Residual Permutation Test for High-Dimensional Regression Coefficient Testing

Kaiyue Wen[1]*, Tengyao Wang[2]*, Yuhao Wang[1,3,4]

[1]Tsinghua University; [2]London School of Economics and Political Science;
[3]Shanghai Artificial Intelligence Laboratory; [4]Shanghai Qi Zhi Institute.

Abstract

We consider the problem of testing whether a single coefficient is equal to zero in high-dimensional fixed-design linear models. In the high-dimensional setting where the dimension of covariates p is allowed to be in the same order of magnitude as sample size n, to achieve finite-population validity, existing methods usually require strong distributional assumptions on the noise vector (such as Gaussian or rotationally invariant), which limits their applications in practice. In this paper, we propose a new method, called *residual permutation test* (RPT), which is constructed by projecting the regression residuals onto the space orthogonal to the union of the column spaces of the original and permuted design matrices. RPT can be proved to achieve finite-population size validity under fixed design with just exchangeable noises, whenever $p < n/2$. Moreover, RPT is shown to be asymptotically powerful for heavy tailed noises with bounded $(1+t)$-th order moment when the true coefficient is at least of order $n^{-t/(1+t)}$ for $t \in [0,1]$. We further proved that this signal size requirement is essentially optimal in the minimax sense. Numerical studies confirm that RPT performs well in a wide range of simulation settings with normal and heavy-tailed noise distributions.

Keywords: distribution-free test, permutation test, finite-population validity, heavy tail distribution, high-dimensional data

1 Introduction

Testing and inference of linear regression coefficients is a fundamental problem in statistics research and has inspired methodological innovations in many other research directions in the statistics community [e.g. Arias-Castro et al., 2011, Zhang and Zhang, 2014, Barber and Candès, 2015, Chernozhukov et al., 2018, Bradic et al., 2019]. In this paper, we consider the setting that we have observations $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n \times \mathbb{R}^n$ generated according to the following model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + b\boldsymbol{Z} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ is an n-dimensional noise vector, and our goal is to test the null hypothesis $H_0 : b = 0$ against the alternative $H_1 : b \neq 0$.

In this paper, we are primarily interested in designing a new coefficient test with finite-population validity. In other words, we require our test to have valid size control with arbitrary magnitude of n, instead

*Equal contribution

of requiring some asymptotic regime assumption that may be unrealistic in practice.When the noise variables are independent and identically distributed (i.i.d.) Gaussian random variables and $p < n$, the ANOVA test [Fisher, 1973] can be used to test H_0 against H_1 with finite-population valid Type-I error control. While the Gaussianity assumption is convenient for theoretical analysis, it is in general not realistic in practical applications, which limits the applicability of the ANOVA test. Indeed, as we will see in Section 3, the size of ANOVA test can be far from the nominal level in the presence of heavy-tailed noises. This motivates us to propose a new test that is finite-population valid without such restrictive distributional assumptions. In particular, instead of the independent Gaussian distribution assumption above, we only assume that the noise $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ have *exchangeable components*:

Assumption 1 (Exchangeable noise)**.** For any permutation σ of indices $1, \ldots, n$,

$$(\varepsilon_1, \ldots, \varepsilon_n) \overset{\mathrm{d}}{=} (\varepsilon_{\sigma(1)}, \ldots, \varepsilon_{\sigma(n)}).$$

A common approach to handle exchangeable noise is through the idea of permutation tests [Pitman, 1937a,b, 1938]. Recently, Lei and Bickel [2021] implemented this idea to the problem of regression coefficient testing. In their seminal work, the authors proposed *cyclic permutation test* that achieved finite population validity under Assumption 1 by exploiting the exchangeability of the noise terms. However, to achieve a size α control, their cyclic permutation test requires that $n/p \geq 1/\alpha - 1$. For example, for a sample size of $n = 300$ and a targeting Type-I error rate is $\alpha = 0.01$, at most $p = 2$ covariates are allowed in $\boldsymbol{X}$. This limits the applicability of their test in moderately large dimensions. In this paper, we consider the more challenging question of finite-population Type-I error control in the high dimensional setting where p is allowed to be of the same order of magnitude as n. We propose *residual permutation test* (RPT), a permutation-based approach that performs hypothesis tests by manipulating the empirical residuals after regression adjustment. The proposed test is guaranteed to have the correct Type-I error control whenever $p < n/2$. Moreover, our result is fixed design and does not require any regularity conditions on the design matrix $\boldsymbol{X}$.

In addition to proving its finite-population validity, we further analyze the statistical power of the proposed test in high dimensions, especially when the ε_i's follow a heavy-tailed distribution. As we will discuss further in Section 2.3, statistical methods with robustness to heavy-tailed data have significant demands in practice [Eklund et al., 2016, Wang et al., 2015, Cont, 2001], and has been actively studied in both modern statistics and theoretical computer science communities. Despite its importance, there is a lack of available tools that can handle high dimensional regression coefficient testing with heavy-tailed noise. In this paper, we fill this gap by showing that when the ε_i's are i.i.d. and have a finite $(1 + t)$-th order moment for any $t \in [0, 1]$, we can still construct a test with non-trivial statistical power in high dimensions. Specifically, we prove that when $n/p \geq 3 + m$ for some $m > 0$, our proposed test is guaranteed to have power converging to 1 whenever the coefficient b is of order at least $n^{-t/(1+t)}$. We also studied the minimax optimality of high-dimensional coefficient testing with heavy-tailed noises; and proved that in the presence of heavy-tailed noise with only a finite $(1 + t)$-th moment, the $n^{-t/(1+t)}$ order requirement for b is essentially optimal.

Since ANOVA has been used extensively in practical applications, as an independent contribution, we provide a more comprehensive analysis of the ANOVA test. Specifically, we show that ANOVA is finite population valid when either the design or the noise follows a *spherically symmetric distribution*, a condition that is slightly weaker than the Gaussianity assumption. On the other hand, our simulation analysis show that ANOVA is indeed not validity when such spherical distributional assumptions are violated. At the same, we propose another permutation-based test: naive residual permutation test (naive RPT), which like ANOVA, is also valid under spherically symmetric noise distribution whenever $p < n$. While naive RPT is still not

valid for non-spherically symmetric noises, it does appear to have smaller Type I error violations compared to ANOVA.

In sum, we make the following contributions

- We propose a new test that has finite population validity with fixed-design linear models and exchangeable noises in the high dimensional setting where $p < n/2$.
- We prove that when the noise variables are heavy-tailed with bounded $(1+t)$-th order moment for $t \in [0, 1]$, our test is asymptotically powerful when b is at least of order $n^{-t/(1+t)}$.
- We perform numerical analysis to show that ANOVA is indeed invalid in general distributions, especially with heavy-tailed data. We also studied other theoretical properties of ANOVA.
- We discuss the minimax optimality of regression coefficient test with heavy-tailed distributions, and show that our test is essentially optimal in the minimax sense.

The rest of this paper is organized as follows. In Section 2, we review existing results in high-dimensional coefficient testing, conditional independence testing and heavy-tailed data. In Section 3, we provide more studies on the finite-sample properties of ANOVA test with non-Gaussian noises, and propose a new test that is easier to implement and more robust to non-Gaussianity. As ANOVA test has been heavily used in practical applications, we believe this is of independent interest. In Section 4, we present our method, and prove its finite population validity. In Sections 5 and 6, we provide power analysis of RPT and study the minimax optimality of high-dimensional coefficient testing under some heavy-tailed assumptions. Finally, in Section 7 we provide numerical analysis. In Section 8, we finalize the manuscript with a discussion.

Notation

We conclude this section by introducing some notation used throughout the paper. For any $n \times p$ dimensional matrix $\boldsymbol{A}$, we denote by $\mathrm{span}(\boldsymbol{A})$ the subspace spanned by the p column vectors of $\boldsymbol{A}$; and we write $\mathrm{span}(\boldsymbol{A})^{\perp}$ as the space that is orthogonal to $\mathrm{span}(\boldsymbol{A})$. Given an n-dimensional vector $\boldsymbol{a}$, we denote by $\mathrm{Proj}_{\boldsymbol{A}}(\boldsymbol{a})$ the projection of $\boldsymbol{a}$ onto the subspace $\mathrm{span}(\boldsymbol{A})$, and denote by $\|\boldsymbol{a}\|_2$ as the ℓ_2-norm of the vector $\boldsymbol{a}$. Given two $n \times q_1$ and $n \times q_2$ dimensional matrices $\boldsymbol{A}$, $\boldsymbol{B}$, we denote by $(\boldsymbol{A}, \boldsymbol{B})$ as the $n \times (q_1 + q_2)$ matrix via column concatenation of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. We write $\mathcal{N}(0, 1)$ as standard normal distribution. For two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n = O(b_n)$, or equivalently $b_n = \Omega(a_n)$, if there exists a universal constant $C > 0$ such that $|a_n| \leq C|b_n|$ for all n; we write $a_n = o(b_n)$, or equivalently $b_n = \omega(a_n)$, if $|a_n|/|b_n| \to 0$.

2 Literature review

Our work spans a wide range of research directions, including high dimensional coefficient testing, permutation-based hypothesis tests and high dimensional heavy-tailed problems. In this section, we compare our research to works within each direction.

2.1 High dimensional testing of regression coefficients

The most classical approach for testing the null hypothesis $b = 0$ is through the analysis of variance (ANOVA) test [Fisher, 1973]. ANOVA test was originally proposed by Sir Ronald Fisher in the 1920s, and

has been widely used in economics [Doane and Seward, 2016], finance [Paolella, 2018] and biology [Lazic, 2008] etc. Under the context of single coefficient testing, when $n > p + 1$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ for some $\sigma^2 > 0$, if $\tilde{\beta} := \text{argmin}_\beta \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2$ and $(\hat{\beta}, \hat{b}) := \text{argmin}_{(\beta,b)} \|\boldsymbol{Y} - \boldsymbol{X}\beta - b\boldsymbol{Z}\|_2^2$, then under H_0, the test statistic

$$\phi_{\text{anova}} := \frac{\|\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta} - \hat{b}\boldsymbol{Z}\|_2^2}{\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta} - \hat{b}\boldsymbol{Z}\|_2^2/(n-p-1)} \sim F_{1,n-p-1} \tag{2}$$

can be used to construct a test where H_0 is rejected when ϕ_{anova} exceeds the $1-\alpha$ quantile of the $F_{1,n-p-1}$ distribution. As the above distributional result is nonasymptotic and holds whenever $n > p + 1$, the associated test is valid even in the high-dimensional setting. However, as we will discuss in Section 3, beyond Gaussianity and some other class of restrictive assumptions on $\boldsymbol{\varepsilon}$, ANOVA test is usually *not* guaranteed to have a valid Type-I error control. This encourages us to construct hypothesis tests with valid Type-I error control allowing a broader class of noise distributions.

As emphasized by Lei and Bickel [2021], this is a challenging problem, with a "century long effort" in the statistical community to alleviate the strong Gaussianity assumption of ANOVA. Some representative works include Hartigan [1970], Meinshausen [2015]. However, the two methods mentioned above still require the noise to follow certain geometric constraint, which is either symmetric about 0 or rotationally invariant. Lei and Bickel [2021] represented, to the best of our knowledge, the first work that established finite-population size control with only exchangeable noise. However, as mentioned in the introduction, despite its striking distribution-free property, the cyclic permutation test proposed in Lei and Bickel [2021] requires the dimension of p to be much smaller than n for valid size control, and no corresponding statistical power analysis was provided. Another alternative with less restrictive assumptions on dimension p was proposed in D'Haultfœuille and Tuvaandorj [2022], where the authors proposed a "stratified randomization test". Different from our test that is fixed design and allows arbitrary $\boldsymbol{X}$, D'Haultfœuille and Tuvaandorj [2022] assume that rows of $\boldsymbol{X}$ must follow a discrete random distribution with a relatively small number of unique values.

Besides finite-population validity, a less demanding criteria for coefficient test is the *asymptotic validity*. By invoking 1) certain sparsity conditions on the regression coefficients; 2) some regularity conditions on the design matrix $\boldsymbol{X}$ and 3) sharp tail bounds on the noise variables, debiased lasso is guaranteed to establish asymptotically valid p-value and confidence intervals for regression coefficients [Zhang and Zhang, 2014, Van de Geer et al., 2014, Javanmard and Montanari, 2014]. Recall that our test is finite population valid with arbitrary design and coefficient and has non-trivial power even with heavy-tailed $\boldsymbol{\varepsilon}$. Other follow up studies of debiased lasso include Zhu and Bradic [2018], Bradic et al. [2019], Shah and Bühlmann [2019], to name a few.

More broadly speaking, regression coefficient test can be viewed as a subdomain of the more general conditional independence testing, i.e., testing the null hypothesis $Y \perp\!\!\!\perp Z \mid X$, treating $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ as i.i.d. realizations from some hypothesized superpopulation. Unfortunately, when one has no assumption on the joint distribution of the random variables X, Y and Z, Shah and Peters [2020] proved that it is a "statistically hard problem", in the sense that a valid test for the null does not have power against *any* alternative. This means that some restrictions must be added to the class of null distributions to have some power. Following this insight, an important research question then, is to propose valid test under minimal distributional assumptions. In this paper, we show that a linear functional relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is sufficient to have exact validity with non-trivial power.

2.2 Permutation based hypothesis tests

As also mentioned in the introduction section, our new method is based on permutation test [Pitman, 1937a,b, 1938]. Permutation test was originally developed for independence testing. Specifically, using the exchangeability properties of the sampled data, permutation test is guaranteed to have finite-sample validity guarantee, without any geometric or moment constraints on the underlying distributions. Thanks to such distribution-free property, permutation tests and its extensions have also been used in coefficient tests [Lei and Bickel, 2021, D'Haultfœuille and Tuvaandorj, 2022], sharp null hypothesis tests [Caughey et al., 2017, 2021] and conditional independence tests [Berrett et al., 2020, Kim et al., 2021] for finite-population or asymptotically valid Type-I error control.

2.3 Heavy-tailed data

To understand the efficiency of the proposed method in heavy tailed data, in this paper, we further provide power analysis when the noise terms follow a heavy-tailed distribution. In classical high-dimensional literature, due to the simplicity of theoretical analysis, existing methods usually focus on data with sharp tail bounds, such as sub-Gaussian or sub-exponential tail bounds [see, e.g. Wainwright, 2019]. However, as also discussed by Sun et al. [2020], such strong tail condition may not be reasonable in real world applications, such as neuroimaging [Eklund et al., 2016], gene expression analysis [Wang et al., 2015], and finance [Cont, 2001].

Since the pioneering work by Catoni [2012], the problem of extracting useful information from heavy-tailed data (or the related adversarially contaminated data) has been an active area of research in mathematical statistics and theoretical computer science literature in the past ten years [Bubeck et al., 2013, Lykouris et al., 2018, Lugosi and Mendelson, 2019, Sun et al., 2020, Fan et al., 2021]. In the high dimensional setting where we allow the dimension p to grow with n, heavy-tailed data has been actively studied in mean estimation [Lugosi and Mendelson, 2019, 2021], regression coefficient estimation [Wang, 2013, Fan et al., 2017, Sun et al., 2020, Pensia et al., 2020] and covariance matrix analysis [Loh and Tan, 2018, Fan et al., 2021]. The definition of "heavy-tail" may vary across different articles. Among all the heavy tail literature, our heavy tail assumption is the same as the one in Sun et al. [2020], Bubeck et al. [2013], which assume that the noise variables has at most a finite $(1+t)$-th order moments for some $t \in (0, 1]$ without any geometric or shape constraints. To our knowledge this is also the weakest heavy tail assumption studied in the literature (or at least in the high dimensional literature).

Nevertheless, under the context of coefficient testing, existing methods on heavy-tailed data seem still limiting. In this paper, we fill this gap by providing statistical power guarantee of our constructed test with heavy tail noises. Our power analysis stems from our new theoretical insight on the asymptotic convergence of heavy-tailed random variables after subspace projections. It would be of interest if these results could be extended to understand the power of permutation-testing based hypothesis tests in other heavy-tailed scenarios.

3 Finite-population validity of ANOVA beyond Gaussianity

As ANOVA has been frequently used in empirical analysis, it would be of interest to provide a more comprehensive analysis on the sensitivity of ANOVA test with respect to the Gaussianity assumption, both empirically and theoretically. First, we show that in fact, by Lemma 1 below, ANOVA is valid as long as either the noise $\boldsymbol{\varepsilon}$ or the design matrix $(\boldsymbol{X}, \boldsymbol{Z})$ follows a spherically symmetric distribution.

Definition 1. We say that a random matrix $\boldsymbol{A} \in \mathbb{R}^{n\times q}$ follows a spherically symmetric distribution if for any $\boldsymbol{Q} \in \mathbb{O}^{n\times n}$, $\boldsymbol{A} \overset{\mathrm{d}}{=} \boldsymbol{Q}\boldsymbol{A}$, where $\mathbb{O}^{n\times n}$ is the set of $n \times n$ orthonormal matrices.

Lemma 1. *Suppose $\boldsymbol{Y}$ is generated under (1) with $\beta \in \mathbb{R}^p, b = 0$. Suppose also that $\boldsymbol{\varepsilon}$ is a random vector that is almost surely not a zero vector, $(\boldsymbol{X}, \boldsymbol{Z})$ is either deterministic or independent from $\boldsymbol{\varepsilon}$. If either $\boldsymbol{\varepsilon}$ or $(\boldsymbol{X}, \boldsymbol{Z})$ follows a spherically symmetric distribution, then the test statistic ϕ_{anova} defined in (2) satisfies $\phi_{\mathrm{anova}} \sim F_{1,n-p-1}$.*

The spherical symmetry in the noise or the design is slightly weaker than the usual Gaussianity constraint, however, it is still too strong for many real data applications. For instance, if we assume that observations (X_i, Z_i, Y_i) are independent, then this assumption amounts to either i.i.d. normal noise or an i.i.d. multivariate normal design.

