\mathbf{U}_2\|_\infty)\lesssim n^{1/2}(\log p_{\boldsymbol{\gamma}})^{1/2}\left\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}})\left(\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right)^{1/2}\vee\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right\}\varphi \\ +\log p_{\boldsymbol{\gamma}}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2). \tag{48}$$

Combining equations (47) and (48), Assumption 3, we have

$$I_3=O_p\left[s\log p_{\boldsymbol{\gamma}}\left\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}})\left(\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right)^{1/2}\vee\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right\}\varphi \\ +n^{-1/2}s(\log p_{\boldsymbol{\gamma}})^{3/2}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2)\right]. \tag{49}$$

Thus $I_3=o_p(\sqrt{\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}})$ when

$$s\log p_{\boldsymbol{\gamma}}\left\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}})\left(\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right)^{1/2}\vee\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_{\mathbf{Z}})\frac{s'\log p_{\boldsymbol{\gamma}}}{n'}\right\}\varphi=o(\sqrt{\boldsymbol{\Lambda}_{\hat{\boldsymbol{\eta}}}})$$

and $n^{-1/2}s(\log p_{\boldsymbol{\gamma}})^{3/2}=o(1)$ hold simultaneously. The proof is concluded.

B.6. The proof of Theorem 5.2

As we can follow almost the same lines of proving Theorem 5.1 to have, for $\boldsymbol{\beta}\in\mathscr{L}(\boldsymbol{\beta})$,

$$\frac{M_n-n\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta}}{\sqrt{2\boldsymbol{\Lambda}_{\boldsymbol{\eta}}}}\to N(0,1)\quad\text{in distribution}$$

as $(n,p_{\boldsymbol{\beta}},p_{\boldsymbol{\gamma}})\to\infty$ conditional on $\mathcal{A}$, we then omit the details here.

Appendix C: Some useful lemmas

Lemma C.1. *Let $\boldsymbol{\xi}=\mathbf{A}\boldsymbol{\nu}$, where $\mathbf{A}$ is a $p_{\boldsymbol{\xi}}\times m$ matrix with $p_{\boldsymbol{\xi}}\leq m$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}=\mathbf{A}\mathbf{A}^\top$. Let $\boldsymbol{\nu}$ be a m-dimensional sub-Gaussian random vector with mean zero and identity covariance matrix. ϵ is a zero mean sub-Gaussian random variable with variance σ^2 and ϵ is independent of $\boldsymbol{\xi}$. Assume that $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^4)=o(\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2))$ and $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)\to\infty$ as $(n,p_{\boldsymbol{\xi}})\to\infty$. Then*

$$\frac{\sum_{i\neq j}\epsilon_i\epsilon_j\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_j}{n\sqrt{2\Lambda_{\boldsymbol{\xi}}}}\to N(0,1)\quad\textit{in distribution}$$

as $(n,p_{\boldsymbol{\xi}})\to\infty$, where $\Lambda_{\boldsymbol{\xi}}=\sigma^4\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)$.

Proof. Denote

$$R_{ni} = 2n^{-1}\sum_{j=1}^{i-1}\epsilon_i\epsilon_j\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_j,$$

$S_{nk} = \sum_{i=2}^{k} R_{ni}$ and $\mathscr{F}_k = \sigma\{(\boldsymbol{\xi}_i, \epsilon_i),\ i = 1, 2, \ldots, k\}$. Obviously, $\mathbb{E}(R_{nk}|\mathscr{F}_{k-1}) = 0$ as $(S_{nk}, \mathscr{F}_k)$ is a zero-mean martingale sequence. Denote v_{ni}=Var$(R_{ni}|\mathscr{F}_{i-1})$ and $V_n = \sum_{i=2}^{n} v_{ni}$. To apply the martingale central limit theorem, it suffices to show the following two conditions:

$$\frac{V_n}{\mathrm{Var}(S_{nn})} \to 1 \quad \text{in probability} \tag{50}$$

and

$$\frac{\sum_{i=2}^{n}\mathbb{E}\big[R_{ni}^2 I\{|R_{ni}|/\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} > \eta\}|\mathscr{F}_{i-1}\big]}{\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} \to 0 \quad \text{in probability} \tag{51}$$

for any $\eta > 0$. To prove equation (50) first. Observe that

$$v_{ni} = 4\sigma^2 n^{-2}\sum_{j=1}^{i-1}\epsilon_j^2\boldsymbol{\xi}_j^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_j + 4\sigma^2 n^{-2}\sum_{1\le j_1\neq j_2\le i-1}\epsilon_{j_1}\epsilon_{j_2}\boldsymbol{\xi}_{j_1}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_{j_2},$$

and

$$V_n = 4\sigma^2 n^{-2}\sum_{i=2}^{n}\Bigg(\sum_{j=1}^{i-1}\epsilon_j^2\boldsymbol{\xi}_j^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_j + \sum_{1\le j_1\neq j_2\le i-1}\epsilon_{j_1}\epsilon_{j_2}\boldsymbol{\xi}_{j_1}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_{j_2}\Bigg).$$

Since $\mathrm{Var}(S_{nn}) = 2n^{-1}(n-1)\sigma^4\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)$, we have

$$\begin{aligned}\frac{V_n}{\mathrm{Var}(S_{nn})} &= \frac{2}{n(n-1)\sigma^2\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)}\\ &\quad\times\Bigg(\sum_{i=2}^{n}\sum_{j=1}^{i-1}\epsilon_j^2\boldsymbol{\xi}_j^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_j + \sum_{i=2}^{n}\sum_{1\le j_1\neq j_2\le i-1}\epsilon_{j_1}\epsilon_{j_2}\boldsymbol{\xi}_{j_1}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_{j_2}\Bigg)\\ &=: G_1 + G_2,\end{aligned}$$

where $\mathbb{E}(G_1) = 1$ and $\mathbb{E}(G_2) = 0$. Observe that

$$\begin{aligned}\mathrm{Var}(G_1) &\lesssim \frac{1}{n^4\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)}\sum_{j=1}^{n-1}(n-j)^2\{\mathbb{E}(\boldsymbol{\xi}_j^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_j)^2 - \mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)\}\\ &\lesssim \frac{1}{n^4}\sum_{j=1}^{n-1}(n-j)^2 \lesssim n^{-1},\end{aligned}$$

where the second inequality holds by using Lemma C.3 to derive

$$\mathbb{E}(\boldsymbol{\xi}_1^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_1)^2 \lesssim \mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2).$$

Similarly, we derive that

$$
\begin{aligned}
\mathrm{Var}(G_2) &\lesssim \frac{1}{n^4\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} \sum_{j_1<k_1}\sum_{j_2<k_2} (n-k_1)(n-k_2)\mathbb{E}\big(\epsilon_{j_1}\epsilon_{j_2}\epsilon_{k_1}\epsilon_{k_2}\boldsymbol{\xi}_{j_1}^\top \boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_{k_1}\boldsymbol{\xi}_{j_2}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_{k_2}\big) \\
&= \frac{\sum_{k=1}^n (n-k)^2(k-1)}{n^4} \cdot \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^4)}{\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} = o(1)
\end{aligned}
$$

where the last equality holds by condition $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^4) = o(\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2))$. The Markov inequality yields $G_1 \to 1$ in probability and $G_2 \to 0$ in probability. Thus equation (50) is proved.