We now perform a numerical experiment to analyze the validity of ANOVA test under general distributional classes of $\boldsymbol{\varepsilon}$. We generate data $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ according to the model specified in (1) and that

$$\boldsymbol{Z} = \boldsymbol{X}\beta^Z + \boldsymbol{e}. \tag{3}$$

In the simulation, we set $b = 0$; since the result of ANOVA is invariant to β, β^Z, we simply set them to be zero vectors. We also set $\boldsymbol{X}$ as $n \times p$ matrices with i.i.d. entries following either $\mathcal{N}(0, 1)$ or t_1 distribution, with $(n, p) = (300, 100), (600, 100)$ or $(600, 200)$; and $\boldsymbol{e}$ and $\boldsymbol{\varepsilon}$ have i.i.d. components from one of $\mathcal{N}(0, 1)$, t_2 or t_1 distributions.

Table 1 summarizes the performance of ANOVA test from 100000 Monte Carlo simulations. For evaluation criterion we consider the sizes of the ANOVA test at nominal levels $\alpha = 0.01, 0.005$ among the 100000 replicates. According to the simulation results, when the noises of $\boldsymbol{e}$ and $\boldsymbol{\varepsilon}$ follows a standard normal distribution, the ANOVA test has the correct size control, which is consistent with Lemma 1. However, when normality is violated, the ANOVA test will be overly optimistic, with an empirical size more than twice as large as the nominal level in some cases. In particular, the performance of noise type t_1 is in general worse than that of t_2, this means that ANOVA test is more vulnerable to heavy-tailed noises. Moreover, the performance of ANOVA is worse with a heavy-tailed design matrix $\boldsymbol{X}$.

To better understand the empirical distribution of the simulated p-values, we plot the histogram of Monte Carlo repetitions in Figure 1. Figure 1(a)-(c) corresponds to the histogram of the p-values from the ANOVA test. Apparently, all the histograms are far from uniform on $[0, 1]$ under the null hypothesis, with a large spike near zero. In addition, the magnitude of the spike gets higher as n becomes smaller or that $\boldsymbol{\varepsilon}$ or $\boldsymbol{X}$ becomes more heavy-tailed. Another interesting property is that the histograms are usually "U-shaped", where the peaks appear at regions near either 1 or 0. In sum, when data are generated from non-Gaussian and in particular heavy-tailed distributions, the ANOVA tests are usually far from the correct level and the aim of the current paper is to propose a new test that 1) is finite-population valid in high dimensions just with exchangeable noises and 2) has power even in heavy-tailed distribution.

It is worth noting that when $\beta = 0$, we can easily construct a valid permutation test by computing the correlation of Y to Z and to its permutations. From this intuition, a straightforward approach is to first regress both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ onto $\boldsymbol{X}$ to eliminate the influence of $\boldsymbol{X}$, and then to use regression residuals for permutation test construction. Specifically, let $\hat{\boldsymbol{R}}_\varepsilon := (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{Y}$ and $\hat{\boldsymbol{R}}_e := (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{Z}$ be the regression residuals after projecting $\boldsymbol{Y}$ and $\boldsymbol{Z}$ onto $\boldsymbol{X}$ respectively. Let $\boldsymbol{V}_0 \in \mathbb{R}^{n\times(n-p)}$ be a matrix with orthonormal columns spanning an $(n-p)$-dimensional subspace of $\mathrm{span}(\boldsymbol{X})^\perp$, then $\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{V}_0\boldsymbol{V}_0^\top$. Hence under $H_0 : b = 0$, the regression residuals $\hat{\boldsymbol{R}}_\varepsilon$ satisfy $\hat{\boldsymbol{R}}_\varepsilon = \boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{Y} = \boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{\varepsilon}$. From above, we construct a test, which we call as *naive residual permutation*

				ANOVA		Naive	
n	p	X type	noise type	0.01	0.005	0.01	0.005
300	100	Guassian	Guassian	$0.0101_{(0.0003)}$	$0.0050_{(0.0002)}$	$0.010_{(0.0003)}$	$0.0049_{(0.0002)}$
300	100	Guassian	t_1	$0.0181_{(0.0004)}$	$0.0160_{(0.0004)}$	$0.0158_{(0.0004)}$	$0.0116_{(0.0003)}$
300	100	Guassian	t_2	$0.0153_{(0.0004)}$	$0.0107_{(0.0003)}$	$0.0139_{(0.0004)}$	$0.0089_{(0.0003)}$
300	100	t_1	Guassian	$0.0101_{(0.0003)}$	$0.0050_{(0.0002)}$	$0.0103_{(0.0003)}$	$0.0049_{(0.0002)}$
300	100	t_1	t1	$0.0243_{(0.0005)}$	$0.0208_{(0.0005)}$	$0.0158_{(0.0004)}$	$0.0107_{(0.0003)}$
300	100	t_1	t_2	$0.0180_{(0.0004)}$	$0.0130_{(0.0004)}$	$0.0141_{(0.0004)}$	$0.0088_{(0.0003)}$
600	100	Guassian	Guassian	$0.0095_{(0.0003)}$	$0.0050_{(0.0002)}$	$0.0096_{(0.0003)}$	$0.0048_{(0.0002)}$
600	100	Guassian	t_1	$0.0163_{(0.0004)}$	$0.0143_{(0.0004)}$	$0.0128_{(0.0004)}$	$0.0080_{(0.0003)}$
600	100	Guassian	t_2	$0.0169_{(0.0004)}$	$0.0120_{(0.0003)}$	$0.0128_{(0.0004)}$	$0.0076_{(0.0003)}$
600	100	t_1	Guassian	$0.0105_{(0.0003)}$	$0.0050_{(0.0002)}$	$0.0102_{(0.0003)}$	$0.0052_{(0.0002)}$
600	100	t_1	t1	$0.0188_{(0.0004)}$	$0.0166_{(0.0004)}$	$0.0106_{(0.0003)}$	$0.0058_{(0.0002)}$
600	100	t_1	t_2	$0.0174_{(0.0004)}$	$0.0130_{(0.0004)}$	$0.0114_{(0.0003)}$	$0.0063_{(0.0002)}$
600	200	Guassian	Guassian	$0.0101_{(0.0003)}$	$0.0049_{(0.0002)}$	$0.0103_{(0.0003)}$	$0.0050_{(0.0002)}$
600	200	Guassian	t_1	$0.0141_{(0.0004)}$	$0.0122_{(0.0003)}$	$0.0124_{(0.0004)}$	$0.0090_{(0.0003)}$
600	200	Guassian	t_2	$0.0150_{(0.0004)}$	$0.0104_{(0.0003)}$	$0.0136_{(0.0004)}$	$0.0089_{(0.0003)}$
600	200	t_1	Guassian	$0.0101_{(0.0003)}$	$0.0049_{(0.0002)}$	$0.0098_{(0.0003)}$	$0.0049_{(0.0002)}$
600	200	t_1	t_1	$0.0202_{(0.0004)}$	$0.0173_{(0.0004)}$	$0.0133_{(0.0004)}$	$0.0086_{(0.0003)}$
600	200	t_1	t_2	$0.0170_{(0.0004)}$	$0.0120_{(0.0003)}$	$0.0134_{(0.0004)}$	$0.0080_{(0.0003)}$

Table 1: Sizes of the ANOVA test and naive residual permutaion test, estimated over 100000 Monte Carlo repetitions, for various noise distributions at nominal levels of $\alpha = 0.01$ and $\alpha = 0.005$. Data are generated by models (1) and (3), with $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ having independent components distributed according to the various X types and noise types described in the table. Standard errors of the estimated sizes are given in parentheses.

test, based on the *projected residuals* $\hat{\boldsymbol{\varepsilon}} := \boldsymbol{V}_0^\top \hat{\boldsymbol{R}}_\varepsilon = \boldsymbol{V}_0^\top \boldsymbol{Y}$ and $\hat{e} := \boldsymbol{V}_0^\top \hat{\boldsymbol{R}}_e = \boldsymbol{V}_0^\top \boldsymbol{Z}$ as

$$\phi_{\text{naive}} = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}(|\hat{\boldsymbol{e}}^\top \hat{\boldsymbol{\varepsilon}}| \le |\hat{\boldsymbol{e}}^\top \boldsymbol{P}_k \hat{\boldsymbol{\varepsilon}}|)\right), \tag{4}$$

where the $\boldsymbol{P}_k \in \mathbb{R}^{(n-p)\times(n-p)}$'s are random permutation matrices that are sampled uniformly at random from the set of all permutation matrices. Lemma 2 shows that under a slightly weaker condition than Lemma 1, ϕ_{naive} is a valid test.

Lemma 2. *Suppose $\boldsymbol{Y}$ is generated under (1) with $\beta \in \mathbb{R}^p$, $b = 0$. If either*

(a) $\boldsymbol{\varepsilon}$ or $(\boldsymbol{X}, \boldsymbol{Z})$ follows a spherically symmetric distribution;

(b) $\boldsymbol{Z}$ is generated under (3) and either $\boldsymbol{e}$ or $(\boldsymbol{X}, \boldsymbol{Y})$ follows a spherically symmetric distribution,

ϕ_{naive} is valid p-value, i.e., for all $\alpha \in (0, 1)$, $\mathbb{P}(\phi_{\text{naive}} \le \alpha) \le \alpha$.

The conditions for Lemma 2 is slightly weaker than Lemma 1. However, Lemma 2 still requires the spherically symmetric distribution. To better understand their empirical performances, we also show the performance of ϕ_{naive} with non-Gaussian noises or non-Gaussian designs in Table 1 and Figures 1(d)-(f) .

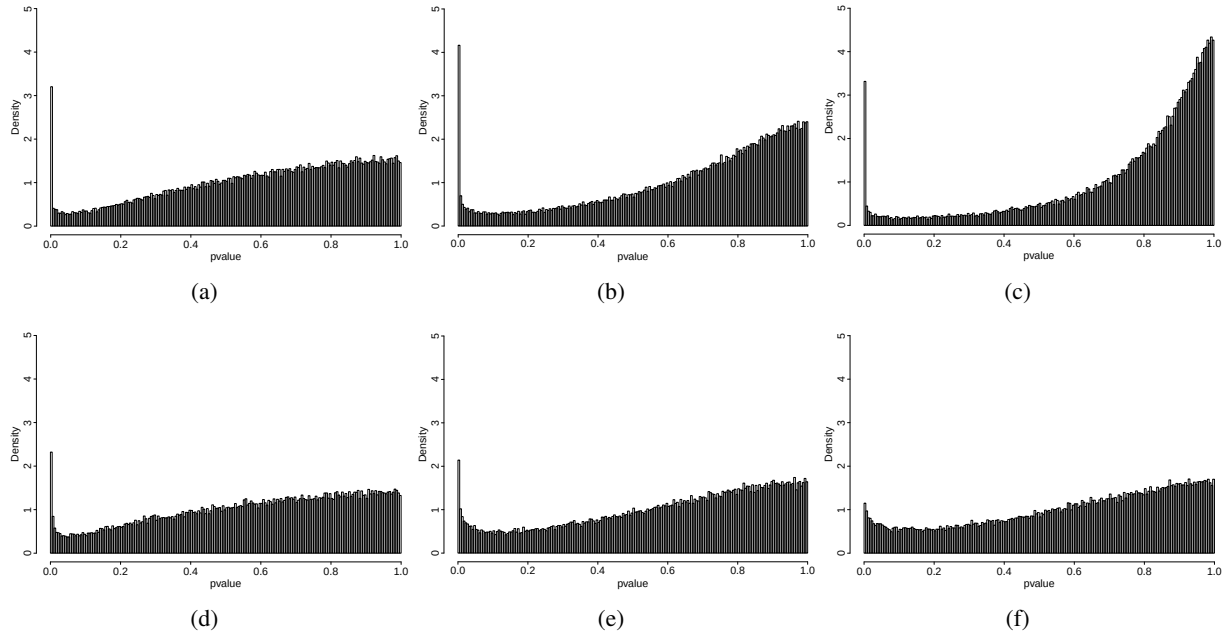


Figure 1: Histogram of p-values under the null for ANOVA test and naive residual permutation test from 100000 Monte-Carol replicates. The first line are the histograms of the ANOVA test under different specifications. Specifically, (a) is the result with Gaussian design, $n = 300, p = 100$ and $\boldsymbol{\varepsilon}$ has independent t_1 components; (b) is the histogram with the same setting as in (a) except that we switch from Gaussian design to t_1 design; (c) is the histogram with Gaussian design, $n = 600, p = 100$ and $\boldsymbol{\varepsilon}$ has independent t_1 components. The second line are the histogram for naive test. (e)-(f) use the same simulation settings as (a)-(c).

Consistent with the theoretical findings, without the strong Gaussianity or shperically symmetry assumption, ϕ_{anova} is also not guaranteed to have finite-population validity. Nevertheless, when both tests are invalid, the size of naive permutation test is closer to the correct level than its competitor. This indicates that naive test is more robust to non-Gaussian distributions. Moreover, the naive test is an intuitive method and is easy to implement. Thus, the naive test could be used as a preferrable alternative to ANOVA in real data analysis when $n/2 \leq p < n$.

Although the empirical finding suggests that the naive RPT is more robust to Gaussian violations, the question remains: how can we construct a hypothesis test that is finite-population valid just with arbitrary exchangeable noises in high dimensions where p can be in the same order of magnitude as n? We answer this question in the next section.

4 Residual permutation test: methodology and validity

In Section 3, we have shown from simulation experiments that a naive permutation test on the residuals, although more robust than ANOVA, is still not guaranteed to have finite-population validity with just exchangeable noises. In this section we describe a more refined test using the projected residuals $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\boldsymbol{e}}$, which we call *residual permutation test* (RPT), and present its finite-population validity guarantee in Theorem 2. For intuitions behind such construction, we refer the readers to Section 4.1.

To describe RPT, we write $\mathcal{P}$ for the set of all permutation matrices in $\mathbb{R}^{n\times n}$ and we denote by $\boldsymbol{P}_0 = \boldsymbol{I} \subseteq \mathcal{P}$ the identity matrix. To successfully perform the regression permutation test, we first need to randomly generate of a sequence of K permutation matrices $\{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\} \subseteq \mathcal{P} \setminus \{\boldsymbol{P}_0\}$, such that together with $\boldsymbol{P_0}$ they form a group:

Assumption 2. The set of permutation matrices $\mathcal{P}_K := \{\boldsymbol{P}_0, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}$ satisfies that for any $\boldsymbol{P}_i, \boldsymbol{P}_j$, there exists a $k \in \{0, \ldots, K\}$ such that $\boldsymbol{P}_k = \boldsymbol{P}_i\boldsymbol{P}_j$.

We write $\boldsymbol{V}_0 \in \mathbb{R}^{n\times(n-p)}$ as a matrix with orthonormal columns spanning an $(n-p)$-dimensional subspace of $\mathrm{span}(\boldsymbol{X})^\perp$ and $\boldsymbol{V}_k := \boldsymbol{P}_k\boldsymbol{V}_0$.[1] In addition, we denote by $\tilde{\boldsymbol{V}}_k \in \mathbb{R}^{n\times(n-2p)}$ a matrix with orthonormal columns spanning a subspace of $\mathrm{span}(\boldsymbol{V}_0) \cap \mathrm{span}(\boldsymbol{V}_k)$. Recall that $\hat{\boldsymbol{e}} := \boldsymbol{V}_0^\top \boldsymbol{Z}$ and $\hat{\boldsymbol{\varepsilon}} := \boldsymbol{V}_0^\top \boldsymbol{Y}$. Given a fixed $T : \mathbb{R}^{n-2p} \times \mathbb{R}^{n-2p} \to \mathbb{R}$, we can calculate the p-value of our coefficient test via:

$$\phi := \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{\varepsilon}}\right) \le T\left(\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\hat{\boldsymbol{\varepsilon}}\right)\right\}\right). \quad (5)$$

Notice that T can be any function. For example, one can choose $T(x, y) = |\langle x, y\rangle|$. As demonstrated in the Supplementary Material, the above definition of ϕ can be simplified as the following equivalent form

$$\phi := \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{Z}, \tilde{\boldsymbol{V}}^\top\boldsymbol{Y}\right) \le T\left(\tilde{\boldsymbol{V}}_k^\top\boldsymbol{Z}, \tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{Y}\right)\right\}\right). \quad (6)$$

The following theorem shows that the proposed p-value is uniformly valid under the null:

Theorem 2. *Suppose that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is generated under model (1) with $p < n/2$ and that the noise $\boldsymbol{\varepsilon}$ satisfies Assumption 1. Suppose $\{\boldsymbol{P}_k : k = 0, \ldots, K\}$ satisfies Assumption 2. Under $H_0 : b = 0$, ϕ defined in (6) is a valid p-value, i.e. $\mathbb{P}(\phi \le \alpha) \le \alpha$ for all $\alpha \in [0, 1]$.*

We remark that as shown in Theorem 2, an important advantage of RPT is that the result is finite-population such that it holds for arbitrary size of n. Moreover, our result assumes a fixed-design matrix and does not require any assumption on $\boldsymbol{X}$ for finite-population validity. For example, the rank of $\boldsymbol{X}$ even does not necessarily need to be p. Also, Theorem 2 shows that RPT has valid size for any choice of function $T(\cdot, \cdot)$ and number of permutations K. However, in practice, to have good power under the alternative, we typically set $T(x, y) = |\langle x, y\rangle|$ and choose a moderate size of $K = O(1/\alpha)$.

4.1 Some intuition of RPT

In this section, we discuss the intuition behind (5). As demonstrated in Section 3, a naive permutation test on the residuals is in general not valid in the finite population setting with just exchangeable noises. This is because under the null, ϕ_{naive} performs permutations on the vector $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_0^\top\boldsymbol{\varepsilon}$ instead of $\boldsymbol{\varepsilon}$ itself. Even if $\boldsymbol{\varepsilon}$ is an exchangeable random vector, $\boldsymbol{V}_0^\top\boldsymbol{\varepsilon}$ may no longer be so, which renders the naive test invalid.

To overcome this challenge, we may want to construct a new test that, under H_0, is equivalent to permuting the noise vector $\boldsymbol{\varepsilon}$ directly, instead of the transformed noise $\boldsymbol{V}_0\boldsymbol{\varepsilon}$. Interestingly, this goal can be

[1] If $\boldsymbol{X}$ is full column rank, then $\boldsymbol{V}_0\boldsymbol{V}_0^\top = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $\mathrm{span}(\boldsymbol{V}_0)$ and $\mathrm{span}(\boldsymbol{X})^\perp$ are the same space. Otherwise, $\mathrm{span}(\boldsymbol{V}_0)$ is a subspace of $\mathrm{span}(\boldsymbol{X})^\perp$.

achieved based on a special transformation of the vector $\boldsymbol{V}_0^\top \boldsymbol{\varepsilon}$. Specifically, given a permutation matrix $\boldsymbol{P}_k$, recall that $\boldsymbol{V}_k = \boldsymbol{P}_k \boldsymbol{V}_0$, we may use the transformation that under H_0,

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_0^\top \boldsymbol{\varepsilon} = \boldsymbol{V}_0^\top \boldsymbol{P}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon} = \boldsymbol{V}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}. \tag{7}$$

In light of this transformation, we further have that under H_0, $\boldsymbol{V}_k \hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_k \boldsymbol{V}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon} = \mathrm{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k \boldsymbol{\varepsilon})$, i.e., a projection of the noise vector $\boldsymbol{P}_k \boldsymbol{\varepsilon}$ onto the space $\mathrm{span}(\boldsymbol{V}_k)$, and equivalently, $\boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}} = \mathrm{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$. However, this is still not enough, as $\mathrm{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$ and $\mathrm{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k \boldsymbol{\varepsilon})$ corresponds to the projections of the vectors $\boldsymbol{\varepsilon}$ and $\boldsymbol{P}_k \boldsymbol{\varepsilon}$ onto different subspaces, which are not directly comparable. This means that we need to further propose a more refined strategy to project $\boldsymbol{\varepsilon}$ and $\boldsymbol{P}_k \boldsymbol{\varepsilon}$ onto some *same space* for a fair comparison.

Now recall that we already have $\mathrm{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$ and $\mathrm{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k \boldsymbol{\varepsilon})$, an ideal choice of such space would then be $\mathrm{span}(\tilde{\boldsymbol{V}}_k)$, i.e., the intersection of $\mathrm{span}(\boldsymbol{V}_0)$ and $\mathrm{span}(\boldsymbol{V}_k)$. Specifically, using that $\tilde{\boldsymbol{V}}_k$ spans a subspace of $\mathrm{span}(\boldsymbol{V}_k)$, it is straightforward that $\tilde{\boldsymbol{V}}_k^\top = \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \boldsymbol{V}_k^\top$. From this and (7), we have that under H_0,

$$\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \hat{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \boldsymbol{V}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon} = \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}$$

and equivalently $\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}$ since $\tilde{\boldsymbol{V}}_k$ spans a subspace of $\mathrm{span}(\boldsymbol{V}_0)$ as well.

From the above analysis, we further have that under H_0,

$$T\left(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}\right) = T\left(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}\right) \quad \text{and} \quad T\left(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \hat{\boldsymbol{\varepsilon}}\right) = T\left(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}\right).$$

This allows us to control ϕ as that

$$\begin{aligned}
\phi &= \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}} \in \{\tilde{\boldsymbol{V}}_1, \ldots, \tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top \boldsymbol{\varepsilon}\right) \le T\left(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}\right)\right\}\right) \\
&\ge \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}} \in \{\tilde{\boldsymbol{V}}_1, \ldots, \tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top \boldsymbol{\varepsilon}\right) \le \min_{\tilde{\boldsymbol{V}} \in \{\tilde{\boldsymbol{V}}_1, \ldots, \tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top \boldsymbol{V}_0 \hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}\right)\right\}\right) \\
&= \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{g(\boldsymbol{\varepsilon}) \le g(\boldsymbol{P}_k \boldsymbol{\varepsilon})\right\}\right)
\end{aligned}$$

for some function $g(\cdot)$ that depends only on $(\boldsymbol{X}, \boldsymbol{Z}, \mathcal{P}_K)$. Since here we consider a deterministic $\mathcal{P}_K$, $g(\cdot)$ is also a deterministic function.

Now our only remaining job is to prove that the p-value displayed at the end of the above inequality is valid. The following lemma shows that once we construct $\mathcal{P}_K$ such that Assumption 2 holds, ϕ is a valid p-value:

Lemma 3. *Suppose $\boldsymbol{\varepsilon}$ satisfies Assumption 1 and let $\{\boldsymbol{P}_0 = \boldsymbol{I}, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}$ be a fixed set of permutation matrices satisfying Assumption 2. Then for any function $g : \mathbb{R}^n \to \mathbb{R}$, we have that*

$$\mathbb{P}\left\{\frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\{g(\boldsymbol{\varepsilon}) \le g(\boldsymbol{P_k}\boldsymbol{\varepsilon})\}\right) \le \alpha\right\} \le \frac{\lfloor \alpha(K+1) \rfloor}{K+1} \le \alpha.$$

5 Analysis of statistical power

This section provides power analysis of RPT under mild moment assumptions of noises ε_i and e_i's where, e.g., the second order moments are not necessarily finite. For simplicity of exposition, throughout this section we assume without loss of generality that n is a multiple of $|\mathcal{P}_K| = K + 1$, where K is a fixed constant that is chosen such that $K \geq 1/\alpha$ for the prespecified Type-I error α. The scenario where n is not divisible by $K + 1$ can be handled by randomly discarding a subset of data of size at most K to make n divisible. We will focus on the version of RPT defined in (6) with $T(x, y) = |\langle x, y\rangle|$. Moreover, we are primarily interested in the dependence of the power of RPT on the tail heaviness of the noise distributions. To this end, we make the following assumption on the model:

Assumption 3. ε_i's are i.i.d. from some distribution $\mathbb{P}_\varepsilon$ with mean 0, $\boldsymbol{Z}$ follows the model in (3) with e_i's i.i.d. from some distribution $\mathbb{P}_e$ with mean 0. $\boldsymbol{\varepsilon}$ is independent from $\boldsymbol{e}$.

In addition, we make following assumption on the permutation matrices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K$.

Assumption 4. For any $k = 1, \ldots, K$, $|\mathrm{tr}[\boldsymbol{V}_0 \boldsymbol{V}_0^\top \boldsymbol{P}_k]| < \sqrt{2p}K$ and $\mathrm{tr}[\boldsymbol{P}_k] = 0$.

Notice that when the covariate matrix $\boldsymbol{X}$ is of full column rank p, Assumption 4 is equivalent to that $|\mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{P}_k]| < \sqrt{2p}K$.

In Theorem 3, we showcase the pointwise signal detection rate of ϕ given any fixed $\mathbb{P}_\varepsilon$ and $\mathbb{P}_e$. Moreover, we just require $\mathbb{P}_\varepsilon$ to have bounded $(1 + t)$-th order moment.

Theorem 3. *Fix $K \in \mathbb{N}$. Assume that $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ satisfy Assumption 3 and*

$$0 < \mathbb{E}[|e_1|^2] < \infty \quad \textit{and} \quad 0 < \mathbb{E}[|\varepsilon_1|^{1+t}] < \infty$$

for some constant $t \in [0, 1]$. Assume $\mathcal{P}_K$ satisfies Assumption 4. In the asymptotic regime where b and p vary with n in a way such that $n > (3 + m)p$ for some constant $m > 0$ and

$$|b| = \Omega(n^{-\frac{t}{1+t}}) \textit{ if } t < 1 \quad \textit{or} \quad |b| = \omega(n^{-\frac{1}{2}}) \textit{ if } t = 1, \tag{8}$$

we have $\lim_{n\to\infty} \mathbb{P}\left(\phi > \frac{1}{K+1}\right) = 0$.

Notice that here we need to assume without loss of generality that $\mathbb{E}[e_i^2] > 0$ and $\mathbb{E}[|\varepsilon_1|^{1+t}] > 0$ to ensure that both two random variables are *not* almost surely equal to zero. Otherwise, ϕ is almost surely equal to 1, and cannot have any statistical power with any size of b. Theorem 3 shows that under certain assumptions on the $\mathcal{P}_K$, RPT has power to reject the alternative classes even with heavy-tailed noises. Moreover, our analysis is high-dimensional and allows the number of covariates to be as large as $n/3$. Remarkably, the statistical power guarantee in Theorem 3 does not require the ε_i's to have a bounded second order moment. This distinguishes us from the class of empirical correlation based approaches, such as debiased lasso or OLS fit based tests, which requires ε_i's to have at least a bounded second order moment or even stronger conditions such as sub-Gaussianity to have statistical power.

As we will see in Section 5.1, Assumption 4 is a mild condition that can be checked in practice. However, an inspection of the proof of Theorem 3 reveals that, even if Assumption 4 does not hold for $\mathcal{P}_K$, RPT is still asymptotically powerful under the same signal strength condition (8) and a slightly stronger requirement on the number of covariates. Specifically, we require that $n > (4 + m)p$ for some constant $m > 0$ that does not depend on n.

In the following theorem, we show that when $p/n \to 0$, we can further relax e_i's finite second order moment condition to a finite first order moment condition.

Algorithm 1: Permutation set construction

Input: The number of permutation matrices K, design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, the maximal number of loops T

repeat

  Generate an independent random permutation π of indices $\{1,\ldots,n\}$

  for $k = 1,\ldots,K$ **do**

    Construct a permutation matrix $\boldsymbol{P}_k \in \mathcal{P}$ by setting its (u,v)-th entries as 1 if and only if

$$\left\lceil \frac{\pi(u)}{K+1} \right\rceil = \left\lceil \frac{\pi(v)}{K+1} \right\rceil \quad \text{and} \quad \pi(u) - \pi(v) \in \{k, k-(K+1)\}.$$

  end

until *(i)* $|\mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{X}^\top\boldsymbol{P}_k)| \le \sqrt{2}Kp^{1/2}$ *for all* $k = 1,\ldots,K$ *or (ii) the number of iterations has reached its limit* T

Output: Set of permutation matrices $\mathcal{P}_K := \{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_K\}$ satisfying the criteria (i). When none of the $\mathcal{P}_K$'s comply, report the $\mathcal{P}_K$ with the smallest $\sum_{k=1}^K |\mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{X}^\top\boldsymbol{P}_k)|$.

Theorem 4. *Fix* $K \in \mathbb{N}$. *Assume that* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{e}$ *satisfy Assumption 3 and*

$$0 < \mathbb{E}[|e_1|] < \infty \quad \textit{and} \quad 0 < \mathbb{E}[|\varepsilon_1|^{1+t}] < \infty$$

for some constant $t \in [0,1]$. *In the asymptotic regime where* b *and* p *vary with* n *in a way such that* $p/n \to 0$ *and* b *satisfies (8),* $\lim_{n\to\infty} \mathbb{P}\left(\phi > \frac{1}{K+1}\right) = 0$.

The statistical power guarantee in Theorem 3 requires the set of permutations to follow Assumption 4, whlist the finite-population validity requires instead Assumption 2. Then an important question is, how to effectively construct a $\mathcal{P}_K$ that satisfies both assumptions. In Section 5.1, we provide an algorithm to answer this question. In order to prove Theorems 3 and 4, we are faced with two questions, the first is that we do not have any assumption on $\boldsymbol{X}$, so that $\tilde{V}_j$ can follow arbitrary pattern; the second is the heavy tails of e_i's and ε_i's. We defer the proof of the two theorems to the Supplementary Material. To help the readers understand the intuitions of the proof, we provide a proof sketch of the two theorems in Sections 5.2 and 5.3.

5.1 An algorithm for construction of permutation set

As demonstrated in Theorems 2 and 3, to successfully perform a test that is valid under the null and has sufficient statistical power to get the rate in (8) in high-dimensional models, one needs a set of permutations satisfying both Assumptions 2 and 4. As demonstrated in Proposition 1 below, such permutation set always exist, so that we can at least apply a brute-forth algorithm to find a desired set. To improve computational efficiency, we further develop a randomized algorithm that can discover the desired permutation set with high probability (Algorithm 1).

In fact, finding a permutation set that just satisfies Assumption 2 is trivial, as we can easily divide the indices $\{1,\ldots,n\}$ into $K+1$ subsets and change the order of subsets for permutation set construction. However, such construct cannot create sufficient randomness to satisfy Assumption 4 for an arbitrary $\boldsymbol{X}$. To generate extra randomness, we need to first shuffle the permutation of all indices and then construct $\mathcal{P}_K$ on the randomized data. In Proposition 1, we show that after doing T-th round of such random shuffling, Algorithm 1 is able to deliver a $\mathcal{P}_K$ satisfying the desired properties with probability at least $1 - \frac{1}{K^T}$.

Algorithm 2: Residual Permutation Test (RPT)

Input: design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, additional covariate of of interest $\boldsymbol{Z} \in \mathbb{R}^n$, response vector $\boldsymbol{Y} \in \mathbb{R}^n$, number of permutations $K \in \mathbb{N}$, maximal number of iterations $T \in \mathbb{N}$.

Apply Algorithm 1 with inputs K, T and $\boldsymbol{X}$ to generate K permutation matrices $\{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}$.

Find an orthonormal matrix $\boldsymbol{V}_0 \in \mathbb{R}^{n\times(n-p)}$ such that $\boldsymbol{V}_0^\top \boldsymbol{X} = 0$.

for $k = 1, \ldots, K$ **do**

  Set $\boldsymbol{V}_k := \boldsymbol{P}_k \boldsymbol{V}_0$.

  Find an orthonormal matrix $\tilde{\boldsymbol{V}}_k \in \mathbb{R}^{n\times(n-2p)}$ such that $\tilde{\boldsymbol{V}}_k^\top (\boldsymbol{X}, \boldsymbol{P}_k \boldsymbol{X}) = 0$.

  Compute

$$a_k := \left|\langle \tilde{\boldsymbol{V}}_k^\top \boldsymbol{Z}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{Y}\rangle\right| \quad \text{and} \quad b_k := \left|\langle \tilde{\boldsymbol{V}}_k^\top \boldsymbol{Z}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{Y}\rangle\right|,$$

  where $< \cdot, \cdot >$ denotes the inner product.

end

Output: p-value $\phi := \frac{1}{K+1}(1 + \sum_{k=1}^K \mathbb{1}\{\min_{1\le j\le K} a_j \le b_k\})$

Proposition 1. *Given K, T and assume that $n \ge 2$, we have that there exists a $\mathcal{P}_K$ satisfying Assumptions 2 and 4. Moreover, with probability at least $1 - \frac{1}{K^T}$, Algorithm 1 returns a $\mathcal{P}_K$ that satisfies Assumptions 2 and 4.*

Notice that throughout this article, we assume that the alternative class is in the form $\boldsymbol{Y} = \boldsymbol{X}\beta + b\boldsymbol{Z} + \boldsymbol{\varepsilon}$ for some $b \neq 0$, whence we invoke Assumption 4 to increase its statistical power. When the alternative class follows other forms, such as $\boldsymbol{Y} = \boldsymbol{X}\beta + f(\boldsymbol{Z}) + \boldsymbol{\varepsilon}$ with some nonlinear function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$, one may not necessarily need Assumption 4 anymore. Instead, one may need other assumptions on $\mathcal{P}_K$ to adapt to the nonlinear function $f(\cdot)$. In light of Algorithm 1 and our theoretical statements, we summarize an implementation of RPT in Algorithm 2.

5.2 Proof sketch of Theorem 3

As K is finite, we mainly need to prove that for any fixed $\boldsymbol{P}_j, \boldsymbol{P}_k \in \mathcal{P}_K$, with probability converging to 1, $|\hat{\boldsymbol{e}}^\top \boldsymbol{V}_0^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{V}_j \hat{\boldsymbol{\varepsilon}}| > |\hat{\boldsymbol{e}}^\top \boldsymbol{V}_0^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \hat{\boldsymbol{\varepsilon}}|$. To achieve this goal, we need to prove that

$$\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} = o_{\mathbb{P}}(1) \tag{9}$$

(i.e., that the empirical correlation between the projection of $\boldsymbol{e}$ and $\boldsymbol{\varepsilon}$ onto the space spanned by $\tilde{\boldsymbol{V}}_j$ is negligible with high probability) and that with high probability,

$$\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} \gtrsim 1 \quad \text{and} \quad \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} \gtrsim 1. \tag{10}$$

There are two main challenges to establish (9), namely the lack of structural assumptions of arbitrary fixed design matrix $\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top$ and the heavy tailed noise $\boldsymbol{\varepsilon}$. To address the first challenge, we introduce a random permutation matrix $\boldsymbol{P} \sim \mathrm{Unif}(\mathcal{P})$ independent from $\boldsymbol{e}$ and $\boldsymbol{\varepsilon}$. From the exchangeability of $\boldsymbol{\varepsilon}$, we have $\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon} \overset{d}{=} \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{P}\boldsymbol{\varepsilon}$, so that we can take expectation over $\boldsymbol{P}$ to "smooth out" the matrix $\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top$ for a tighter control of (9).