To handle equation (51). Notice that for any $\eta > 0$,

$$
\frac{\sum_{i=2}^n \mathbb{E}\big[R_{ni}^2 I\{|R_{ni}|/\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} > \eta\}|\mathscr{F}_{i-1}\big]}{\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} \le \frac{1}{\eta^2}\cdot \frac{\sum_{i=2}^n \mathbb{E}(R_{ni}^4|\mathscr{F}_{i-1})}{\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)}, \qquad (52)
$$

and

$$
\begin{aligned}
&\mathbb{E}\bigg\{\frac{\sum_{i=2}^n \mathbb{E}(R_{ni}^4|\mathscr{F}_{i-1})}{\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)}\bigg\} \\
&\quad = \frac{\sum_{i=2}^n \mathbb{E}(R_{ni}^4)}{\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} \\
&\quad \lesssim \frac{1}{n^4\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} \sum_{i=2}^n \mathbb{E}\bigg(\sum_{j=1}^{i-1}\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_j\bigg)^4 \\
&\quad \lesssim \frac{1}{n^4\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)} \bigg[\sum_{i=2}^n\sum_{s\neq t}\mathbb{E}\big\{(\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_s)^2(\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_t)^2\big\} + \sum_{i=2}^n\sum_{j=1}^{i-1}\mathbb{E}(\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_j)^4\bigg] \\
&\quad = o(1), \qquad (53)
\end{aligned}
$$

where the last equality holds by the fact that

$$
\sum_{i=2}^n\sum_{s\neq t}\mathbb{E}\big\{(\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_s)^2(\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_t)^2\big\} \lesssim n^3\mathbb{E}\big\{(\boldsymbol{\xi}_3^\top\boldsymbol{\xi}_1)^2(\boldsymbol{\xi}_3^\top\boldsymbol{\xi}_2)^2\big\} \lesssim n^3\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2),
$$

Lemma C.3 ensures

$$
\begin{aligned}
\mathbb{E}\big\{(\boldsymbol{\xi}_3^\top\boldsymbol{\xi}_1)^2(\boldsymbol{\xi}_3^\top\boldsymbol{\xi}_2)^2\big\} &= \mathbb{E}\big[\mathbb{E}\big\{(\boldsymbol{\xi}_3^\top\boldsymbol{\xi}_1)^2(\boldsymbol{\xi}_3^\top\boldsymbol{\xi}_2)^2|\boldsymbol{\xi}_3\big\}\big] \\
&\lesssim \mathbb{E}(\boldsymbol{\xi}_3^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_3)^2 \\
&\lesssim \mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2),
\end{aligned}
$$

and Lemma C.4 yields

$$
\sum_{i=2}^n\sum_{j=1}^{i-1}\mathbb{E}(\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_j)^4 \lesssim n^2\mathbb{E}(\boldsymbol{\xi}_1^\top\boldsymbol{\xi}_2)^4 \lesssim n^2\mathrm{tr}^2(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2).
$$

Combining inequalities (52) and (53), equation (51) is verified. The proof is concluded. □

Lemma C.2. *(A maximal inequality for U-statistics of order two). Let $\{X_i\}_{i=1}^n$ be a sample of i.i.d. random variables in a separable and measurable space $(S, \mathcal{S})$. Let $\boldsymbol{f} : S \times S \to \mathbb{R}^p$ be an $\mathcal{S} \bigotimes \mathcal{S}$-measurable, symmetric kernel such that $\mathbb{E}\{|f_k(X_1, X_2)|\} < \infty$ for all $k = 1, \dots, p$. Let $\mathbf{U}_n = \{n(n-1)\}^{-1} \sum_{1 \le i \ne j \le n} \boldsymbol{f}(X_i, X_j)$ and $\boldsymbol{\theta} = \mathbb{E}\{\boldsymbol{f}(X_1, X_2)\}$. Suppose $2 \le p \le \exp(bn)$ for some constant $b > 0$, then*

$$\mathbb{E}(\|\mathbf{U}_n\|_\infty) \lesssim \|\boldsymbol{\theta}\|_\infty + \left(\frac{\log p}{n}\right)^{1/2} D_2' + \frac{\log p}{n} M_2' + \frac{\log p}{n} D_2'' + \left(\frac{\log p}{n}\right)^{5/4} D_4'' + \left(\frac{\log p}{n}\right)^{3/2} M_4'',$$

where $D_2' = \max_{1 \le k \le p}[E\{f_{1,k}^2(X_1)\}]^{1/2}$, $M_2' = (\mathbb{E}[\{\max_{1 \le i \le n} \max_{1 \le k \le p} |f_{1,k}(X_i)|\}^2])^{1/2}$, $D_q'' = \max_{1 \le k \le p}[E\{f_{2,k}^q(X_1, X_2)\}]^{1/q}$ for $q = 2, 4$, $M_4'' = (\mathbb{E}[\{\max_{1 \le i \ne j \le n} \max_{1 \le k \le p} |f_{2,k}(X_i, X_j)|\}^4])^{1/4}$ with kernel function

$$\boldsymbol{f}_1(x_1) = \mathbb{E}\{\boldsymbol{f}(X_1, X_2) | X_1 = x_1\} - \boldsymbol{\theta}$$

and

$$\boldsymbol{f}_2(x_1, x_2) = \boldsymbol{f}(x_1, x_2) - \boldsymbol{f}_1(x_1) - \boldsymbol{f}_1(x_2) - \boldsymbol{\theta}.$$

Proof. By the Hoeffding decomposition of U-statistic, we derive

$$\mathbf{U}_n = \boldsymbol{\theta} + 2\boldsymbol{S}_{1n} + \boldsymbol{S}_{2n},$$

where

$$\boldsymbol{S}_{1n} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{f}_1(X_i)$$

and

$$\boldsymbol{S}_{2n} = \frac{1}{n(n-1)} \sum_{i \ne j} \boldsymbol{f}_2(X_i, X_j).$$

By the triangle inequality of norm, we derive

$$\mathbb{E}(\|\mathbf{U}_n\|_\infty) \lesssim \|\boldsymbol{\theta}\|_\infty + \mathbb{E}(\|\boldsymbol{S}_{1n}\|_\infty) + \mathbb{E}(\|\boldsymbol{S}_{2n}\|_\infty). \tag{54}$$

Note that $\boldsymbol{S}_{1n}$ is the sample mean of an i.i.d. random vector sequence. Thus

$$\mathbb{E}(\|\boldsymbol{S}_{1n}\|_\infty) \lesssim \left(\frac{\log p}{n}\right)^{1/2} D_2' + \frac{\log p}{n} M_2' \tag{55}$$

by Lemma 8 in [13]. Observe that $\boldsymbol{S}_{2n}$ is a degenerate U-statistic, we derive

$$\mathbb{E}(\|\boldsymbol{S}_{2n}\|_\infty) \lesssim \frac{\log p}{n} D_2'' + \left(\frac{\log p}{n}\right)^{5/4} D_4'' + \left(\frac{\log p}{n}\right)^{3/2} M_4'' \tag{56}$$

by Theorem 5 in [11]. Altogether, the proof is finished. □

Lemma C.3. *Let $\{\mathbf{A}_i\}_{i=1}^k$ be a sequence of $p \times p$ semi-positive matrices and k is a fixed positive integer. Suppose $\boldsymbol{\nu}$ is a p-dimensional sub-Gaussian random vector with sub-Gaussian norm $\|\boldsymbol{\nu}\|_{\psi_2}$. Then*

$$\mathbb{E}\bigg(\prod_{i=1}^k \boldsymbol{\nu}^\top \mathbf{A}_i \boldsymbol{\nu}\bigg) \lesssim \|\boldsymbol{\nu}\|_{\psi_2}^{2k} \prod_{i=1}^k \mathrm{tr}(\mathbf{A}_i).$$