We now focus on the second challenge. To illustrate intuitions, throughout this section we assume that the ε_i's are one-sided heavy tail, i.e., there exists a constant $B > 0$ such that almost surely, $\varepsilon_i \geq -B$, and defer its extension to the two-sided heavy tail data to the Supplementary Material. To deal with such one-sided heavy tail noise, we truncate ε_i to obtain $f_i := \varepsilon_i \mathbb{1}(|\varepsilon_i| \leq Bi^{\frac{1}{1+t}})$. As using such truncation the expectation of f_i may not necessarily be zero, we further construct f_i' as a mean-zero random variable such that $f_i' = f_i$ almost surely under the event $f_i \geq 0$ and $f_i \leq f_i' \leq 0$ almost surely under the event $f_i < 0$. As we will demonstrate in the Supplementary Material, such construction is always possible. Also, we define $\boldsymbol{f} := (f_1, \ldots, f_n)$ and $\boldsymbol{f}' := (f_1', \ldots, f_n')$.

Such construction allows us to solve (9) by tackling the terms $\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{P}\boldsymbol{f}'$, $\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{P}(\boldsymbol{f} - \boldsymbol{f}')$ and $\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{P}(\boldsymbol{\varepsilon} - \boldsymbol{f})$ separately. The third term can be controlled via applying the Borel–Cantelli Lemma [Durrett, 2019]; and the second term can be controlled via an analgous argument as the first term. Thus, we mainly discuss the first term $\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{P}\boldsymbol{f}'$.

Now that all the f_i' are bounded, we may apply Chebyshev's inequality for this term. As will be demonstrated in the supplement, to apply Chebyshev's inequality, we just need to prove that given any fixed $\boldsymbol{e} \in \mathbb{R}^n$,

$$\mathbb{E}[(\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{P}\boldsymbol{f}')^2] = o(n^{\frac{1-t}{1+t}} \cdot \|\boldsymbol{e}\|_2^2). \tag{11}$$

To prove (11), exploiting the randomness of $\boldsymbol{P}$, we can prove that $\mathbb{E}[(\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{P}\boldsymbol{f}')^2] \leq \|\boldsymbol{e}\|_2^2 \mathbb{E}[\|\boldsymbol{f}'\|_2^2]/n$ (this is how $\boldsymbol{P}$ "smoothes" $\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top$). Now to control $\mathbb{E}[\|\boldsymbol{f}'\|_2^2]/n$, we further let a_n be a sequence of integers such that $a_n \to \infty$ but $a_n/n \to 0$. Then we may decompose $\mathbb{E}[\|\boldsymbol{f}'\|_2^2]/n$ as that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{f}'\|_2^2]/n =& \frac{1}{n}\sum_{i=1}^{a_n} \mathbb{E}[(f_i')^2] + \frac{1}{n}\sum_{i=a_n+1}^{n} \mathbb{E}[(f_i')^2] \leq \mathbb{E}[\varepsilon_1^2 \mathbb{1}(\varepsilon_1 \leq Ba_n^{\frac{1}{1+t}})] \\ &+ \frac{1}{n}\sum_{i=a_n+1}^{n} \mathbb{E}[\varepsilon_i^2 \mathbb{1}(Ba_n^{\frac{1}{1+t}} < \varepsilon_1 \leq Bi^{\frac{1}{1+t}})],\end{aligned} \tag{12}$$

where the first term on the right hand side of (12) can be bounded as

$$\mathbb{E}[\varepsilon_1^2 \mathbb{1}(\varepsilon_1 \leq Ba_n^{\frac{1}{1+t}})] = \mathbb{E}[|\varepsilon_1|^{1+t}|\varepsilon_1|^{1-t} \mathbb{1}(\varepsilon_1 \leq Ba_n^{\frac{1}{1+t}})] \lesssim \mathbb{E}[|\varepsilon_1|^{1+t}] a_n^{\frac{1-t}{1+t}} = o(n^{\frac{1-t}{1+t}}).$$

Using an analogous argument, we may control the second term of (12) as well. Putting together, we get the desired bound in (9).

Thanks to the bounded second order moment of e_i's, the analysis of (10) is simpler. Specially, by using a weak law of large number to control the weighted sum of e_i^2's [Van Thanh, 2006] and a Chebyshev's inequality to control the sum of cross term $e_i e_j$'s, we can have that with probability converging to 1,

$$\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} \gtrsim \frac{n - 3p - \mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_k]}{n}.$$

Using that $\boldsymbol{P}_k$ satisfies Assumption 4, we easily obtain the desired result.

5.3 Proof sketch of Theorem 4

Due to the heavy-tailedness of $\boldsymbol{e}$, $\|\boldsymbol{e}\|_2^2/n$ is not statistically convergent anymore. Hence, our goal now becomes proving that

$$\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2} = o_{\mathbb{P}}(1) \tag{13}$$

and

$$\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{\|\boldsymbol{e}\|_2^2} \gtrsim 1 \quad \text{and} \quad \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{\|\boldsymbol{e}\|_2^2} \gtrsim 1. \tag{14}$$

We note that (13) can be controlled using a similar argument as (9). Therefore, in the rest of this proof sketch we focus on (14). Without loss of generality we focus on the first inequality of (14). In the regime $p \ll n$, the matrices $\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top$ and $\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top$ are expected to be not too faraway from the identity matrix. This encourages us to prove (14) via (i) bounding the differences between the quantity in (14) and $\frac{\boldsymbol{e}^\top \boldsymbol{e} - \boldsymbol{e}^\top \boldsymbol{P}_k \boldsymbol{e}}{\|\boldsymbol{e}\|_2^2}$; and (ii) proving that $\boldsymbol{e}^\top \boldsymbol{P}_k \boldsymbol{e} = o_{\mathbb{P}}(\|\boldsymbol{e}\|_2^2)$. To tackle Step (i), we use that

$$\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e} \geq \boldsymbol{e}^\top \boldsymbol{e} - \boldsymbol{e}^\top \boldsymbol{P}_k \boldsymbol{e} + \boldsymbol{e}^\top \boldsymbol{M} \boldsymbol{e}$$

for some positive semidefinite matrix $\boldsymbol{M}$ depending on $\boldsymbol{P}_k$ such that $\mathrm{tr}[\boldsymbol{M}] \leq 2p$. Now it remains to control $\boldsymbol{e}^\top \boldsymbol{M} \boldsymbol{e}$, which seems impossible as entries of $\boldsymbol{e}$ do not even have finite second moment. Fortunately, if we further restrict $\mathbb{P}_e$ to follow a *symmetric distribution*, we may use the following lemma to prove $\boldsymbol{e}^\top \boldsymbol{M} \boldsymbol{e} = o_{\mathbb{P}}(\|\boldsymbol{e}\|_2^2)$ knowing that $\mathrm{tr}[\boldsymbol{M}]/n \to 0$:

Lemma 4. *Consider $\mathbb{P}_w$ as a distribution that is symmetric around zero and $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ as a positive semi-definite matrix. Let $\boldsymbol{w} := (w_1, \ldots, w_n)^\top$ be n i.i.d. realizations from $\mathbb{P}_w$. Then we have that for any $\delta > 0$,*

$$\mathbb{P}\left(\boldsymbol{w}^\top \boldsymbol{U} \boldsymbol{w} > \delta \|\boldsymbol{w}\|_2^2\right) \leq \frac{\mathrm{tr}[\boldsymbol{U}]}{\delta n}.$$

We now discuss Step (ii). We write σ_k for the permutation of $\{1, \ldots, p\}$ corresponding to $\boldsymbol{P}_k$. Observe that $\boldsymbol{e}^\top \boldsymbol{P}_k \boldsymbol{e} = \sum_{i=1}^n e_i e_{\sigma_k(i)}$ have dependent summands. The following combinatorial lemma allows us to circumvent this difficulty by partitioning the summands into three similar sized subsets so that summands within each subset are i.i.d. and can be controlled using standard concentration inequalities.

Lemma 5. *Consider a permutation σ of $\{1, \ldots, n\}$ such that for any $i \in \{1, \ldots, n\}$, $\sigma(i) \neq i$. Then there exists a partition U_1, U_2, U_3 of the set $\{1, \ldots, n\}$ such that*

$$\forall j \in \{1, 2, 3\}, \quad |U_j| \in \left[\frac{n}{4} - 1, \frac{n}{2} + 1\right] \ \& \ |U_j \cap \sigma(U_j)| = 0.$$

The above analysis illustrates how we prove (14) with symmetric $\mathbb{P}_e$. To generalize to asymmetric case, we can simply let $\boldsymbol{e}'$ be an independent replicate of $\boldsymbol{e}$. Apparently, $\boldsymbol{e} - \boldsymbol{e}'$ is a mean zeroed random vector where all the indices are symmetric around zero. Thus allows us to use the previous arguments find a control of $(\boldsymbol{e} - \boldsymbol{e}')^\top \left(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k\right) (\boldsymbol{e} - \boldsymbol{e}')$. Then we can control $\boldsymbol{e}^\top \left(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k\right) \boldsymbol{e}$ using the independence between $\boldsymbol{e}$ and $\boldsymbol{e}'$ and the fact that they have the same distribution.

6 Minimax optimality of coefficient tests

In this section, we investigate the minimax optimality of RPT by deriving the statistical efficiency limit of coefficient tests with heavy-tailed noises. Without loss of generality, we denote $\mathcal{D}_t$ as the class of distributions with t-th order moment bounded between $[1, 2]$, i.e., for some $t > 0$ and some random variable ξ with distribution $\mathbb{P}_\xi$,

$$\mathbb{P}_\xi \in \mathcal{D}_t \quad \text{iff} \quad \mathbb{E}[\xi] = 0 \text{ and } 1 \leq \mathbb{E}[|\xi|^t] \leq 2.$$

Notice that in the above definition, the thresholds 1 and 2 are chosen for notational simplicity, in fact, the general conclusions in this section still hold for $\eta_1 \leq \mathbb{E}[|\xi|^t] \leq \eta_2$ with arbitrary $\eta_1, \eta_2 > 0$. We further let $\tilde{\mathcal{D}}$ denote the class of distributions such that

$$\mathbb{P}_\xi \in \tilde{\mathcal{D}} \quad \text{iff} \quad \mathbb{P}\left(|\xi| > \frac{1}{2}\right) > \frac{1}{2}.$$

With a slight abuse of notation, given $b_0 \in \mathbb{R}$, we write $\mathbb{P}_{b_0}$ as a distribution of $(\boldsymbol{Y}, \boldsymbol{Z})$ such that the b in (1) is equal to b_0. Note that we have suppressed the dependence of $\mathbb{P}_{b_0}$ on $\boldsymbol{X}, \beta, \beta^Z, \mathbb{P}_\varepsilon$ and $\mathbb{P}_e$ for notational simplicity. In particular, $\mathbb{P}_0$ corresponds to the null hypothesis.

From above, we define the minimax testing risk indexed by $t, \boldsymbol{X}$ as

$$\mathcal{R}_{t,\boldsymbol{X}}(\tau) := \inf_{\varphi \in \Phi} \left\{ \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_t} \sup_{\mathbb{P}_e \in \mathcal{D}_1 \cap \tilde{\mathcal{D}}} \sup_{\beta, \beta^Z \in \mathbb{R}^p} \mathbb{P}_0(\varphi = 1) + \sup_{|b| \geq \tau} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_t} \sup_{\mathbb{P}_e \in \mathcal{D}_1 \cap \tilde{\mathcal{D}}} \sup_{\beta, \beta^Z \in \mathbb{R}^p} \mathbb{P}_b(\varphi = 0) \right\}.$$

Here Φ corresponds to the class of measurable functions of data $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ taking value in $\{0, 1\}$. We first establish the following finite-population minimax lower bound for testing $H_0 : b = 0$ against $H_1 : b \neq 0$ in the presence of heavy-tailed noises.

Theorem 5. *Let $t \in [0, 1]$ be given and assume that $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ satisfy Assumption 3. For any $\eta \in (0, 1)$, there exists a constant $c_\eta > 0$ depending only on η such that for any fixed design $\boldsymbol{X}$,*

$$\mathcal{R}_{1+t,\boldsymbol{X}}\left(c_\eta n^{-\frac{t}{1+t}}\right) \geq 1 - \eta.$$

Theorem 5 shows that when entries of $\boldsymbol{\varepsilon}$ have finite $(1 + t)$-th moment, the minimax separation in b for testing H_0 against H_1 is at least of order $n^{-\frac{t}{1+t}}$, which matches the upper bound in Theorem 3. This indicates that the rate $n^{-\frac{t}{1+t}}$ may be a tight lower bound, and that our constructed test may be an optimal test. Nevertheless, Theorems 3 and 4 are pointwise convergence results, where both $\mathbb{P}_\xi$ and $\mathbb{P}_e$ are considered as fixed and does not depend on n, p. To match the lower bound in Theorem 5, we further provide a power control of RPT uniformly over classes of noise distributions of $\mathbb{P}_\varepsilon$ and $\mathbb{P}_e$. Just as in Section 5, we assume without loss of generality that n is divisible by $K + 1$.

Theorem 6. *Fix $K \in \mathbb{N}$. Suppose that $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ satisfy Assumption 3 and that $\mathcal{P}_K$ satisfies Assumption 4. In an asymptotic regime where b and p vary with n in a way such that $n > (3 + m)p$ for some constant $m > 0$ and $|b| = \Omega(n^{-\frac{t}{1+t}+\delta})$ for some constants $t \in (0, 1]$ and $\delta > 0$, we have for any constant $\nu > 0$ that,*

$$\lim_{n \to \infty} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu} \cap \tilde{\mathcal{D}}} \mathbb{P}\left(\phi > \frac{1}{K+1}\right) = 0. \tag{15}$$

If we drop Assumption 4 and instead assume $p/n \to 0$, then we have for any constant $\nu > 0$ that,

$$\lim_{n \to \infty} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \sup_{\mathbb{P}_e \in \mathcal{D}_{1+\nu} \cap \tilde{\mathcal{D}}} \mathbb{P}\left(\phi > \frac{1}{K+1}\right) = 0. \tag{16}$$

In Theorem 6, the separation rate is slightly worse than (8) by a factor of n^δ, where δ can be any positive constant. Also, it is slightly worse than the lower bound in Theorem 5. This shows that the separation rate $n^{-\frac{t}{1+t}}$ is a nearly-optimal rate of coefficient testing in the minimax sense. At the same time, it also shows that our residual permutation test is a nearly-optimal hypothesis test in the minimax sense.

7 Numerical studies

7.1 Experimental setups

In this section, we evaluate the performance of RPT, together with several competitors, in the following synthetic datasets. The observations $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{n\times p}\times\mathbb{R}^n\times\mathbb{R}^n$ are generated according to the models (1) and (3) where

- $\boldsymbol{X}$ is generated according to $\boldsymbol{X} := \boldsymbol{W}\Sigma^{1/2}$, where $\Sigma = (2^{-|j-k|})_{j,k\in[p]}$ is the Toeplitz matrix and $\boldsymbol{W}$ is an $n \times p$ dimensional matrix with i.i.d. entries from either $\mathcal{N}(0, 1)$ or t_1 distribution;
- β and β^Z are p-dimensional vectors with the first 10 components equal to $1/\sqrt{10}$ and the rest components equal to 0;
- $\boldsymbol{e}$ and $\boldsymbol{\varepsilon}$ have independent and identically distributed components drawn from $\mathcal{N}(0, 1), t_1$ or t_2.

We vary $n \in \{300, 600\}$, $p \in \{100, 200\}$ and b in different simulation experiments.

In practice, we find that the p-value calculated by Algorithm 2 is slightly on the conservative side. Hence, in addition to the test with p-value constructed by Algorithm 2, we also study a variant in our numerical experiments, where the p-value is computed as $\frac{1}{K+1}(1 + \sum_{k=1}^{K} \mathbb{1}\{a_k \le b_k\})$ instead (we call this variant as RPT_{EM}, where "EM" stands for empirical). To benchmark the performance of RPT and RPT_{EM}, we also look at the naive residual permutation test in (4). Other tests used for comparison include the ANOVA test described in the introduction and a debiased Lasso based test (dbLasso) using the implementation of Javanmard and Montanari [2018] in our numerical studies.

We note that RPT relies on tuning parameters K and T. For a test to have a size of α, we need to have $K+1$ at least $1/\alpha$. We suggest using $K+1 = \lceil 1/\alpha \rceil$ in practice, though empirical simulation results suggest that our method is robust to the choice of K. We also set $T = 1$ to boost the computational efficiency of Algorithm 1.

7.2 Numeric analysis of validity under the null

We start by analysing the validity of various tests under the null described in Section 7.1. We estimate the size of RPT, RPT_{EM} and dbLasso at nominal levels of 0.01 and 0.005 for $(n, p) \in \{(300, 100), (600, 100), (600, 200)\}$. The results are summarised in Table 2. Notice that since the p-values of both ANOVA and the naive test are invariant with respect to the choices of β, β^Z and Σ, the results in Table 1 are directly comparable to the ones in Table 2. Therefore, we do not repeat the simulations of the two tests here. Besides, notice that since the test of Lei and Bickel [2021] is not well-defined for $p \ge n/(\lceil 1/\alpha \rceil + 1)$, we did not include CPT in our numerical experiment.

From Table 2, we see that all tests considered have valid size guarantees when the noise has Gaussian components, regardless of the choices of $\boldsymbol{X}$. However, in the presence of heavy-tailed noises, dbLasso reports empirical sizes much larger than the nominal levels, which is just like ANOVA and the naive RPT in Table 1. Interestingly, when $\boldsymbol{X}$ follows the t_1 design, the size of dbLasso is even above 70%, which is much larger than the size of ANOVA and the naive RPT displayed in Table 1. This indicates that dbLasso is more sensitive to heavy-tailed design matrices than the other competing methods.

On the other hand, RPT exhibits valid size controls in all settings, which is consistent with our theoretical findings. What's more interesting, the size of RPT_{EM} is also valid across all the simulation settings, even with heavy-tailed noises and heavy-tailed design. In Section 7.3, we further study the empirical power of RPT and RPT_{EM}.

				RPT_{EM}		RPT		dbLasso	
n	p	$\boldsymbol{X}$	noise	0.01	0.005	0.01	0.005	0.01	0.005
300	100	Gauss.	Gauss.	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0.31_{(0.02)}$	$0.15_{(0.01)}$
300	100	Gauss.	t_1	$0.51_{(0.02)}$	$0.12_{(0.01)}$	$0.24_{(0.02)}$	$0_{(0)}$	$1.83_{(0.04)}$	$1.56_{(0.04)}$
300	100	Gauss.	t_2	$0.14_{(0.01)}$	$0.02_{(0)}$	$0.04_{(0.01)}$	$0_{(0)}$	$1.20_{(0.03)}$	$0.84_{(0.03)}$
300	100	t_1	Gauss.	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$99.71_{(0.02)}$	$99.71_{(0.02)}$
300	100	t_1	t_1	$0.01_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$75.98_{(0.14)}$	$75.44_{(0.14)}$
300	100	t_1	t_2	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$93.20_{(0.08)}$	$93.19_{(0.08)}$
600	100	Gauss.	Gauss.	$0.21_{(0.01)}$	$0.07_{(0.01)}$	$0.01_{(0)}$	$0_{(0)}$	$0.62_{(0.02)}$	$0.28_{(0.02)}$
600	100	Gauss.	t_1	$0.73_{(0.03)}$	$0.43_{(0.02)}$	$0.48_{(0.02)}$	$0.28_{(0.02)}$	$1.59_{(0.04)}$	$1.41_{(0.04)}$
600	100	Gauss.	t_2	$0.61_{(0.02)}$	$0.33_{(0.02)}$	$0.20_{(0.01)}$	$0.12_{(0.01)}$	$1.16_{(0.03)}$	$0.85_{(0.03)}$
600	100	t_1	Gauss.	$0.23_{(0.02)}$	$0.07_{(0.01)}$	$0.01_{(0)}$	$0_{(0)}$	$99.93_{(0.01)}$	$99.93_{(0.01)}$
600	100	t_1	t_1	$0.13_{(0.01)}$	$0.03_{(0.01)}$	$0_{(0)}$	$0_{(0)}$	$72.21_{(0.14)}$	$71.78_{(0.14)}$
600	100	t_1	t_2	$0.10_{(0.01)}$	$0.03_{(0.01)}$	$0_{(0)}$	$0_{(0)}$	$92.95_{(0.08)}$	$92.94_{(0.08)}$
600	200	Gauss.	Gauss.	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0.86_{(0.03)}$	$0.43_{(0.02)}$
600	200	Gauss.	t_1	$0.46_{(0.02)}$	$0.34_{(0.02)}$	$0.26_{(0.02)}$	$0.17_{(0.01)}$	$1.51_{(0.04)}$	$1.30_{(0.04)}$
600	200	Gauss.	t_2	$0.12_{(0.01)}$	$0.10_{(0.01)}$	$0.04_{(0.01)}$	$0.03_{(0)}$	$1.26_{(0.04)}$	$0.87_{(0.03)}$
600	200	t_1	Gauss.	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$100.00_{(0)}$	$100.00_{(0)}$
600	200	t_1	t_1	$0.01_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$88.97_{(0.10)}$	$88.55_{(0.10)}$
600	200	t_1	t_2	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$99.39_{(0.02)}$	$99.38_{(0.02)}$

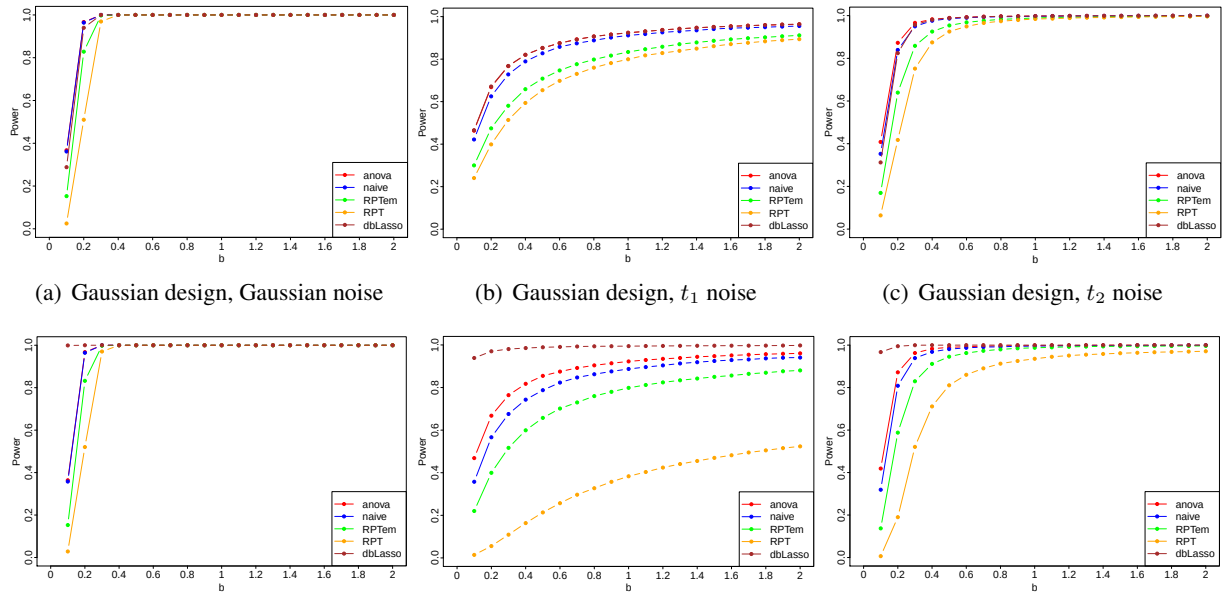
Table 2: Sizes of various tests under the null, estimated over 100000 Monte Carlo repetitions, for various noise distributions at nominal levels of $\alpha = 0.01$ and $\alpha = 0.005$. Data are generated from the model in (1) and (3) with $b = 0$. $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ are generated according to the various distribution types prescribed in the table. Here "Gauss." stands for standard normal distribution. For ease of presentation, the estimated sizes are multiplied by 100 in the table. Standard errors of the estimated size are given in parentheses.

7.3 Numeric analysis of alternative power

In Section 5, we established asymptotic power guarantees of RPT under fixed design and heavy tailed noises. In this section, we validate these theoretical insights via numerical analysis. To benchmark the results, we investigate the power of all tests considered in Section 7.1. We set $n = 600$, $p = 100$ and vary the b in (1) as $b \in \{0.1, 0.2, \ldots, 1.9, 2.0\}$. We analyze the power of all methods with design following Gaussian and t_1 distributions and noises following Gaussian, t_1, and t_2 distributions. The estimated power curves over 10000 repetitions are displayed in Figure 2.

From Figures 2(a)-(c), (d) and (f), we can conclude that in most of the settings, the power of RPT is slightly worse than the competing approaches. The difference is more pronounced when both the design and the noise follow a heavy-tailed distribution (Figure 2(e)). However, bearing in mind the lack of valid size control of ANOVA, naive RPT and debiased Lasso, we would argue that the gap in power between RPT and these competitors is the price to pay for distribution-free finite-population validity in high dimensions. Moreover, we observe that RPT is nevertheless still guaranteed to reject the alternative with high probability given sufficiently large b.

Another interesting phenomenon is that the power of RPT_{EM} is generally stronger than RPT, especially in the setting displayed in Figure 2(e), where both design and noise follow t_1 distribution. This, together with the validity display in Section 7.2, suggests RPT_{EM}, although being lack of theoretical support, can



(a) Gaussian design, Gaussian noise (b) Gaussian design, t_1 noise (c) Gaussian design, t_2 noise

(d) t_1 design, Gaussian noise (e) t_1 design, t_1 noise (f) t_1 design, t_2 noise

Figure 2: Proportion of p-values below $\alpha = 0.01$ over 10000 replicates for $b = 0.1, 0.2, \ldots, 1.9, 2$. Here $\boldsymbol{X}, \boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ are generated according to various distribution types prescribed in the caption of each figure.

serve as a viable alternative of RPT in empirical analysis. We leave the theoretical investigations of RPT_{EM} as future work.

Finally, the power of naive RPT is nearly the same as ANOVA in most of the settings. Sometimes, the two power curves are even indistinguishable. Recall that in Section 3, we have shown that empirically, the size control of the naive RPT under the null is more robust to non-Gaussian noises than ANOVA. In practice, we recommend using the naive RPT for single coefficient tests when $n/2 \leq p < n$.

8 Discussion

In this paper, we propose a new method for high-dimensional fixed design regression coefficient test. RPT is a permutation-based approach that exploits the exchangeability of the noise terms to achieve finite-population validity control. Our approach uses the fact that the empirical residuals of the classical OLS fit is equivalent to the projection of the n-dimensional noise vector onto an $(n-p)$-dimensional subspace to construct a valid test for $p < n/2$ based on multiple subspace projection. At the same time, we provide power analysis of RPT, and derived the signal detection rate of the coefficient b in the presence of heavy-tailed noise vector $\boldsymbol{\varepsilon}$. As a by product, we propose RPT_{EM} and demonstrate its validity and power via numerical experiments. It would be of interest to understand the theoretical properties of RPT_{EM} in future study.

In the higher dimensional regime $n/2 \leq p < n$, we propose the naive RPT, and prove its finite-population validity under spherically invariant distributions, and compare it with ANOVA as well as other competing approaches via numerical experiments. In the meanwhile, we provide a more profound theoretical and empirical analysis of ANOVA test, which is of independent interest for practitioners interested in ANOVA.

Distribution-free inference and test is an important topic in statistics research. In this paper, permutation test facilitates an important basis for construction of finite-population tests hypothesis tests with distribution-free validity. This sheds light on extending permutation tests to solve other distribution-free problems in modern statistics, which we leave as future work. In addition, permutation tests and its related the rank based tests have also been applied in model-free uncertainty quantification of machine learning predictions [Lei et al., 2013, Balasubramanian et al., 2014, Romano et al., 2019]. It would be of interest if the power analysis techniques invented in this paper could be used to understand the efficiency of these approaches in modern machine learning applications.

References

Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

Jelena Bradic, Victor Chernozhukov, Whitney K Newey, and Yinchu Zhu. Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*, 2019.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP Probabilit'es et statistiques*, 48(4):1148–1185, 2012.

Devin Caughey, Allan Dafoe, and Luke Miratrix. Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. *arXiv preprint arXiv:1709.07339*, 2017.

Devin Caughey, Allan Dafoe, Xinran Li, and Luke Miratrix. Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. *arXiv preprint arXiv:2101.09195*, 2021.

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

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

Xavier D'Haultfœuille and Purevdorj Tuvaandorj. A robust permutation test for subvector inference in linear regressions. *arXiv preprint arXiv:2205.06713*, 2022.

David P Doane and Lori E Seward. *Applied statistics in business and economics, 5th*. Mcgraw-Hill, 2016.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905, 2016.

Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239, 2021.

Sir Ronald Aylmer Fisher. *Statistical methods for research workers... revised and enlarged*. Hafner Publishing Company, 1973.

JA Hartigan. Exact confidence intervals in regression problems with independent symmetric errors. *The Annals of Mathematical Statistics*, pages 1992–1998, 1970.

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

Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *arXiv preprint arXiv:2112.11666*, 2021.

Stanley E Lazic. Why we should use simpler models if the data allow this: relevance for anova designs in experimental biology. *BMC physiology*, 8(1):1–7, 2008.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Lihua Lei and Peter J Bickel. An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, 108(2):397–412, 2021.

Po-Ling Loh and Xin Lu Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12(1):1429–1467, 2018.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

Nicolai Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):923–945, 2015.

Marc S Paolella. *Linear models and time-series analysis: regression, ANOVA, ARMA and GARCH*. John Wiley & Sons, 2018.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

Edwin JG Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937a.

Edwin JG Pitman. Significance tests which may be applied to samples from any populations. II. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4(2):225–232, 1937b.

Edwin JG Pitman. Significance tests which may be applied to samples from any populations: III. the analysis of variance test. *Biometrika*, 29(3/4):322–335, 1938.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Rajen D Shah and Peter Bühlmann. Double-estimation-friendly inference for high-dimensional misspecified models. *arXiv preprint arXiv:1909.10828*, 2019.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Le Van Thanh. Mean convergence theorems and weak laws of large numbers for double arrays of random variables. *Journal of Applied Mathematics and Stochastic Analysis*, 2006, 2006.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Lan Wang, Bo Peng, and Runze Li. A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669, 2015.

Lie Wang. The ℓ_1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

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

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

SUPPLEMENT TO "RESIDUAL PERMUTATION TEST FOR HIGH-DIMENSIONAL REGRESSION COEFFICIENT TESTING"

Section A1 provides validity analysis of the ANOVA test, naive RPT and RPT. It includes the proof of the theoretical statements in Sections 3 and 4 and also the discussion of the equivalence between (5) and (6).

Section A2 studies the statistical power of RPT. It includes the proof of the theoretical statements in Section 5.

Section A3 studies the minimax optimality of coefficient test with heavy-tailed noises. It includes proof of the theoretical statements in Section 6.

A1 Theoretical analysis of finite-population validity

A1.1 ANOVA validity

Proof of Lemma 1. Recall that $\mathrm{Proj}_{\boldsymbol{X}} := \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}} := (\boldsymbol{X},\boldsymbol{Z})\{(\boldsymbol{X},\boldsymbol{Z})^\top(\boldsymbol{X},\boldsymbol{Z})\}^{-1}(\boldsymbol{X},\boldsymbol{Z})^\top$. First assume that $\boldsymbol{\varepsilon}$ is spherically symmetric. Since $\boldsymbol{\varepsilon}$ has a spherically symmetric distribution, we can write $\boldsymbol{\varepsilon} = \rho\boldsymbol{\xi}$, such that $\boldsymbol{\xi} \sim \mathrm{Unif}(\mathcal{S}^n)$, i.e., a random vector that is sampled uniformly from the unit sphere with respect to the Haar measure; and that ρ is some random variable taking value in $[0,\infty)$ and is independent from $\boldsymbol{\xi}$. Then, we have almost surely,

$$\phi_{\mathrm{anova}} = \frac{\|(\boldsymbol{I}-\mathrm{Proj}_{\boldsymbol{X}})(\boldsymbol{\varepsilon})\|_2^2 - \|(\boldsymbol{I}-\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}})(\boldsymbol{\varepsilon})\|_2^2}{\|(\boldsymbol{I}-\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}})(\boldsymbol{\varepsilon})\|_2^2/(n-p-1)} = \frac{\|(\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}}-\mathrm{Proj}_{\boldsymbol{X}})(\boldsymbol{\xi})\|_2^2}{\|(\boldsymbol{I}-\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}})(\boldsymbol{\xi})\|_2^2/(n-p-1)}. \quad \text{(A1.1)}$$

Hence, the distribution of ϕ_{anova} does not depend on ρ.

By Cochran's theorem, we know that $\phi_{\mathrm{anova}} \sim F_{1,n-p-1}$ when $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\boldsymbol{I})$, i.e., a multivariate standard normal distribution. Moreover, when $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\boldsymbol{I})$, we have $\boldsymbol{\varepsilon}$ satisfies the above decomposition $\boldsymbol{\varepsilon} = \rho\boldsymbol{\xi}$ for some random variable ρ. Now recall that ϕ_{anova} does not depend on ρ (as shown in (A1.1)), we must have $\phi_{\mathrm{anova}} \sim F_{1,n-p-1}$ for all spherically symmetric $\boldsymbol{\varepsilon}$ as desired.