Proof. Matrix eigen-decomposition yields, for $i = 1, \ldots, k$,

$$\mathbf{A}_i = \mathbf{C}_i^\top \mathbf{D}_i \mathbf{C}_i,$$

where $\mathbf{D}_i = \mathrm{diag}\{\lambda_1(\mathbf{A}_i), \lambda_2(\mathbf{A}_i), \cdots, \lambda_p(\mathbf{A}_i)\}$, $\lambda_j(\mathbf{A}_i)$ is the j-th eigenvalue of $\mathbf{A}_i$ and $\mathbf{C}_i$ is a unit orthogonal matrix. Denote $\boldsymbol{S}_i = \mathbf{C}_i \boldsymbol{\nu}_1$. Then

$$\begin{aligned}
&\mathbb{E}\bigg(\prod_{i=1}^k \boldsymbol{\nu}^\top \mathbf{A}_i \boldsymbol{\nu}\bigg) \\
&\quad = \mathbb{E}\bigg(\prod_{i=1}^k \sum_{j_i=1}^p \lambda_{j_i}(\mathbf{A}_i) S_{i,j_i}^2\bigg) \\
&\quad = \sum_{j_1=1}^p \sum_{j_2=1}^p \cdots \sum_{j_k=1}^p \lambda_{j_1}(\mathbf{A}_1)\lambda_{j_2}(\mathbf{A}_2)\cdots\lambda_{j_k}(\mathbf{A}_k)\mathbb{E}\big(S_{1,j_1}S_{2,j_2}\cdots S_{k,j_k}\big)^2 \\
&\quad \le \max_{1\le j_1,j_2,\ldots,j_k\le p} \mathbb{E}\big(S_{1,j_1}S_{2,j_2}\cdots S_{k,j_k}\big)^2 \prod_{i=1}^k \mathrm{tr}(\mathbf{A}_i),
\end{aligned}$$

where S_{i,j_i} is the j_i-th element of $\boldsymbol{S}_i$. We then verify the following to conclude the proof:

$$\max_{1\le j_1,j_2,\ldots,j_k\le p} \mathbb{E}\big(S_{1,j_1}S_{2,j_2}\cdots S_{k,j_k}\big)^2 \lesssim \|\boldsymbol{\nu}\|_{\psi_2}^{2k}. \tag{57}$$

Observe that

$$\begin{aligned}
&\max_{1\le j_1,j_2,\ldots,j_k\le p} \mathbb{E}\big(S_{1,j_1}S_{2,j_2}\cdots S_{k,j_k}\big)^2 \\
\le& \max_{1\le j_1,j_2,\ldots,j_k\le p} \Big\{\mathbb{E}\big(S_{1,j_1}\big)^{2^k}\Big\}^{1/2^{k-1}} \Big\{\mathbb{E}\big(S_{2,j_2}\big)^{2^k}\Big\}^{1/2^{k-1}} \cdots \Big\{\mathbb{E}\big(S_{k,j_k}\big)^{2^k}\Big\}^{1/2^{k-1}} \\
\le& \Big\{\max_{1\le i\le k}\max_{1\le j\le p} \mathbb{E}\big(S_{i,j}\big)^{2^k}\Big\}^{k/2^{k-1}},
\end{aligned}$$

where the first inequality holds by Cauchy-Schwartz inequality and Liapounov inequality. By the sub-Gaussian assumption of $\boldsymbol{\nu}$, $\max_{1\le i\le k}\max_{1\le j\le p}\mathbb{E}\big(S_{i,j}\big)^{2^k} \lesssim \|\boldsymbol{\nu}\|_{\psi_2}^{2^k}$. Thus the inequality (57) is proved. □

Lemma C.4. *(A corollary of Lemma C.3) Let $\boldsymbol{\xi} = \mathbf{A}\boldsymbol{\nu}$, where $\mathbf{A}$ is a $p_{\boldsymbol{\xi}} \times m$ matrix with $p_{\boldsymbol{\xi}} \le m$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \mathbf{A}\mathbf{A}^\top$. Under Assumption 2, for a fixed positive integer q,*

$$\mathbb{E}(\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_2)^{2q} \lesssim \mathrm{tr}^q(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2).$$

Proof. Denote $\mathbb{E}_{\boldsymbol{\xi}_1}(\cdot) = \mathbb{E}(\cdot|\boldsymbol{\xi}_1)$, it can be shown that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_2)^{2q} &= \mathbb{E}(\boldsymbol{\xi}_2^\top \boldsymbol{\xi}_1 \boldsymbol{\xi}_1^\top \boldsymbol{\xi}_2)^q = \mathbb{E}\{\mathbb{E}_{\boldsymbol{\xi}_1}(\boldsymbol{\nu}_2^\top \mathbf{A}^\top \boldsymbol{\xi}_1 \boldsymbol{\xi}_1^\top \mathbf{A}\boldsymbol{\nu}_2)^q\} \\ &\lesssim \mathbb{E}(\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma}_{\boldsymbol{\xi}} \boldsymbol{\xi}_1)^q = \mathbb{E}(\boldsymbol{\nu}_1 \mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{\xi}} \mathbf{A}\boldsymbol{\nu}_1)^q \lesssim \operatorname{tr}^q(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^2)\end{aligned}$$

where the first inequality holds by Lemma C.3, $\boldsymbol{\xi}_1$ is independent of $\boldsymbol{\xi}_2$, and the second inequality holds also by Lemma C.3. □

Definition C.5. (Some notations used in Lemmas C.6, C.7 and C.8.) Recall that $\boldsymbol{W}$ is defined in (7) in subsection 2.2 and $\hat{\boldsymbol{W}}$ is defined in (15) in section 5. We give some notations here to avoid repeating the notations in the following proofs. Let $\boldsymbol{\eta}_i = \mathbf{X}_i - \boldsymbol{W}^\top \mathbf{Z}_i$, $\hat{\boldsymbol{\eta}}_i = \mathbf{X}_i - \hat{\boldsymbol{W}}^\top \mathbf{Z}_i$, $\boldsymbol{\Gamma}_{\boldsymbol{\eta}} = \boldsymbol{\Gamma}_{\mathbf{X}} - \boldsymbol{W}^\top \boldsymbol{\Gamma}_{\mathbf{Z}}$, $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}} = \boldsymbol{\Gamma}_{\mathbf{X}} - \hat{\boldsymbol{W}}^\top \boldsymbol{\Gamma}_{\mathbf{Z}}$ and $\breve{\boldsymbol{W}} = \boldsymbol{W} - \hat{\boldsymbol{W}}$, where $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_{\mathbf{X}}^\top, \boldsymbol{\Gamma}_{\mathbf{Z}}^\top)^\top$ and $\boldsymbol{\Gamma}$ is defined in Assumption 2. Write $\mathbf{X}_i = \boldsymbol{\Gamma}_{\mathbf{X}}\boldsymbol{\nu}_i$, $\mathbf{Z}_i = \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\nu}_i$, $\boldsymbol{\eta}_i = \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\nu}_i$ and $\hat{\boldsymbol{\eta}}_i = \boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{\nu}_i$. Further, write $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}_{\mathbf{X}}\boldsymbol{\Gamma}_{\mathbf{X}}^\top$, $\boldsymbol{\Sigma}_{\mathbf{Z}} = \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}} = \boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}^\top$.

Lemma C.6. *(Some technical results for the proof of Theorem 2.2.) Under the conditions in Theorem 2.2,*

$$\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta})^2 \lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}, \tag{58}$$

$$\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \mathbf{X}_2 \boldsymbol{\eta}_2^\top \boldsymbol{\beta})^2 \lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}, \tag{59}$$

$$\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \mathbf{X}_2)^2 \lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} \operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{X}}^2) \tag{60}$$

Proof. Denote $\mathbb{E}_{\boldsymbol{\nu}_1}(\cdot) = \mathbb{E}(\cdot|\boldsymbol{\nu}_1)$ where ν is defined in Assumption 2. To prove the inequality (58), we have

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta})^2 &= \mathbb{E}(\boldsymbol{\nu}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}} \boldsymbol{\nu}_1 \boldsymbol{\nu}_1^\top \boldsymbol{\Gamma}_{\mathbf{X}}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Gamma}_{\mathbf{X}} \boldsymbol{\nu}_1) \\ &\lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}.\end{aligned}$$

The inequality then follows from Lemma C.3. To prove the inequality (59), we have

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \mathbf{X}_2 \boldsymbol{\eta}_2^\top \boldsymbol{\beta})^2 &= \mathbb{E}\{\mathbb{E}_{\boldsymbol{\nu}_1}(\boldsymbol{\nu}_2^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \boldsymbol{\Gamma}_{\mathbf{X}} \boldsymbol{\nu}_2)^2\} \\ &\lesssim \mathbb{E}(\mathbf{X}_1^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\eta}_1)^2 = \mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta})^2.\end{aligned}$$

Combining the inequality (58), the inequality (59) is proved. Consider the inequality (60). Observe that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \mathbf{X}_2)^2 &= \mathbb{E}\{\mathbb{E}_{\boldsymbol{\nu}_1}(\boldsymbol{\nu}_2^\top \boldsymbol{\Gamma}_{\mathbf{X}}^\top \mathbf{X}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\eta}_1 \mathbf{X}_1^\top \boldsymbol{\Gamma}_{\mathbf{X}} \boldsymbol{\nu}_2)\} \\ &\lesssim \mathbb{E}(\boldsymbol{\nu}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}} \boldsymbol{\nu}_1 \boldsymbol{\nu}_1^\top \boldsymbol{\Gamma}_{\mathbf{X}}^\top \boldsymbol{\Gamma}_{\mathbf{X}} \boldsymbol{\Gamma}_{\mathbf{X}}^\top \boldsymbol{\Gamma}_{\mathbf{X}} \boldsymbol{\nu}_1) \\ &\lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\beta} \operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{X}}^2).\end{aligned}$$

Thus the inequality (60) is proved. □

Lemma C.7. *(Some technical results for the proof of Theorem 4.2.) Under conditions in Theorem 4.2,*

$$\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\eta}_1 \boldsymbol{\eta}_1^\top \boldsymbol{\beta})^2 \lesssim (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2 \boldsymbol{\beta})^2, \tag{61}$$

$$\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top \boldsymbol{\beta})^2 \lesssim (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2 \boldsymbol{\beta})^2, \tag{62}$$

$$\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\eta}_2)^2 \lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2). \tag{63}$$

Proof. Similar to Lemma C.6, we can prove this lemma by Lemma C.3. At first, we prove inequality (61). Note that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\beta})^2 &= \mathbb{E}(\boldsymbol{\nu}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\nu}_1)^2\\ &\lesssim (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2\boldsymbol{\beta})^2.\end{aligned}$$

Where the equality holds by Assumption 2 and the inequality follows from Lemma C.3. Now we prove inequality (62). Denote $\mathbb{E}_{\boldsymbol{\nu}_1}(\cdot) = \mathbb{E}(\cdot|\boldsymbol{\nu}_1)$. Observe that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top \boldsymbol{\beta})^2 &= \mathbb{E}\{\mathbb{E}_{\boldsymbol{\nu}_1}(\boldsymbol{\nu}_2^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\nu}_2)^2\}\\ &\lesssim \mathbb{E}(\boldsymbol{\eta}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\eta}_1)^2 = \mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\beta})^2.\end{aligned}$$

Combing inequality (61), expression (62) is proved. Now we prove inequality (63). It can be shown that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\eta}_2)^2 &= \mathbb{E}\{\mathbb{E}_{\boldsymbol{\nu}_1}(\boldsymbol{\nu}_2^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\nu}_2)\}\\ &\lesssim \mathbb{E}(\boldsymbol{\eta}_1^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\eta}_1)\\ &= \mathbb{E}(\boldsymbol{\nu}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\nu}_1\boldsymbol{\nu}_1^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\nu}_1)\\ &\lesssim \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{\beta}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2).\end{aligned}$$

The inequalities follow from Lemma C.3 and Assumption 2. □

Lemma C.8. *(Some technical results for the proof of Theorem 5.1.) Under conditions in Theorem 5.1,*

$$\mathrm{tr}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2) \le \{\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2) + \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})\|\breve{\boldsymbol{W}}\|_F^2\}^2 \tag{64}$$

$$\lambda_{\max}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}) \le \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})\|\breve{\boldsymbol{W}}\|_F^2 \tag{65}$$

$$\max_{1\le k\le p_{\boldsymbol{\gamma}}} \|\mathbb{E}(Z_k\hat{\boldsymbol{\eta}})\|_2^2 \le \|\breve{\boldsymbol{W}}\|_F^2\varphi^2. \tag{66}$$

Proof. Note that

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}} &= \boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\eta}}}^\top = (\boldsymbol{\Gamma}_{\boldsymbol{\eta}} + \breve{\boldsymbol{W}}^\top \boldsymbol{\Gamma}_{\mathbf{Z}})(\boldsymbol{\Gamma}_{\boldsymbol{\eta}} + \breve{\boldsymbol{W}}^\top \boldsymbol{\Gamma}_{\mathbf{Z}})^\top\\ &= \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top + \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top \breve{\boldsymbol{W}} + \breve{\boldsymbol{W}}^\top \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top + \breve{\boldsymbol{W}}^\top \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top \breve{\boldsymbol{W}}\\ &= \boldsymbol{\Sigma}_{\boldsymbol{\eta}} + \breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}}\breve{\boldsymbol{W}}\end{aligned}$$

where the last equality holds because of

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^\top &= \boldsymbol{\Gamma}_{\mathbf{Z}}(\boldsymbol{\Gamma}_{\mathbf{X}} - \boldsymbol{\Gamma}_{\mathbf{X}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top(\boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top)^{-1}\boldsymbol{\Gamma}_{\mathbf{Z}})^\top\\ &= \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{X}}^\top - \boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top(\boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{Z}}^\top)^{-1}\boldsymbol{\Gamma}_{\mathbf{Z}}\boldsymbol{\Gamma}_{\mathbf{X}}^\top = \mathbf{0}.\end{aligned}$$

Now we prove the inequality (64). It can be shown that

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}^2) &= \operatorname{tr}\big\{(\boldsymbol{\Sigma}_{\boldsymbol{\eta}} + \breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}})^2\big\} \\
&= \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2) + \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}} \breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}}) + \operatorname{tr}(\breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \operatorname{tr}\big\{(\breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}})^2\big\} \\
&\leq \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2) + 2\operatorname{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2)\operatorname{tr}^{1/2}\big\{(\breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}})^2\big\} + \operatorname{tr}\big\{(\breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}})^2\big\} \\
&\leq \big\{\operatorname{tr}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^2) + \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})\|\breve{\boldsymbol{W}}\|_F^2\big\}^2,
\end{aligned}$$

where the first inequality holds by the matrix Cauchy-Schwartz inequality, the second inequality holds by the fact that

$$\operatorname{tr}^{1/2}\big\{(\breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}})^2\big\} \leq \operatorname{tr}(\breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})\|\breve{\boldsymbol{W}}\|_F^2.$$

Similarly, (65) can be derive as

$$\lambda_{\max}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}) = \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \operatorname{tr}(\breve{\boldsymbol{W}}^\top \boldsymbol{\Sigma}_{\mathbf{Z}} \breve{\boldsymbol{W}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{Z}})\|\breve{\boldsymbol{W}}\|_F^2.$$