If instead $(\boldsymbol{X},\boldsymbol{Z})$ is spherically symmetric, let $\boldsymbol{Q}$ be an indepedent random matrix that is sampled uniformly from $\mathbb{O}^{n\times n}$ with respect to the Haar measure, then

$$\phi_{\mathrm{anova}} \overset{\mathrm{d}}{=} \frac{\|(\mathrm{Proj}_{\boldsymbol{Q}\boldsymbol{X},\boldsymbol{Q}\boldsymbol{Z}}-\mathrm{Proj}_{\boldsymbol{Q}\boldsymbol{X}})(\boldsymbol{\varepsilon})\|_2^2}{\|(I_n-\mathrm{Proj}_{\boldsymbol{Q}\boldsymbol{X},\boldsymbol{Q}\boldsymbol{Z}})(\boldsymbol{\varepsilon})\|_2^2/(n-p-1)} = \frac{\|(\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}}-\mathrm{Proj}_{\boldsymbol{X}})(\boldsymbol{Q}^{-1}\boldsymbol{\varepsilon})\|_2^2}{\|(I_n-\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}})(\boldsymbol{Q}^{-1}\boldsymbol{\varepsilon})\|_2^2/(n-p-1)}.$$

Since $\boldsymbol{Q}^{-1}\boldsymbol{\varepsilon}$ has a spherically symmetric distribution, the desired conclusion follows from the first case. □

A1.2 Validity of naive residual permutation test

Proof of Lemma 2. Without loss of generality we just prove the lemma with Condition (a). We first consider the case where $\boldsymbol{\varepsilon}$ follows a spherically symmetric distribution. Then using an analogous analysis as in Lemma 1, we have

$$\begin{aligned}\phi_{\mathrm{naive}} &= \frac{1}{K+1}\left(1+\sum_{k=1}^{K}\mathbb{1}(|\hat{\boldsymbol{e}}^\top\hat{\boldsymbol{\varepsilon}}| \le |\hat{\boldsymbol{e}}^\top\boldsymbol{P}_k\hat{\boldsymbol{\varepsilon}}|)\right)\\ &= \frac{1}{K+1}\left(1+\sum_{k=1}^{K}\mathbb{1}(|\boldsymbol{Z}^\top\boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{\xi}| \le |\boldsymbol{Z}^\top\boldsymbol{V}_0\boldsymbol{P}_k\boldsymbol{V}_0^\top\boldsymbol{\xi}|)\right).\end{aligned}$$

This means that just like ϕ_{anova}, the distribution of ϕ_{naive} does not depend on ρ. Moreover, when $\boldsymbol{\varepsilon}$ follows a multivariate standard normal distribution, $\boldsymbol{V}_0\boldsymbol{\varepsilon}$ is a $n-p$ dimensional multivariate standard normal random vector and thus ϕ_{naive} is a valid p-value. Then using an analogous argument as in the proof of Lemma 1, we have that ϕ_{naive} is a valid p-value for all spherically symmetric noises.

If instead $(\boldsymbol{X}, \boldsymbol{Z})$ is spherically symmetric, again let $\boldsymbol{Q}$ be an independent matrix sampled uniformly from $\mathbb{O}^{n\times n}$, then

$$\begin{aligned}\phi_{\text{naive}} &\overset{\text{d}}{=} \frac{1}{K+1}\left(1+\sum_{k=1}^{K}\mathbb{1}(|(\boldsymbol{QZ})^\top \boldsymbol{QV}_0\boldsymbol{V}_0^\top\boldsymbol{Q}^\top\boldsymbol{\varepsilon}| \le |(\boldsymbol{QZ})^\top \boldsymbol{QV}_0\boldsymbol{P}_k\boldsymbol{V}_0^\top\boldsymbol{Q}^\top\boldsymbol{\varepsilon}|)\right)\\ &= \frac{1}{K+1}\left(1+\sum_{k=1}^{K}\mathbb{1}(|\boldsymbol{Z}^\top\boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{Q}^\top\boldsymbol{\varepsilon}| \le |\boldsymbol{ZV}_0\boldsymbol{P}_k\boldsymbol{V}_0^\top\boldsymbol{Q}^\top\boldsymbol{\varepsilon}|)\right).\end{aligned}$$

Then using an analogous argument, we prove the validity of ϕ_{naive}. □

A1.3 Validity of residual permutation test

We first show that the two definitions of RPT defined in (5) and (6) are equivalent. Since by definition, $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_0^\top\boldsymbol{Y}$, we easily have $\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_0\hat{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{Y} = \tilde{\boldsymbol{V}}_k^\top\boldsymbol{Y}$, where for the last equality we apply Lemma A1. Using an analogous argument, we can prove that $\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_0\hat{\boldsymbol{e}} = \tilde{\boldsymbol{V}}_k^\top\boldsymbol{Z}$. Now for $\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\hat{\boldsymbol{\varepsilon}}$, we apply that

$$\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\hat{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\boldsymbol{V}_0^\top\boldsymbol{Y} = \tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\boldsymbol{V}_0^\top\boldsymbol{P}_k^\top\boldsymbol{P}_k\boldsymbol{Y} = \tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\boldsymbol{V}_k^\top\boldsymbol{P}_k\boldsymbol{Y} = \tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{Y},$$

where for the last equality we apply again Lemma A1. Putting together, we see that the two definitions of ϕ in (5) and (6) are numerically equivalent.

In the rest of this section, our goal is to prove Theorem 2. We start with the following preliminary lemmas. Recall that for any matrix $\boldsymbol{U} \in \mathbb{R}^{n\times q}$ with orthonormal columns and any vector $\boldsymbol{a} \in \mathbb{R}^n$, $\text{Proj}_{\boldsymbol{U}}(\boldsymbol{a}) := \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{a}$.

Lemma A1. *Let $\boldsymbol{U} \in \mathbb{R}^{n\times p_1}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times p_2}$ be two matrices with orthonormal columns spanning subspaces of $\mathbb{R}^n$. Let $\boldsymbol{W} \in \mathbb{R}^{n\times q}$ be a matrix with orthonormal columns spanning a subspace of $\text{span}(\boldsymbol{U}) \cap \text{span}(\boldsymbol{V})$. Then for any vector $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{W}^\top\boldsymbol{a} = \boldsymbol{W}^\top\text{Proj}_{\boldsymbol{U}}(\boldsymbol{a}) = \boldsymbol{W}^\top\text{Proj}_{\boldsymbol{V}}(\boldsymbol{a})$.*

Proof. This is straghtfoward using that

$$\boldsymbol{W}^\top = \boldsymbol{W}^\top\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{W}^\top\boldsymbol{V}\boldsymbol{V}^\top$$

since $\boldsymbol{V}$ spans a subspace of $\text{span}(\boldsymbol{U})$ and $\text{span}(\boldsymbol{V})$. □

Lemma A2. *Under H_0, $\boldsymbol{V}_0\hat{\boldsymbol{\varepsilon}} = \text{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$. Moreover, for any permutation matrix $\boldsymbol{P}_k$, we have that $\boldsymbol{V}_k\hat{\boldsymbol{\varepsilon}} = \text{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k\boldsymbol{\varepsilon})$.*

Proof. Since we are under the H_0, we have that

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_0^\top\boldsymbol{Y} = \boldsymbol{V}_0^\top(\boldsymbol{X}\beta + \boldsymbol{\varepsilon}).$$

Then as a direct consequence of that $\text{span}(\boldsymbol{V}_0)$ is orthogonal to $\text{span}(\boldsymbol{X})$, we have that $\boldsymbol{V}_0^\top\boldsymbol{X} = 0$ and thus $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_0^\top\boldsymbol{\varepsilon}$. From above, we have

$$\boldsymbol{V}_0\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{\varepsilon} = \text{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$$

and that

$$\boldsymbol{V}_k\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_k\boldsymbol{V}_0^\top\boldsymbol{\varepsilon} = \boldsymbol{V}_k\boldsymbol{V}_0^\top\boldsymbol{P}_k^\top\boldsymbol{P}_k\boldsymbol{\varepsilon} = \boldsymbol{V}_k\boldsymbol{V}_k^\top\boldsymbol{P}_k\boldsymbol{\varepsilon} = \mathrm{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k\boldsymbol{\varepsilon}).$$

□

Proof of Theorem 2. Throughout the proof we work on a fixed $(\boldsymbol{X}, \boldsymbol{Z})$ and a fixed set of permutation matrices $\{\boldsymbol{P}_0, \ldots, \boldsymbol{P}_K\}$ satisfying Assumption 2.

From Lemmas A1 and A2, we have that for any $\alpha \in [0, 1]$,

$$\begin{aligned}\mathrm{I}_\alpha &:= \mathbb{P}\left(\frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{\varepsilon}}\right) \le T\left(\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\hat{\boldsymbol{\varepsilon}}\right)\right\}\right) \le \alpha\right) \\ &= \mathbb{P}\left(\frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{\varepsilon}\right) \le T\left(\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{\varepsilon}\right)\right\}\right) \le \alpha\right).\end{aligned}$$

Then using that for any $k \in \{1, \ldots, K\}$,

$$\begin{aligned}&\mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{\varepsilon}\right) \le T\left(\tilde{\boldsymbol{V}}_k^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{\varepsilon}\right)\right\} \\ &\ge \mathbb{1}\left\{\min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{\varepsilon}\right) \le \min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{P}_k\boldsymbol{\varepsilon}\right)\right\},\end{aligned}$$

we have

$$\begin{aligned}\mathrm{I}_\alpha \le \mathbb{P}\Bigg(\frac{1}{K+1}\Bigg(1 + \sum_{k=1}^{K} \mathbb{1}\Bigg\{&\min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{\varepsilon}\right) \\ &\le \min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{P}_k\boldsymbol{\varepsilon}\right)\Bigg\}\Bigg) \le \alpha\Bigg).\end{aligned}$$

By defining $g : \mathbb{R}^n \mapsto \mathbb{R}$ as a fixed projection depending only on $(\boldsymbol{X}, \boldsymbol{Z})$ and $\mathcal{P}_K$ such that for any $\boldsymbol{a} \in \mathbb{R}^n$,

$$g(\boldsymbol{a}) = \min_{\tilde{\boldsymbol{V}}\in\{\tilde{\boldsymbol{V}}_1,\ldots,\tilde{\boldsymbol{V}}_K\}} T\left(\tilde{\boldsymbol{V}}^\top\boldsymbol{V}_0\hat{\boldsymbol{e}}, \tilde{\boldsymbol{V}}^\top\boldsymbol{a}\right),$$

we can further rewrite the above inequality as

$$\mathrm{I}_\alpha \le \mathbb{P}\left(\frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{g(\boldsymbol{\varepsilon}) \le g(\boldsymbol{P}_k\boldsymbol{\varepsilon})\right\}\right) \le \alpha\right).$$

Using Lemma 3, we can finally have that

$$\mathrm{I}_\alpha \le \mathbb{P}\left(\frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\left\{g(\boldsymbol{\varepsilon}) \le g(\boldsymbol{P}_k\boldsymbol{\varepsilon})\right\}\right) \le \alpha\right) \le \alpha,$$

which proves the desired results. □

A1.3.1 Proof of Lemma 3

Proof. Let $\xi_0, \dots, \xi_K \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and independent of all other randomness in the problem. Let

$$R_k := \sum_{k'=0}^{K} \mathbb{1}\{g(\boldsymbol{P}_k \boldsymbol{\varepsilon}) \le g(\boldsymbol{P}_{k'} \boldsymbol{\varepsilon})\},$$

and

$$\tilde{R}_k := \sum_{k'=0}^{K} \Big(\mathbb{1}\{g(\boldsymbol{P}_k \boldsymbol{\varepsilon}) < g(\boldsymbol{P}_{k'} \boldsymbol{\varepsilon}) \ + \mathbb{1}\{g(\boldsymbol{P}_k \boldsymbol{\varepsilon}) = g(\boldsymbol{P}_{k'} \boldsymbol{\varepsilon}) \text{ and } \xi_k \le \xi_{k'}\} \Big).$$

In other words, $\tilde{R}_k$ is the rank of $g(\boldsymbol{P}_k \boldsymbol{\varepsilon})$ among $(g(\boldsymbol{P}_{k'} \boldsymbol{\varepsilon}) : k' = 0, \dots, K)$ in a decreasing order, with random tie-breaking. Also, observe that $R_k \ge \tilde{R}_k$. By Assumptions 1 we have $\boldsymbol{\varepsilon} \overset{\text{d}}{=} \boldsymbol{P}_k \boldsymbol{\varepsilon}$ for all k, hence

$$R_0 \overset{\text{d}}{=} \sum_{k'=0}^{K} \mathbb{1}\{g(\boldsymbol{P}_k \boldsymbol{\varepsilon}) < g(\boldsymbol{P}_{k'} \boldsymbol{P}_k \boldsymbol{\varepsilon})\} = \sum_{k'=0}^{K} \mathbb{1}\{g(\boldsymbol{P}_k \boldsymbol{\varepsilon}) < g(\boldsymbol{P}_{k'} \boldsymbol{\varepsilon})\} = R_k,$$

where we used Assumption 2 in the penultimate equality. Thus, for all $k \in \{0, \dots, K\}$ and $x \in \{1, \dots, K+1\}$,

$$\mathbb{P}(R_k \le x) = \frac{1}{K+1} \sum_{k'=0}^{K} \mathbb{P}(R_{k'} \le x) \le \frac{1}{K+1} \sum_{k'=0}^{K} \mathbb{P}(\tilde{R}_{k'} \le x). \tag{A1.2}$$

On the other hand, almost surely $(\tilde{R}_0, \tilde{R}_1, \dots, \tilde{R}_K)$ is a re-arrangement of $(1, \dots, K+1)$. This means that for any fixed $j \in \{1, \dots, K+1\}$, almost surely there is a k' such that $\tilde{R}_{k'} = j$. In other words, for $j \in \{1, \dots, K+1\}$,

$$\sum_{k'=0}^{K} \mathbb{P}(\tilde{R}_{k'} = j) = 1.$$

By taking this back to (A1.2), we may further bound (A1.2) as

$$\mathbb{P}(R_k \le x) \le \frac{x}{K+1}.$$

Then

$$\mathbb{P}\left\{ \frac{1}{K+1} \left(1 + \sum_{k=1}^{K} \mathbb{1}\{g(\boldsymbol{\varepsilon}) \le g(\boldsymbol{P}_k \boldsymbol{\varepsilon})\} \right) \le \alpha \right\} = \mathbb{P}\left(\frac{R_0}{K+1} \le \alpha \right) \le \frac{\lfloor \alpha(K+1) \rfloor}{K+1} \le \alpha,$$

as desired. □

A2 Theoretical analysis of Type-II error

Notations we define $\|\|_{\text{op}}$ as operator norm, $\|\|_2$ as ℓ_2-norm, $\|\|_F$ as Frobenius norm. We define $a/b = \infty$ if $a = b = 0$ or $b = 0$. We denote $\tilde{\mathcal{D}}_{\ge,B}$ as the class of distributions that is lower bounded by a threshold B, and $\tilde{\mathcal{D}}_{\le,B}$ as the class of distributions upper bounded by some threshold B. Without loss of generality, we assume $b > 0$. Let $\boldsymbol{w} := (w_1, \dots, w_n)^\top$ and $\boldsymbol{\xi} := (\xi_1, \dots, \xi_n)^\top$ be two independent random vectors with i.i.d. entries from some distributions $\mathbb{P}_w$ and $\mathbb{P}_\xi$ respectively (where restrictions on $\mathbb{P}_w$ and $\mathbb{P}_\xi$ differ from lemma to lemma).

A2.1 Preliminary lemmas

Lemma A3. *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a deterministic matrix that varies with n and satisfies $\|\boldsymbol{M}\|_{\rm op} \le 1$. Then if $b = \omega\big(n^{-1/2}\big)$, we have that for any fixed $\delta > 0$,*

$$\forall \mathbb{P}_w \in \mathcal{D}_1 \cup \mathcal{D}_2, \mathbb{P}_\xi \in \mathcal{D}_2, \quad \lim_{n\to\infty} \mathbb{P}\bigg(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\bigg) = 0.$$

Proof. For any fixed $\boldsymbol{w}_0 \in \mathbb{R}^n$, we have

$$\begin{aligned}\mathbb{E}[|\boldsymbol{w}^\top M\boldsymbol{\xi}|^2|\boldsymbol{w} = \boldsymbol{w}_0] &= \mathbb{E}[\boldsymbol{w}^\top M\boldsymbol{\xi}\boldsymbol{\xi}^\top M^\top \boldsymbol{w}|\boldsymbol{w} = \boldsymbol{w}_0] \\ &= 2\mathbb{E}[\boldsymbol{w}^\top MM^\top \boldsymbol{w}|\boldsymbol{w} = \boldsymbol{w}_0] \le 2\|\boldsymbol{w}_0\|_2^2,\end{aligned}$$

and thus by Chebyshev's inequality, for any $\delta > 0$,

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top M\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta \mid \boldsymbol{w} = \boldsymbol{w}_0\right) \le \frac{\mathbb{E}[|\boldsymbol{w}^\top M\boldsymbol{\xi}|^2|\boldsymbol{w} = \boldsymbol{w}_0]}{\delta^2 b^2\|\boldsymbol{w}_0\|_2^4} \le \frac{2}{\delta^2 b^2\|\boldsymbol{w}_0\|_2^2}.$$

From above, we have

$$\begin{aligned}\mathbb{P}\left(\frac{|\boldsymbol{w}^\top M\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) &\le \mathbb{E}\left[\mathbb{P}\left(\frac{|\boldsymbol{w}^\top M\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta \mid \boldsymbol{w}\right)\mathbb{1}\left(\|\boldsymbol{w}\|_2^2 \ge \frac{1}{2}n\right)\right] + \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 < \frac{1}{2}n\right) \\ &\le \mathbb{E}\left[\frac{2}{\delta^2 b^2\|\boldsymbol{w}\|_2^2}\mathbb{1}\left(\|\boldsymbol{w}\|_2^2 \ge \frac{1}{2}n\right)\right] + \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 < \frac{1}{2}n\right) \\ &\le \frac{4}{\delta^2 b^2 n} + \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 < \frac{1}{2}n\right).\end{aligned}$$

Then using that $b^2 n = \omega(1)$ and Lemma A9, we obtain the desired result. □