For (66), we derive

$$\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} \|\mathbb{E}(Z_k\hat{\boldsymbol{\eta}})\|_2^2 = \max_{1\leq k\leq p_{\boldsymbol{\gamma}}} \|\mathbb{E}(Z_k\breve{\boldsymbol{W}}^\top\mathbf{Z})\|_2^2 \leq \|\breve{\boldsymbol{W}}\|_F^2\varphi^2,$$

where the inequality holds by the fact that

$$\max_{1\leq k\leq p_{\boldsymbol{\gamma}}} \|\mathbb{E}(Z_k\breve{\boldsymbol{W}}^\top\mathbf{Z})\|_2^2 \leq \lambda_{\max}(\breve{\boldsymbol{W}}\breve{\boldsymbol{W}}^\top)\varphi^2$$

and $\lambda_{\max}(\breve{\boldsymbol{W}}\breve{\boldsymbol{W}}^\top) \leq \|\breve{\boldsymbol{W}}\|_F^2$. □

Lemma C.9. *For any random variables $X_1, \ldots, X_m$(without independence assumption) and any fixed integer q,*

$$\|\max_{1\leq i\leq m} X_i\|_q \lesssim \log m \max_{1\leq i\leq m} \|X_i\|_{\psi_1}, \quad \textit{and} \quad \|\max_{1\leq i\leq m} X_i\|_q \lesssim \sqrt{\log m} \max_{1\leq i\leq m} \|X_i\|_{\psi_2}.$$

Proof. By the property of Orlicz norm and Lemma 2.2.2 in [34] (the first is on Page 95), we have

$$\|\max_{1\leq i\leq m} X_i\|_q \lesssim \|\max_{1\leq i\leq m} X_i\|_{\psi_1} \lesssim \log m \max_{1\leq i\leq m} \|X_i\|_{\psi_1}.$$

Similarly,

$$\|\max_{1\leq i\leq m} X_i\|_q \lesssim \|\max_{1\leq i\leq m} X_i\|_{\psi_2} \lesssim \sqrt{\log m} \max_{1\leq i\leq m} \|X_i\|_{\psi_2}.$$ □

Lemma C.10. *Suppose $\mathbf{A}$ is a $p \times p$ dimensional matrix. For any $\boldsymbol{x} \in \mathbb{R}^p$, it can be shown that*

$$|\boldsymbol{x}^\top\mathbf{A}\boldsymbol{x}| \leq \|\mathbf{A}\|_\infty\|\boldsymbol{x}\|_1^2.$$

Proof. Observe that

$$|\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}| = \left|\sum_{i,j=1}^{p} a_{ij}x_i x_j\right| \leq \sum_{i,j=1}^{p} |a_{ij}| \cdot |x_i| \cdot |x_j|$$

$$\leq \|\mathbf{A}\|_\infty \sum_{i,j=1}^{p} |x_i| \cdot |x_j| = \|\mathbf{A}\|_\infty \|\boldsymbol{x}\|_1^2.$$

Where the first inequality holds by the triangle inequality of norm. □

Appendix D: Additional simulation results

Figure 5 displays the empirical size-power curves of $WALD$ in Scenario 1 in section 6 of the main text. We find that with different values $\tau = 0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5$ the test can be either very liberal or very conservative. Thus selecting a proper tuning parameter is difficult in general.

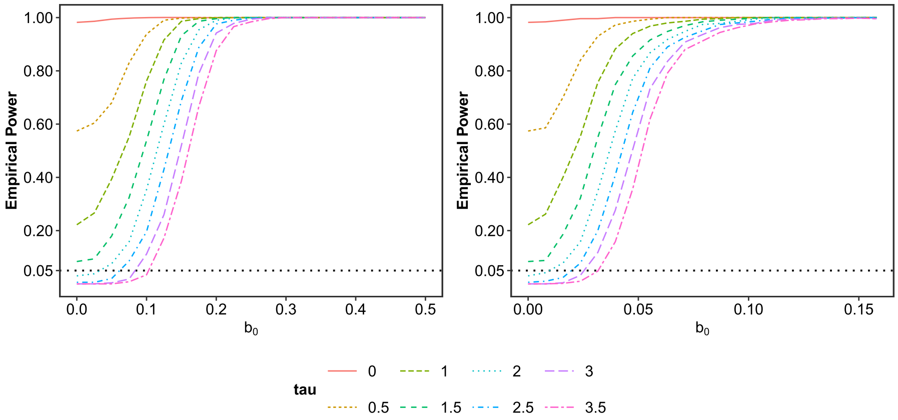

Fig 5. *This figure presents the finite sample performance of $WALD$ in Scenario 1. Left panel presents the empirical sizes and powers of $WALD$ with different tuning parameter τ(τ = $0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5$) in the sparse alternative (setting 2). Right panel corresponds to dense alternative (setting 3).*

Based on the settings in section 6 in the main text, we consider the following five scenarios.

Scenario 3. This scenario investigates the performance of our tests thoroughly when the correlation between covariates of interest and nuisance covariates is weak. Generate the covariates from the multivariate normal distribution $N_p(\mathbf{0}_p, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}, i, j = 1, \ldots, p$ and $\rho = 0.3, 0.5, 0.7$. The sample size $n = 100, 200$, the covariate dimension $p = 600, 1000, 2000, 4000$ and $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}}$. In the sparse alternative (setting 2) and the dense alternative (setting 3), we set $b_0 = \sqrt{\|\boldsymbol{\gamma}\|_2^2 / s_{\boldsymbol{\beta}}}$.

Scenario 4. This scenario is same to Example 2.3 in the main text.

Scenario 5. This scenario investigates the finite performance of our methods thoroughly when $\mathbf{X}$ and $\mathbf{Z}$ are not weakly correlated. Covariates are generated according to the model defined in Scenario 3. Throughout the simulation study, we set $d_{\mathbf{Z}} = 3$ and vary $d_{\mathbf{X}} = 0, 5, 10, 15$, where $d_{\mathbf{X}} = 0$ represents there exists no correlation between $\mathbf{X}$ and $\mathbf{Z}$. The sample size $n = 100, 200$, the predictor dimension $p = 600, 1000, 2000, 4000$ and $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}}$. In the sparse alternative (setting 2) and the dense alternative (setting 3), we set $b_0 = \sqrt{\|\boldsymbol{\gamma}\|_2^2 / s_{\boldsymbol{\beta}}}$.

Scenario 6. We aim to compare our tests with other testing methods in this scenario. The covariates $\boldsymbol{V} = (\mathbf{X}^\top, \mathbf{Z}^\top)^\top$ are generated according to $\boldsymbol{V} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\nu}$, where $\boldsymbol{\nu} = (\nu_1, \dots, \nu_p)^\top$ and $\{\nu_k\}_{k=1}^p$ are generated independently from the $U(-2, 2)$ distribution. Here $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ follows the Toeplitz design, that is, $\sigma_{ij} = 0.5^{|i-j|}, i, j = 1, \dots, p$. The regression error $\epsilon \sim U(-2, 2)$ is independent of $\boldsymbol{V}$. The sample size $n = 100$, the covariate dimension $p = 600$ and $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}} = 300$. In the sparse alternative (setting 2) and the dense alternative (setting 3), we vary b_0 from 0 to $\sqrt{\|\boldsymbol{\gamma}\|_2^2 / s_{\boldsymbol{\beta}}}$.