Lemma A4. *Let $\boldsymbol{g} := (g_1, \ldots, g_n)^\top$ be a n-dimensional vector satisfying $\|\boldsymbol{g}\|_2^2 = O(n^{\frac{1-t}{1+t}})$. Then if $b = \Omega(n^{-\frac{t}{1+t}})$, we have that for any fixed $\delta > 0$,*

$$\forall \mathbb{P}_w \in \mathcal{D}_1 \cup \mathcal{D}_2, \quad \lim_{n\to\infty} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) = 0$$

Proof. Let $\boldsymbol{P}$ be a uniformly random permutation matrix, then $\boldsymbol{P}\boldsymbol{w} \overset{d}{=} \boldsymbol{w}$, and our task becomes proving $\forall \mathbb{P}_w \in \mathcal{D}_1 \cup \mathcal{D}_2$,

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) \to 0.$$

For any $i \ne j$, we have

$$\mathbb{E}[(\boldsymbol{P}\boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{P}^\top|\boldsymbol{w} = \boldsymbol{w}_0)_{i,j}] = \frac{1}{n(n-1)}\sum_{k\ne\ell} w_{0,k}w_{0,\ell} = \frac{(\mathbf{1}^\top \boldsymbol{w}_0)^2}{n(n-1)} - \frac{\|\boldsymbol{w}_0\|_2^2}{n(n-1)},$$

where $\mathbf{1}$ denotes the n dimensional vector with all entries equal to 1. From above and Lemma A13, we have from basic matrix algebra that

$$\mathbb{E}[\boldsymbol{P}\boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{P}^\top|\boldsymbol{w} = \boldsymbol{w}_0] = \left(\frac{\|\boldsymbol{w}_0\|_2^2}{n} - \frac{(\sum \mathbf{1}^\top \boldsymbol{w}_0)^2}{2n(n-1)} + \frac{\|\boldsymbol{w}_0\|_2^2}{2n(n-1)}\right)\boldsymbol{I} + \left(\frac{(\mathbf{1}^\top \boldsymbol{w}_0)^2}{2n(n-1)} - \frac{\|\boldsymbol{w}_0\|_2^2}{2n(n-1)}\right)\mathbf{1}\mathbf{1}^\top,$$

where $\boldsymbol{I}$ denotes the $n \times n$ identity matrix. This allows us to further control $\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g})^2 | \boldsymbol{w} = \boldsymbol{w}_0]$ via that

$$\begin{aligned}
&\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g})^2 | \boldsymbol{w} = \boldsymbol{w}_0] = \mathrm{tr}[\mathbb{E}[\boldsymbol{P}\boldsymbol{w}_0\boldsymbol{w}_0^\top \boldsymbol{P}^\top]\boldsymbol{g}\boldsymbol{g}^\top] \\
&= \mathrm{tr}\left[\left(\frac{\|\boldsymbol{w}_0\|_2^2}{n} - \frac{(\mathbf{1}^\top \boldsymbol{w}_0)^2}{2n(n-1)} + \frac{\|\boldsymbol{w}_0\|_2^2}{2n(n-1)}\right)\boldsymbol{g}\boldsymbol{g}^\top + \left(\frac{(\mathbf{1}^\top \boldsymbol{w}_0)^2}{2n(n-1)} - \frac{\|\boldsymbol{w}_0\|_2^2}{2n(n-1)}\right)\mathbf{1}\mathbf{1}^\top \boldsymbol{g}\boldsymbol{g}^\top\right] \\
&\le \frac{2}{n}\|\boldsymbol{w}_0\|_2^2 \boldsymbol{g}^\top \boldsymbol{g} + \frac{(\mathbf{1}^\top \boldsymbol{w}_0)^2}{2n(n-1)}(\mathbf{1}^\top \boldsymbol{g})^2.
\end{aligned}$$

Now applying again Cauchy-Schwartz inequality,

$$\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g})^2 | \boldsymbol{w} = \boldsymbol{w}_0] \le \frac{2}{n}\|\boldsymbol{w}_0\|_2^2 \|\boldsymbol{g}\|_2^2 + \frac{(\mathbf{1}^\top \boldsymbol{w}_0)^2}{2(n-1)}\|\boldsymbol{g}\|_2^2. \tag{A2.3}$$

Thus by defining $\mathcal{E}(s) := \left\{\|\boldsymbol{w}\|_2^2 > \frac{1}{2}n, \frac{|\mathbf{1}^\top \boldsymbol{w}|}{n} \le s\right\}$ indexed by a $s > 0$, which, using Lemma A9 and the strong law of large number, holds with probability converging to 1 for any constant $s > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) \le \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta \mid \mathcal{E}(s)\right) + \mathbb{P}(\mathcal{E}^c(s)) \\
&\le \mathbb{E}\left[\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta \mid \boldsymbol{w}\right) \mid \mathcal{E}(s)\right] + \mathbb{P}(\mathcal{E}^c(s)) \le \mathbb{E}\left[\frac{\frac{2}{n}\|\boldsymbol{w}\|_2^2\|\boldsymbol{g}\|_2^2 + \frac{(\mathbf{1}^\top \boldsymbol{w})^2}{2(n-1)}\|\boldsymbol{g}\|_2^2}{b^2\delta^2\|\boldsymbol{w}\|_2^4} \mid \mathcal{E}(s)\right] + \mathbb{P}(\mathcal{E}^c(s)).
\end{aligned} \tag{A2.4}$$

Here for the last step we apply Chebyshev's inequality and (A2.3). The lemma statement yields that for sufficiently large n, there exists some $c_g, \Delta > 0$ such that $\|\boldsymbol{g}\|_2^2 \le c_g n^{\frac{1-t}{1+t}}$ and $b \ge \Delta n^{-\frac{t}{1+t}}$. From this, we have that under event $\mathcal{E}(s)$, for n sufficiently large,

$$\frac{\frac{2}{n}\|\boldsymbol{w}\|_2^2\|\boldsymbol{g}\|_2^2 + \frac{(\mathbf{1}^\top \boldsymbol{w})^2}{2(n-1)}\|\boldsymbol{g}\|_2^2}{b^2\|\boldsymbol{w}\|_2^4} \le \frac{4\|\boldsymbol{g}\|_2^2}{b^2 n^2} + \frac{2\|\boldsymbol{g}\|_2^2 s^2}{b^2(n-1)} \le o(1) + \frac{4c_g s^2}{\Delta^2}.$$

Putting back to (A2.4) yields

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) \le \frac{4c_g s}{\Delta^2} + o(1) + \mathbb{P}(\mathcal{E}^c(s)).$$

From above, and that $\mathbb{P}(\mathcal{E}^c(s)) \to 0$ for any choice of constant $s > 0$, we have for any constant $\eta \in (0, 1)$, there exists some constant s small enough such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{P}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) \le \eta.$$

Since the above result holds for any η, we obtain the desired result. □

Lemma A5. *Consider the $\boldsymbol{M}$ in Lemma A3 and let $t \in [0, 1), B > 0$ be given. Then if $b = \Omega(n^{-\frac{t}{1+t}})$, we have that for any fixed $\delta > 0$,*

$$\forall \mathbb{P}_w \in \mathcal{D}_1 \cup \mathcal{D}_2, \mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\ge,-B}, \quad \lim_{n\to\infty} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) = 0.$$

Proof. Let $\boldsymbol{P}$ be a uniformly random permutation matrix (i.e., a random matrix generated by sampling uniformly at random from the set of permutation matrices). Moreover, we require $\boldsymbol{P}$ to be independent from $\boldsymbol{w}$ and $\boldsymbol{\xi}$. Then we have that conditioning on $\boldsymbol{w}$, $\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi} \overset{d}{=} \boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{\xi}$, whence

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) = \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right).$$

Let $f_i := \xi_i \mathbb{1}(|\xi_i| \le Bi^{\frac{1}{1+t}})$ denote a truncated random variable of ξ_i. Since $\mathbb{P}_\xi \in \tilde{\mathcal{D}}_{\ge,-B}$, we have from Lemma A16 that there exists a f_i' satisfying $\mathbb{E}[f_i'] = 0$, $f_i' = f_i$ almost surely when $f_i \ge 0$ and $f_i \le f_i' < 0$ almost surely when $f_i < 0$. Moreover, we write $\boldsymbol{f} := (f_1, \dots, f_n)^\top$ and $\boldsymbol{f}' := (f_1', \dots, f_n')^\top$.

Using the new notations, we have the decomposition $\boldsymbol{\varepsilon} = \boldsymbol{f}' + (\boldsymbol{f} - \boldsymbol{f}') + (\boldsymbol{\varepsilon} - \boldsymbol{f})$, and morevoer,

$$\begin{aligned}
\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) &\le \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right) + \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 \le \frac{1}{2}n\right) \\
&\le \underbrace{\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{f}'|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right)}_{=:\mathrm{I}} + \underbrace{\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{f} - \boldsymbol{f}')|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right)}_{=:\mathrm{II}} \\
&\quad + \underbrace{\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi} - \boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right)}_{\mathrm{III}} + \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 \le \frac{1}{2}n\right).
\end{aligned}$$

From Lemma A9, we have that $\mathbb{P}\left(\|\boldsymbol{w}\|_2^2 \le \frac{1}{2}n\right) \to 0$ as $n \to \infty$, which bounds the last term. In the rest of the proof we focus on controlling the terms I – III.

We first consider I. From Lemma A14, we have for any $i \ne j$, $\mathbb{E}[(\boldsymbol{P}\boldsymbol{f}'(\boldsymbol{f}')^\top \boldsymbol{P}^\top)_{i,j}] = 0$; from Lemma A13 and that $\boldsymbol{P}$ and $\boldsymbol{f}'$ are indepenedent, we have $\mathbb{E}[(\boldsymbol{P}\boldsymbol{f}'(\boldsymbol{f}')^\top \boldsymbol{P}^\top)_{i,i}] = \frac{1}{n}\sum_{j=1}^n \mathbb{E}[(f_j')^2]$. Now to control $\frac{1}{n}\sum_{j=1}^n \mathbb{E}[(f_j')^2]$, let a_n be a sequence of integers such that as $n \to \infty$, $a_n \to \infty$ and $a_n/n \to 0$, we have that

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}[(f_i)'^2] &= \sum_{i=1}^{a_n} \mathbb{E}[(f_i')^2] + \sum_{i=a_n+1}^n \mathbb{E}[(f_i)'^2] \\
&\le \sum_{i=1}^{a_n} \mathbb{E}[\xi_i^2 \mathbb{1}(|\xi_i| \le Bi^{\frac{1}{1+t}})] + \sum_{i=a_n+1}^n \mathbb{E}[\xi_i^2 \mathbb{1}(|\xi_i| \le Ba_n^{\frac{1}{1+t}})] \\
&\quad + \sum_{i=a_n+1}^n \mathbb{E}[\xi_i^2 \mathbb{1}(Ba_n^{\frac{1}{1+t}} \le |\xi_i| \le Bi^{\frac{1}{1+t}})] \\
&\le \sum_{i=1}^n \mathbb{E}[\xi_i^2 \mathbb{1}(|\xi_i| \le Ba_n^{\frac{1}{1+t}})] + \sum_{i=a_n+1}^n \mathbb{E}[\xi_i^2 \mathbb{1}(Ba_n^{\frac{1}{1+t}} \le |\xi_i| \le Bi^{\frac{1}{1+t}})].
\end{aligned} \tag{A2.5}$$

For the first term in the above inequality,

$$\sum_{i=1}^n \mathbb{E}[\xi_i^2 \mathbb{1}(|\xi_i| \le Ba_n^{\frac{1}{1+t}})] = n\mathbb{E}[|\xi_1|^{1+t}|\xi_1|^{1-t}\mathbb{1}(|\xi_1| \le Ba_n^{\frac{1}{1+t}})] \le n\mathbb{E}[|\xi_1|^{1+t}B^{1-t}a_n^{\frac{1-t}{1+t}}] = O(n \cdot a_n^{\frac{1-t}{1+t}}) = o(n^{\frac{2}{1+t}}),$$

where for the first equality we use that the ξ_i's are i.i.d.

For the second term on the right hand side of (A2.5), we have

$$\sum_{i=a_n+1}^{n} \mathbb{E}[\xi_i^2 \mathbb{1}(Ba_n^{\frac{1}{1+t}} \le |\xi_i| \le Bi^{\frac{1}{1+t}})] = \sum_{i=a_n+1}^{n} \mathbb{E}[|\xi_i|^{1+t}|\xi_i|^{1-t}\mathbb{1}(Ba_n^{\frac{1}{1+t}} \le |\xi_i| \le Bi^{\frac{1}{1+t}})]$$
$$\le \sum_{i=a_n+1}^{n} \mathbb{E}[|\xi_i|^{1+t}\mathbb{1}(Ba_n^{\frac{1}{1+t}} \le |\xi_i| \le Bi^{\frac{1}{1+t}})] \cdot B^{1-t} n^{\frac{1-t}{1+t}} \le B^{1-t} n^{\frac{2}{1+t}} \mathbb{E}[|\xi_1|^{1+t}\mathbb{1}(|\xi_1| \ge Ba_n^{\frac{1}{1+t}})].$$

where the last inequality uses again that the ξ_i's are i.i.d. random variables. Putting together yields

$$\sum_{i=1}^{n} \mathbb{E}[(f_i)'^2] = o(n^{\frac{2}{1+t}}) + O(n^{\frac{2}{1+t}})\mathbb{E}[|\xi_1|^{1+t}\mathbb{1}(|\xi_1| \ge Ba_n^{\frac{1}{1+t}})].$$

Notice further that $|\xi_1|^{1+t}\mathbb{1}(|\xi_1| \ge Ba_n^{\frac{1}{1+t}}) \to 0$ almost surely, and almost surely $|\xi_1|^{1+t}\mathbb{1}(|\xi_1| \ge Ba_n^{\frac{1}{1+t}}) \le |\xi_1|^{1+t}$ where $\mathbb{E}[|\xi_1|^{1+t}] < \infty$. Therefore, by dominated convergence theorem, $\mathbb{E}[|\xi_1|^{1+t}\mathbb{1}(|\xi_1| \ge Ba_n^{\frac{1}{1+t}})] \to 0$. Thus $\sum_{i=1}^n \mathbb{E}[(f_i)'^2] = o(n^{\frac{2}{1+t}})$, which means that $\mathbb{E}[(\boldsymbol{P}\boldsymbol{f}'(\boldsymbol{f}')^\top \boldsymbol{P}^\top)_{i,i}] = o(n^{\frac{1-t}{1+t}})$.

In light of our control of all the (i,j)'s entries of the matrix $\mathbb{E}[\boldsymbol{P}\boldsymbol{f}'(\boldsymbol{f}')^\top \boldsymbol{P}^\top]$, we have for any fixed $\boldsymbol{w}_0 \in \mathbb{R}^n$,

$$\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{f}')^2 | \boldsymbol{w} = \boldsymbol{w}_0] = \boldsymbol{w}_0^\top \boldsymbol{M}\mathbb{E}[\boldsymbol{P}\boldsymbol{f}'(\boldsymbol{f}')^\top \boldsymbol{P}^\top]\boldsymbol{M}^\top \boldsymbol{w}_0$$
$$= o(n^{\frac{1-t}{1+t}})\boldsymbol{w}_0^\top \boldsymbol{M}\boldsymbol{M}^\top \boldsymbol{w}_0 = o(n^{\frac{1-t}{1+t}}) \cdot \|\boldsymbol{w}_0\|_2^2.$$

From above, and by Chebyshev's inequality,

$$\mathrm{I} = \mathbb{E}\left[\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{f}'|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \boldsymbol{w}\right) \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right] \le \mathbb{E}\left[\frac{9\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{f}')^2|\boldsymbol{w}]}{\delta^2(\|\boldsymbol{w}\|_2^2)^2 b^2} \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right]$$
$$= o(n^{\frac{1-t}{1+t}}) \cdot \mathbb{E}\left[\frac{1}{\|\boldsymbol{w}\|_2^2 b^2} \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right] = o\left(\frac{n^{\frac{1-t}{1+t}}}{b^2 n}\right) = o(1).$$

We second consider II. Notice that again

$$\mathrm{II} \le \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{f} - \boldsymbol{f}' - \mathbb{E}[\boldsymbol{f} - \boldsymbol{f}'])|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{6} \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right) + \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\mathbb{E}[\boldsymbol{f} - \boldsymbol{f}']|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{6} \mid \|\boldsymbol{w}\|_2^2 > \frac{1}{2}n\right)$$
$$=: \mathrm{II}_1 + \mathrm{II}_2.$$

For II_1, as $f_i - f_i'$ is bounded between $[-B, B]$, using the same analysis as in Lemma A3, we have $\mathrm{II}_1 \to 0$. For II_2, observe first that

$$\|\mathbb{E}[\boldsymbol{f} - \boldsymbol{f}']\|_2^2 = \sum_i (\mathbb{E}[f_i - f_i'])^2 = \sum_i (\mathbb{E}[f_i])^2 = \sum_i (\mathbb{E}[\xi_i \mathbb{1}(|\xi_i| > Bi^{\frac{1}{1+t}})])^2$$
$$\overset{(i)}{\le} \sum_i (\mathbb{E}[|\xi_i|^{1+t}])^{\frac{2}{1+t}} (\mathbb{E}[\mathbb{1}(|\xi_i| > Bi^{\frac{1}{1+t}})])^{\frac{2t}{1+t}} \le 2^{\frac{2}{1+t}} \sum_i \mathbb{P}(|\xi_i|^{1+t} > B^{1+t} i)^{\frac{2t}{1+t}}$$
$$\overset{(ii)}{\le} 2^{\frac{2}{1+t}} n \left(\frac{\sum_i \mathbb{P}(|\xi_i|^{1+t} > B^{1+t} i)}{n}\right)^{\frac{2t}{1+t}} \overset{(iii)}{=} O(n^{\frac{1-t}{1+t}}),$$

where (i) uses Hölder's inequality; (ii) uses Jensen's inequality; (iii) uses Lemma A10. Then applying Lemma A4 with $\boldsymbol{P}\mathbb{E}[\boldsymbol{f}-\boldsymbol{f}']$ as $\boldsymbol{g}$ and noticing also Lemma A9, we have $\mathrm{II}_2 \to 0$ as $n \to \infty$. Putting together yields $\mathrm{II} \to 0$.