Scenario 7. This scenario investigates the performance of our tests when $\mathbf{X}$ and $\mathbf{Z}$ are highly correlated. The covariates are generated according to the following model:

$$\mathbf{X} = \boldsymbol{W}^\top \mathbf{Z} + \boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is a $p_{\boldsymbol{\beta}}$-dimensional random vector and $\boldsymbol{\eta}$ is independent of $\mathbf{Z}$. $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{1/2}\boldsymbol{\nu}_{\boldsymbol{\eta}}$ and the elements of $\boldsymbol{\nu}_{\boldsymbol{\eta}}$ are generated independently from the $U(-2, 2)$ distribution, $\mathbf{Z} = \boldsymbol{\Sigma}_{\mathbf{Z}}^{1/2}\boldsymbol{\nu}_{\mathbf{Z}}$ and the elements of $\boldsymbol{\nu}_{\mathbf{Z}}$ are generated independently from the $U(-2, 2)$ distribution. $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}}$ follow the Toeplitz design with $\rho = 0.5$ respectively. $\boldsymbol{W}$ is defined in (10) in subsection 2.3. Throughout the scenario, $d_{\mathbf{Z}} = 3$ and $d_{\mathbf{X}} = 10$. The regression error $\epsilon \sim U(-2, 2)$ is independent of covariates. The sample size $n = 100$, the predictor dimension $p = 600$ and $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}} = 300$. In the sparse alternative (setting 2) and the dense alternative (setting 3), we vary b_0 from 0 to $\sqrt{\|\boldsymbol{\gamma}\|_2^2 / s_{\boldsymbol{\beta}}}$.

Scenario 8. This scenario investigates the power performance of T_n when $\|\boldsymbol{\beta}\|_2$ is fixed but the orientation of $\boldsymbol{\beta}$ is changed. The settings of this scenario are the same as those in Scenario 2 of the main text, except for the following modifications. We set $\boldsymbol{W} = \mathbf{0}$ and $\boldsymbol{\beta} = b_0 \cdot \boldsymbol{e}_k$, where e_k represents a $p_{\boldsymbol{\beta}}$-dimensional vector where the k-th element is 1, and all other elements are 0. We set $b_0 = 3$. As k varies from 1 to 10, the power performances of T_n are 0.808, 0.914, 0.910, 0.906, 0.932, 0.902, 0.910, 0.890, 0.928, and 0.926, respectively. And the difference in power is appreciable with the orientation of $\boldsymbol{\beta}$.

Table 1 reports the simulation results of Scenario 3. We have the following observations. First, T_n and M_n control the type I error well, even when the dimension is 4000. Second, the empirical powers increase when the dimension decreases and the sample size increases. Third, there is no significant difference in power between sparse and dense alternatives as long as $\|\boldsymbol{\beta}\|_2$ stays the same.

Table 1
The empirical type-I errors and powers for T_n and M_n.

			Type-I Error			Power (Sparse)			Power (Dense)		
n	p	ρ	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
100	600	T_n	0.046	0.070	0.060	0.950	0.998	1.000	0.956	1.000	1.000
		M_n	0.050	0.058	0.054	0.964	0.998	1.000	0.950	1.000	1.000
100	1000	T_n	0.038	0.050	0.076	0.820	0.972	1.000	0.826	0.980	1.000
		M_n	0.042	0.050	0.076	0.808	0.974	1.000	0.824	0.980	1.000
100	2000	T_n	0.052	0.058	0.048	0.552	0.800	0.982	0.522	0.760	0.996
		M_n	0.042	0.062	0.042	0.532	0.796	0.984	0.504	0.780	0.990
100	4000	T_n	0.052	0.066	0.068	0.382	0.534	0.816	0.356	0.546	0.848
		M_n	0.046	0.066	0.064	0.376	0.516	0.826	0.348	0.528	0.852
200	600	T_n	0.048	0.056	0.060	1.000	1.000	1.000	1.000	1.000	1.000
		M_n	0.048	0.052	0.052	1.000	1.000	1.000	1.000	1.000	1.000
200	1000	T_n	0.062	0.062	0.054	1.000	1.000	1.000	1.000	1.000	1.000
		M_n	0.070	0.064	0.060	1.000	1.000	1.000	1.000	1.000	1.000
200	2000	T_n	0.070	0.048	0.068	0.982	1.000	1.000	0.978	0.998	1.000
		M_n	0.070	0.042	0.060	0.986	1.000	1.000	0.976	0.998	1.000
200	4000	T_n	0.040	0.050	0.052	0.818	0.956	1.000	0.796	0.986	1.000
		M_n	0.040	0.054	0.052	0.818	0.960	1.000	0.794	0.980	1.000

"Type-I error", "Power (Sparse)," and "Power (Dense)" correspond to Setting 1, Setting 2 and Setting 3.

The left panel of Figure 6 displays the empirical sizes of M_n in Scenario 4. The empirical sizes are close to the nominal level $\alpha = 0.05$ regardless of the increase of $d_{\mathbf{X}}$. The right panel of Figure 6 presents the empirical probability density function of M_n, which can be well approximated by the standard normal distribution no matter what $d_{\mathbf{X}}$ is. These results illustrate that M_n still maintains type I error well when $\mathbf{X}$ and $\mathbf{Z}$ are not weakly dependent.

Table 2 summarizes the empirical sizes and powers of the proposed tests in Scenario 5. Both perform satisfactorily when $d_{\mathbf{X}} = 0$. But when $d_{\mathbf{X}} \neq 0$, M_n performs better than T_n in terms of both the empirical type I error rate and empirical power. As $d_{\mathbf{X}}$ increases, the empirical size of the T_n test deviates significantly from the significant level while the M_n test maintains the empirical size well. Besides, the empirical power of M_n is significantly larger than T_n when $d_{\mathbf{X}} \neq 0$. The results show that the M_n test is more applicable when covariates of interest and nuisance covariates have a strong relationship.

Figure 7 displays the empirical size-power curves of the four tests in Scenario 6. The empirical results in Scenario 6 are similar to those in Scenario 1 of subsection 6.1. It can be observed that T_n, M_n and ST tests control the size well. T_n and M_n are generally more powerful than ST under the sparse and dense alternative hypotheses. Under the dense alternative, the empirical powers of ST can be as low as the significance level. The empirical powers of T_n and M_n increase quickly as the signal strength b_0 becomes stronger. On the other hand, the $WALD$ test is very liberal to have very large empirical size when we use the tuning parameter $\tau = 1$ recommended by [23]. It is worth noticing that T_n and M_n have similar performances in this scenario. T_n performs well enough when the correlation between $\mathbf{X}$ and $\mathbf{Z}$ is relatively weak.

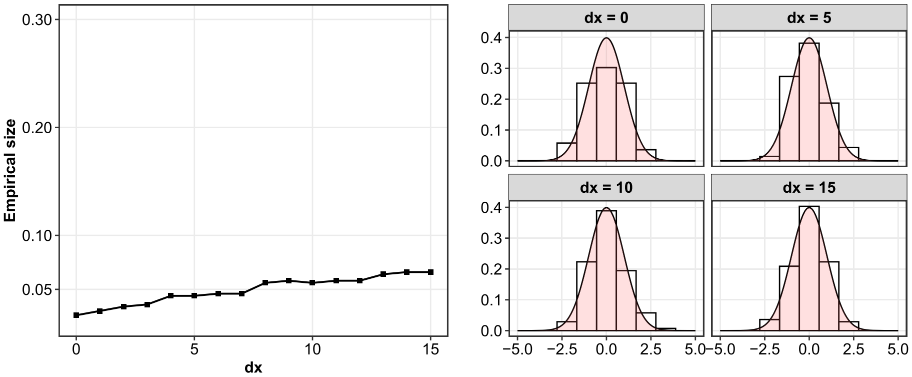

Fig 6. *Left panel presents the empirical sizes of M_n with different $d_{\mathbf{X}}$. Right panel presents the empirical probability density function of M_n with different $d_{\mathbf{X}}$ ($d_{\mathbf{X}} = 0, 5, 10, 15$). The pink shade represents the probability density function of the standard normal distribution. Set $n = 100$, $p = 600$ and $p_{\boldsymbol{\beta}} = p_{\boldsymbol{\gamma}}$. We generate 500 replications and reject the null hypothesis at the significance level $\alpha = 0.05$. More details can be seen in Scenario 3 in Section 6 in the main text.*

Table 2

The empirical type-I errors and powers for T_n test and M_n test.

			Type-I Error				Power (Sparse)				Power (Dense)			
n	p	$d_{\mathbf{X}}$	0	5	10	15	0	5	10	15	0	5	10	15
100	600	T_n	0.038	0.226	0.282	0.296	0.996	0.976	0.816	0.660	1.000	0.982	0.850	0.668
		M_n	0.038	0.052	0.062	0.062	0.994	0.994	0.996	0.998	0.998	1.000	0.998	0.998
100	1000	T_n	0.042	0.210	0.270	0.296	0.980	0.914	0.738	0.584	0.990	0.948	0.748	0.576
		M_n	0.044	0.046	0.054	0.074	0.980	0.986	0.984	0.988	0.992	0.988	0.990	0.990
100	2000	T_n	0.054	0.168	0.242	0.274	0.770	0.706	0.536	0.412	0.786	0.720	0.532	0.426
		M_n	0.052	0.050	0.048	0.062	0.758	0.764	0.772	0.780	0.770	0.776	0.780	0.782
100	4000	T_n	0.038	0.098	0.144	0.192	0.514	0.508	0.380	0.316	0.536	0.476	0.382	0.300
		M_n	0.032	0.038	0.054	0.050	0.520	0.528	0.528	0.520	0.524	0.520	0.530	0.532
200	600	T_n	0.062	0.246	0.294	0.308	1.000	1.000	1.000	0.994	1.000	1.000	1.000	1.000
		M_n	0.062	0.070	0.072	0.076	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
200	1000	T_n	0.072	0.230	0.282	0.298	1.000	1.000	1.000	0.994	1.000	1.000	1.000	0.994
		M_n	0.080	0.084	0.086	0.096	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
200	2000	T_n	0.060	0.184	0.256	0.278	1.000	1.000	0.996	0.954	1.000	1.000	0.992	0.960
		M_n	0.060	0.068	0.072	0.080	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
200	4000	T_n	0.066	0.152	0.248	0.288	0.980	0.962	0.910	0.772	0.974	0.958	0.894	0.772
		M_n	0.064	0.068	0.074	0.080	0.974	0.978	0.980	0.978	0.974	0.976	0.974	0.974

“Type-I error”, “Power (Sparse),” and “Power (Dense)” correspond to Setting 1, Setting 2, and Setting 3.

Figure 8 displays the empirical size-power curves of T_n and M_n in Scenario 7. It can be verified that the empirical results in Scenario 7 are similar to those in Scenario 2 of subsection 6.1. We find that T_n is too liberal to maintain the significance level. On the contrary, M_n maintains the level well. As b_0 increases, although the empirical powers of T_n and M_n increase rapidly, the empirical power of T_n does not go to 1 as the increase of b_0. In contrast, the empirical power of M_n can increase to 1 quickly. The results show that M_n can also improve the power compared with T_n. This confirms the theory.

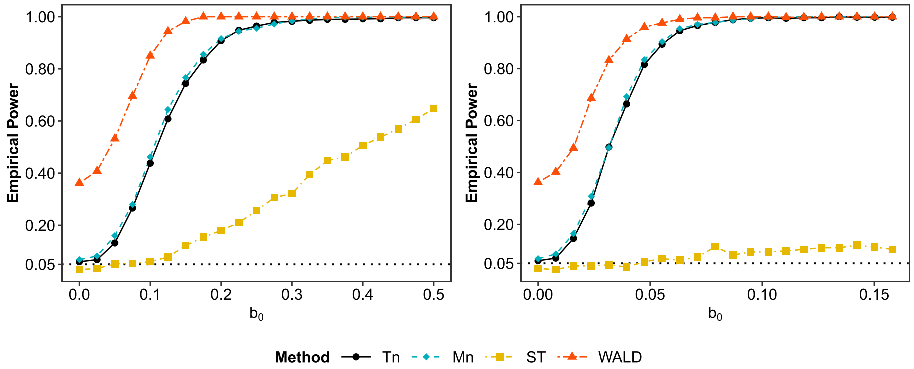

FIG 7. *The left panel represents empirical sizes and powers of the T_n, M_n, ST and $WALD$ in the sparse alternative (setting 2). The right panel corresponds to dense alternative (setting 3). The solid line with circle points, dash line with diamond points, dot-dash line with square points and two-dash line with triangle points represent the empirical sizes and powers of T_n, M_n, ST and $WALD$, respectively.*

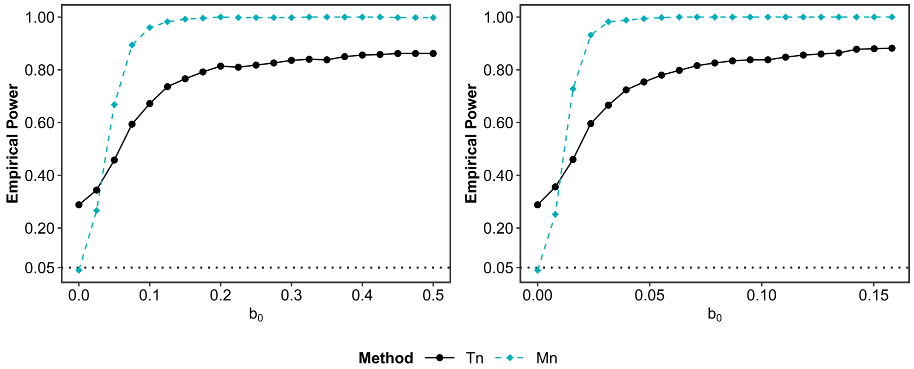

FIG 8. *The left panel represents empirical sizes and powers of the T_n and M_n in the sparse alternatives (setting 2). The right panel corresponds to dense alternatives (setting 3). The solid line with circle points and dash line with diamond points represent the empirical sizes and powers of T_n and M_n, respectively.*

Appendix E: Significant DNA methylation sets in real data analysis in the main text

Table 3 represents the significant DNA methylation sets for T_n in real data analysis in section 6 in the main text. Table 4 correspond to M_n.

Table 3
Numbers of 126 significant DNA methylation sets for T_n.

6 28 33 46 62 83 86 95 103 117 118 144 155 166 168 175 196 197 323 346 365 367 374 386
391 392 407 413 417 424 440 448 457 471 489 495 496 518 539 557 564 587 594 678 706
713 717 718 719 739 757 779 802 807 808 839 843 858 883 924 977 1008 1036 1043 1080
1106 1114 1120 1165 1192 1203 1209 1218 1224 1234 1237 1250 1255 1290 1309 1320 1327
1328 1333 1357 1379 1385 1409 1410 1414 1415 1467 1503 1534 1543 1550 1585 1614 1632
1638 1648 1652 1654 1672 1692 1712 1731 1733 1743 1757 1761 1781 1782 1786 1787 1788
1797 1807 1825 1862 1863 1865 1887 1909 1924 1927

Table 4
Numbers of 149 significant DNA methylation sets for M_n.

6 9 26 28 33 35 62 83 86 95 103 117 118 144 155 159 166 168 175 178 196 197 201 323 329
346 359 365 367 374 386 391 392 394 407 413 417 424 427 432 440 448 457 468 474 489
495 496 510 518 539 557 564 587 594 662 678 706 713 717 719 739 757 779 802 807 808
843 858 871 880 883 891 924 977 1008 1024 1036 1043 1080 1106 1114 1120 1191 1192
1203 1209 1218 1224 1234 1237 1250 1255 1290 1309 1320 1327 1328 1333 1357 1379 1385
1393 1409 1410 1414 1415 1426 1441 1461 1467 1475 1503 1522 1534 1543 1550 1585 1614
1632 1638 1648 1652 1654 1672 1692 1693 1712 1731 1733 1737 1743 1757 1761 1781 1782
1786 1787 1788 1797 1802 1807 1825 1863 1865 1887 1909 1924 1927

Appendix F: Additional real data analysis

We apply our tests to a data set about riboflavin (vitamin B2) production rate with Bacillus Subtilis. This data set was made publicly by [5] and analyzed by several authors, for instance [29], [33], [25], [16], and [17]. It consists of 71 observations of strains of Bacillus Subtilis and 4088 covariates, measuring the log-expression levels of 4088 genes. The response variable is the logarithm of the riboflavin production rate.

Several genes have been previously identified as having associations with the response variable. [29] identified YXLD_at using a multisample-splitting method, while [25] detected YXLD_at and YXLE_at. [33] claimed none. Additionally, [17] reported the presence of YCKE_at, XHLA_at, YXLD_at, YDAR_at, and YCGN_at. In summary, prior literature has identified six genes: YXLD_at, YXLE_at, YCKE_at, XHLA_at, YDAR_at, and YCGN_at, as being associated with the response variable. Denote $\mathcal{G}$ as the set of six selected genes and $\mathcal{G}^c$ as its complement. A natural question is whether the selected genes contribute to the response given the other genes? Consider the following regression modeling:

$$Y = \boldsymbol{\beta}^\top \boldsymbol{V}_{\mathcal{G}} + \boldsymbol{\gamma}^\top \boldsymbol{V}_{\mathcal{G}^c} + \epsilon,$$

where Y is the response variable, $\boldsymbol{V}_{\mathcal{G}}$ is the vector of selected genes, $\boldsymbol{V}_{\mathcal{G}^c}$ denotes the genes in set $\mathcal{G}^c$, and ϵ is the regression error. Further $\|\boldsymbol{\beta}\|_0 = |\mathcal{G}| = 6$, $\|\boldsymbol{\gamma}\|_0 = |\mathcal{G}^c| = 4082$. The null hypothesis of interest is $\mathbb{H}_{01} : \boldsymbol{\beta} = \mathbf{0}$. Thus the testing parameter is $\boldsymbol{\beta}$, and the nuisance parameter is $\boldsymbol{\gamma}$. To verify whether some genes in $\mathcal{G}^c$ contribute to the response given the other genes, we also consider the null hypothesis $\mathbb{H}_{02} : \boldsymbol{\gamma} = \mathbf{0}$. In this testing problem $\mathbb{H}_{02}$, the testing parameter is changed to be $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ is the nuisance parameter correspondingly.

Apply T_n, M_n, ST, and $WALD$. We standardize the data and report the p-values in Table 5. For the testing problem $\mathbb{H}_{01}$, only T_n and M_n reject the null hypothesis at the significance level $\alpha = 0.05$. For the testing problem $\mathbb{H}_{02}$, all tests do not reject the null hypothesis at the significance level $\alpha = 0.05$. The results suggest that the selected gene set $\mathcal{G}$ contributes to the response, and there is no significant gene in $\mathcal{G}^c$.

Table 5
The p-values for a riboflavin data.

Method	T_n	M_n	ST	$WALD$
		p-value		
$\mathbb{H}_{01}$	0.021	0.045	0.160	0.279
$\mathbb{H}_{02}$	0.644	0.632	0.430	0.082

Funding

Xu Guo was supported by the National Key R & D Program of China (grant No. 2023YFA1011100), National Natural Science Foundation of China (grant No. 12322112, 12071038), and the Fundamental Research Funds for the Central Universities. Lixing Zhu was supported by the National Natural Science Foundation of China (grant No. 12131006).

References

[1] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6** 311–329. MR1399305

[2] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. MR3001131

[3] Belloni, A., Chernozhukov, V., Chetverikov, D. and Wei, Y. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Annals of statistics* **46** 3643–3675. MR3852664

[4] Belloni, A., Chernozhukov, V. and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097

[5] Bühlmann, P., Kalisch, M. and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application* **1** 255–278. MR3432840

[6] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., of the Psychiatric Genomics Consortium, S. W. G., Patterson, N., Daly, M. J., Price, A. L. and Neale, B. M. (2015).

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47** 291–295.

[7] Candes, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 551–577. MR3798878

[8] Chen, J., Li, Q. and Chen, H. Y. (2023). Testing generalized linear models with high-dimensional nuisance parameters. *Biometrika* **110** 83–99. MR4565445

[9] Chen, S. X., Peng, L. and Qin, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96** 711–722. MR2538767

[10] Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics* **38** 808–835. MR2604697

[11] Chen, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Annals of Statistics* **46** 642–678. MR3782380

[12] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1–C68. MR3769544

[13] Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields* **162** 47–70. MR3350040

[14] Cui, H., Guo, W. and Zhong, W. (2018). Test for high-dimensional regression coefficients using refitted cross-validation variance estimation. *Annals of Statistics* **46** 958–988. MR3797993

[15] Cui, S., Guo, X. and Zhang, Z. (2024). Estimation and Inference in Ultrahigh Dimensional Partially Linear Single-Index Models. *arXiv preprint arXiv:2404.04471*.

[16] Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test* **26** 685–719. MR3713586

[17] Fei, Z., Zhu, J., Banerjee, M. and Li, Y. (2019). Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach. *Biometrics* **75** 551–561. MR3999178

[18] Friedman, J., Tibshirani, R. and Hastie, T. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

[19] Goeman, J. J., Van De Geer, S. A. and Van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 477–493. MR2278336

[20] Guo, B. and Chen, S. X. (2016). Tests for high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 1079-1102. MR3557190

[21] Guo, W., Zhong, W., Duan, S. and Cui, H. (2022). Conditional Test for Ultrahigh Dimensional Linear Regression Coefficients. *Statistica Sinica* **32** 1381–1409. MR4449899

[22] Guo, X., Li, R., Liu, J. and Zeng, M. (2022). High-dimensional mediation analysis for selecting DNA methylation Loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association* **117** 1110–1121. MR4480694

[23] Guo, Z., Renaux, C., Bühlmann, P. and Cai, T. (2021). Group inference in high dimensions with applications to hierarchical testing. *Electronic Journal of Statistics* **15** 6633–6676. MR4357274

[24] Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., Van Lier, P. A., Meeus, W., Branje, S., Heim, C. M., Nemeroff, C. B., Mill, J. et al. (2016). Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nature communications* **7** 10967.

[25] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15** 2869–2909. MR3277152

[26] Listgarten, J., Kadie, C., Schadt, E. E. and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences* **107** 16465–16470.

[27] Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* **16** 559–616. MR3335800

[28] Ma, R., Tony Cai, T. and Li, H. (2021). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association* **116** 984–998. MR4270038

[29] Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104** 1671–1681. MR2750584

[30] Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics* **45** 158–195. MR3611489

[31] Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons. MR0595165

[32] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[33] Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* **42** 1166–1202. MR3224285

[34] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

[35] van Kesteren, E.-J. and Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal* **26** 710–723. MR4010078

[36] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press. MR3967104

[37] Wu, Y., Wang, L. and Fu, H. (2023). Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension. *Journal of the American Statistical Association* **118** 305–314. MR4571123

[38] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242. MR3153940

[39] Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association* **112** 757–768. MR3671768

[40] Zhong, P.-S. and Chen, S. X. (2011). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association* **106** 260–274. MR2816719

[41] Zhu, L. and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 549–570. MR2278341