We third consider III. For any n, from Lemma A10,

$$\sum_{i=1}^{n} \mathbb{P}(f_i \neq \xi_i) = \sum_{i=1}^{n} \mathbb{P}(|\xi_i|^{1+t} > B^{1+t} i) < \infty.$$

By Borel-Cantelli Lemma (see e.g. Lemma A12), the event $f_i \neq \xi_i$ happens finite time almost surely. That is by setting the random variable $V_i := \mathbb{1}(f_i \neq \xi_i)$, V_i converges to zero almost surely. Hence $V_i \to 0$ in probability, which means that for any $\eta \in (0,1)$, there exists a constant N_η depending on η such that $\mathbb{P}(V_n = 0,\ \forall n \geq N_\eta) \leq \frac{\eta}{3}$. This is equivalent to that

$$\mathbb{P}(\exists k > N_\eta \text{ s.t. } f_k \neq \xi_k) \leq \frac{\eta}{3}.$$

Using that N_η is finite, we further have there exists a constant C_η such that

$$\mathbb{P}(\exists \ell \leq N_\eta, \text{ s.t. } |\xi_\ell| > C_\eta) \leq \frac{\eta}{3}.$$

Writing the event $\mathcal{E} := \{\forall k > N_\eta, f_k = \xi_k, \forall \ell \leq N_\eta, |\xi_i| < C_\eta\}$. From above, we have $\mathbb{P}(\mathcal{E}^c) \leq \frac{3\eta}{2}$, which gives us that

$$\begin{aligned}
\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3}\right) &\leq \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\leq \frac{2\eta}{3} + \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \mathcal{E}\right).
\end{aligned}$$

Under the random event $\mathcal{E}$, using Cauchy-Schwartz inequality, we have

$$\begin{aligned}
|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})| &\leq \|\boldsymbol{w}\|_2 \|\boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})\|_2 \leq \|\boldsymbol{w}\|_2 \|\boldsymbol{\xi}-\boldsymbol{f}\|_2 = \|\boldsymbol{w}\|_2 \sqrt{\sum_{i=1}^{N_\eta} \xi_i^2 \mathbb{1}(|\xi_i|^{1+t} > B^{1+t} i)} \\
&\leq \|\boldsymbol{w}\|_2 C_\eta \sqrt{N_\eta}.
\end{aligned}$$

Putting back yields

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3}\right) \leq \frac{2\eta}{3} + \mathbb{P}\left(\frac{C_\eta \sqrt{N_\eta}\|\boldsymbol{w}\|_2}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \mathcal{E}\right) \leq \frac{2\eta}{3} + o(1),$$

where the last inequality uses Lemma A9. Since it works for any η, we have $\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3}\right) \to 0$. Applying again Lemma A9, we have $\mathrm{III} \to 0$.

In light of our control of I – III, our desired result follows. □

Lemma A6. *Consider the $\boldsymbol{M}$ in Lemma A3; let $t \in [0,1]$ be given and assume that b satisfies (8). Then for any fixed $\delta > 0$,*

$$\forall \mathbb{P}_e \in \mathcal{D}_1 \cup \mathcal{D}_2, \mathbb{P}_\xi \in \mathcal{D}_{1+t} \quad \lim_{n\to\infty} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) = 0.$$

Proof. When $t = 1$, the result follows from Lemma A3. Otherwise, we apply the decomposition

$$\xi_i = (\xi_i \mathbb{1}(\xi_i \geq 0) - \mathbb{E}[\xi_i \mathbb{1}(\xi_i \geq 0)]) - ((-\xi_i)\mathbb{1}(\xi_i < 0) - \mathbb{E}[(-\xi_i)\mathbb{1}(\xi_i < 0)]) =: \xi_{1,i} - \xi_{2,i};$$

and define $\boldsymbol{\xi}_1 := (\xi_{1,1}, \ldots, \xi_{1,n})^\top$, $\boldsymbol{\xi}_2 := (\xi_{2,1}, \ldots, \xi_{2,n})^\top$. Then our desired result follows by applying Lemma A5 but with $\boldsymbol{\xi}$ replaced by $\boldsymbol{\xi}_1$ or $\boldsymbol{\xi}_2$ and taking a union bound. □

Lemma A7. *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a matrix with all diagonal entries equal to zero. Then for any $\mathbb{P}_w \in \mathcal{D}_2$, we have for any fixed $\delta > 0$,*

$$\mathbb{P}\left(\frac{\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{w}}{n} > \delta\right) \leq \frac{4\|\boldsymbol{M}\|_F^2}{n^2\delta^2}.$$

Proof. Observe that

$$\mathbb{E}\left[\left(\frac{\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{w}}{n}\right)^2\right] = \mathbb{E}\left[\left(\sum_{i\neq j} \boldsymbol{M}_{i,j}\frac{w_i w_j}{n}\right)^2\right].$$

Using that for any $i \neq j$, $w_i \perp\!\!\!\perp w_j$, we have

$$\mathbb{E}\left[\left(\frac{\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{w}}{n}\right)^2\right] = \mathbb{E}\left[\sum_{i,j} \boldsymbol{M}_{i,j}^2\frac{w_i^2 w_j^2}{n^2}\right] \leq \frac{4\|\boldsymbol{M}\|_F^2}{n^2}.$$

Then by applying Chebyshev's inequality, we obtain the desired result. □

Lemma A8. *Consider a deterministic permutation matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ that varies with n and $\mathrm{tr}[\boldsymbol{P}] = 0$. We have that for any fixed $\delta > 0$*

$$\forall \mathbb{P}_w \in \mathcal{D}_1, \quad \lim_{n\to\infty} \mathbb{P}(|\boldsymbol{w}^\top \boldsymbol{P}\boldsymbol{w}|/n > \delta) = 0$$

.

Proof. Let σ be the permutation corresponding to $\boldsymbol{P}$. From Lemma 5, we have there exists a partition U_1, U_2, U_3 with $|U_j \cap \sigma(U_j)| = 0$ and that $|U_j| \geq \frac{n}{4} - 1$ for $j = 1, 2, 3$ such that

$$\frac{\boldsymbol{w}^\top \boldsymbol{P}\boldsymbol{w}}{n} = \frac{1}{n}\sum_{j=1}^{3}\sum_{i\in U_j} w_i w_{\sigma(i)}.$$

Then

$$\lim_{n\to\infty} \mathbb{P}(|\boldsymbol{w}^\top \boldsymbol{P}\boldsymbol{w}|/n > \delta) \leq \sum_{j=1}^{3} \mathbb{P}\left(\frac{1}{|U_j|}\left|\sum_{i\in U_j} w_i w_{\sigma(i)}\right| > \frac{\delta}{3}\right).$$

From above, it remains to prove that for any j and any fixed $\delta > 0$,

$$\mathbb{P}\left(\frac{1}{|U_j|}\left|\sum_{i\in U_j} w_i w_{\sigma(i)}\right| > \delta\right) \to 0.$$

Let $\tilde{w}_i$ be a sequence of i.i.d. random variables that is independent from $\boldsymbol{w}$ and that $\tilde{w}_i \overset{\mathrm{d}}{=} w_1 w_2$. Then we easily have that $\tilde{w}_i$ are i.i.d. random variables with zero mean and bounded first order moment. Then using the weak law of large number, we have that with $a_n(\delta) := \sup_{m \geq n} \mathbb{P}(|\sum_{i=1}^m \tilde{w}_i/m| > \delta)$,

$$\lim_{n\to\infty} a_n(\delta) = \limsup_{n\to\infty} \mathbb{P}\left(\left|\sum_{i=1}^n \tilde{w}_i/n\right| > \delta\right) = \lim_{n\to\infty} \mathbb{P}\left(\left|\sum_{i=1}^n \tilde{w}_i/n\right| > \delta\right) = 0.$$

Using that the U_j and $\sigma(U_j)$ has no overlap, we have

$$\sum_{i=1}^{|U_j|} \tilde{w}_i \overset{\mathrm{d}}{=} \sum_{i\in U_j} w_i w_{\sigma(i)}$$

and thus

$$\mathbb{P}\left(\frac{1}{|U_j|}\left|\sum_{i\in U_j} w_i w_{\sigma(i)}\right| > \delta\right) \leq a_{|U_j|}(\delta) \leq a_{\lceil n/4-1\rceil}(\delta) \to 0,$$

where for the last inequality we use that $a_n(\delta)$ is non-increasing and $|U_j| \geq n/4 - 1$. □

Proof of Lemma 4. Let $\boldsymbol{J}$ be a random diagonal matrix where all diagonal entries $\boldsymbol{J}_{i,i}$ are i.i.d. binary random variables with $\mathbb{P}(\boldsymbol{J}_{i,i} = 1) = \mathbb{P}(\boldsymbol{J}_{i,i} = -1) = \frac{1}{2}$. We write $\boldsymbol{P}$ for a uniformly random permutation matrix that is independent from $\boldsymbol{J}$. Recalling that $\mathbb{P}_w$ is symmetric and all the w_i's are independent, we have that $\boldsymbol{w} \overset{d}{=} \boldsymbol{P}\boldsymbol{J}\boldsymbol{w}$, i.e., they are equal in distribution.

This allows us to prove the statement by controlling $\mathbb{P}(\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} > \delta\|\boldsymbol{w}\|_2^2)$ due to that

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{w}^\top \boldsymbol{U}\boldsymbol{w} \geq \delta\|\boldsymbol{w}\|_2^2\right) &= \mathbb{P}\left(\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} > \delta \boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{J}\boldsymbol{w}\right) \\ &= \mathbb{P}\left(\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} > \delta\|\boldsymbol{w}\|_2^2\right). \end{aligned} \tag{A2.6}$$

First, for any fixed $\boldsymbol{w}_0 \in \mathbb{R}^n$, we have

$$\mathbb{E}[\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} | \boldsymbol{w} = \boldsymbol{w}_0] = \boldsymbol{w}_0^\top \mathbb{E}[\boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}]\boldsymbol{w}_0.$$

Second, for any fixed matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, we have $\mathbb{E}[(\boldsymbol{J}^\top \boldsymbol{M}\boldsymbol{J})_{i,j}] = \mathbb{E}[\boldsymbol{J}_{i,i}\boldsymbol{M}_{i,j}\boldsymbol{J}_{j,j}] = 0$ whenever $i \neq j$ and $\mathbb{E}[(\boldsymbol{J}^\top \boldsymbol{M}\boldsymbol{J})_{i,i}] = \mathbb{E}[\boldsymbol{J}_{i,i}\boldsymbol{M}_{i,i}\boldsymbol{J}_{i,i}] = \boldsymbol{M}_{i,i}$. Putting together and apply Lemma A13, we have

$$\mathbb{E}[\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} | \boldsymbol{w} = \boldsymbol{w}_0] = \boldsymbol{w}_0^\top \mathbb{E}[\boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}]\boldsymbol{w}_0 = \frac{\mathrm{tr}(\boldsymbol{U})}{n}\|\boldsymbol{w}_0\|_2^2.$$

From above and Markov's inequality, we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} > \delta\|\boldsymbol{w}\|_2^2) &= \mathbb{E}\left[\mathbb{P}\left(\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} > \delta\|\boldsymbol{w}\|_2^2 \mid \boldsymbol{w}\right)\right] \\ &\leq \mathbb{E}\left[\frac{\mathbb{E}[\boldsymbol{w}^\top \boldsymbol{J}^\top \boldsymbol{P}^\top \boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{w} \mid \boldsymbol{w}]}{\delta\|\boldsymbol{w}\|_2^2}\right] \\ &= \mathbb{E}\left[\frac{\mathrm{tr}[\boldsymbol{U}]}{\delta n}\right] = \frac{\mathrm{tr}[\boldsymbol{U}]}{\delta n}.\end{aligned}$$

In light of the above equality and (A2.6), we obtain the desired result. □

Proof of Lemma 5. Let G be a directed graph on vertices $\{1, \ldots, n\}$ where there exists a directed edge $i \to j$ in G if and only if $j = \sigma(i)$. Then the cycles in G are of length at least 2.

Let U denote a set with the maximum number of notes such that $|U \cap \sigma(U)| = 0$, then apparently $|U| < \frac{n}{2} + 1$. Let G' denote the subgraph of G removing all the edges of the type $(u, \sigma(u))$ for $u \in U$. Then we must have that a node is in U^c if and only if the node has an out edge in G'. Moreover, we claim that (i) G' does not contain a circle with length 2; (ii) all the connected component of G' has no more than 2 edges. To prove claim (i), suppose in contradiction there exists a circle $a \to b \to a$ in G', then we must have that $a, b \notin U$. This means that the set $U' = U \cup \{b\}$ can still satisfy that $|U' \cap \sigma(U')| = 0$, which contradicts that U is maximal. To prove claim (ii), suppose in contradiction there exists a connected component with at least 3 edges, then in this component there must exists a path $a \to b \to c \to d$ or $a \to b \to c \to a$. Then we easily have that $b, c \notin U$. This means that the set $U' = U \cup \{b\}$ can still satisfy that $|U' \cap \sigma(U')| = 0$, which contradicts that U is maximal.

From the two claims, we must have that all the connected components in G' must be of the form $a \to b$ or $a \to b \to c$. We now introduce three sets of nodes A, B, C, where A consists of all the nodes a such that $a \to b$ formalizes a connected component in G'; B consists of all the nodes a such that $a \to b \to c$ is a connected component in G'; and C consists of all the nodes b such that $a \to b \to c$ is a connected component in G'. Now recall the claim that a node is in U^c if and only if the node has an out edge in G', we have that the four disjoint sets A, B, C, U formalizes a partition of all the nodes; moreover, $\sigma(A) \subseteq U$, $\sigma(B) = C$, $\sigma(C) \subseteq U$, $\sigma(U) = A \cup B$.

From above, we split A into two sets A_1, A_2 with size $|A_1|$ and $|A_2|$ differ by at most 1; and set $U_1 = U, U_2 = A_1 \cup B, U_3 = A_2 \cup C$. Then it is straightforward that for all $i = 1, 2, 3$,

$$\frac{n}{4} - 1 \leq \frac{n - |U_1| - 1}{2} \leq |U_i| \leq |U_1| \leq \frac{n}{2} + 1$$

and that

$$|U_i \cap \sigma(U_i)| = 0,$$

which proves the desired result. □

A2.2 Proof of Theorem 3

Proof. Without loss of generality, we assume throughout that $\mathbb{P}_e \in \mathcal{D}_2$ and $\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}$. Since K is finite, we only need to prove that for any $j, k \in \{1, \ldots, K\}$, as $n \to \infty$,

$$\mathbb{P}\left(|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top (\boldsymbol{\varepsilon} + b\boldsymbol{e})| \leq |\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k (\boldsymbol{\varepsilon} + b\boldsymbol{e})|\right) \to 0.$$

In this proof, we tackle this problem via proving that for all $\delta > 0$,

$$\begin{aligned} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} \geq \delta\right) &\to 0; \\ \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}|}{bn} \geq \delta\right) &\to 0; \end{aligned} \tag{A2.7}$$

and that with probability converging to 1,

$$\begin{aligned}\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} &\geq \frac{m}{2(4+m)};\\ \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} &\geq \frac{m}{2(4+m)}.\end{aligned} \tag{A2.8}$$

To prove the first claim of (A2.7), since $\mathbb{P}_e \in \mathcal{D}_2$, we have from the law of large number that

$$\mathbb{P}\left(\frac{1}{2}n \leq \|\boldsymbol{e}\|_2^2 \leq \frac{5}{2}n\right) \to 1.$$

Let $\mathcal{E}$ denote the above event. Then applying basic inequalities of random events, we have

$$\begin{aligned}\mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} \geq \delta\right) &\leq \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} \geq \delta \mid \mathcal{E}\right)\mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c)\\ &\overset{(i)}{\leq} \mathbb{P}\left(\frac{5|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{2b\|\boldsymbol{e}\|_2^2} \geq \delta \mid \mathcal{E}\right)\mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}\left(\frac{5|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{2b\|\boldsymbol{e}\|_2^2} \geq \delta\right) + \mathbb{P}(\mathcal{E}^c),\end{aligned}$$

where for the inequality (i) we apply that we are under $\mathcal{E}$. Then as a direct consequence of Lemma A6, we prove the first claim of (A2.7). For the second claim of (A2.7), using that $\boldsymbol{P}_k\boldsymbol{\varepsilon} \overset{d}{=} \boldsymbol{\varepsilon}$, The result follows the same argument as the first claim of (A2.7).

In the rest of the proof we focus on proving the first statement of (A2.8), and the second statement can be proven via a similar argument. To prove this statement, we apply the decomposition

$$\begin{aligned}\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} =& \frac{\boldsymbol{e}^\top(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \mathrm{diag}(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top))\boldsymbol{e} - \boldsymbol{e}^\top(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k - \mathrm{diag}(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k))\boldsymbol{e}}{n}\\ &+ \frac{\boldsymbol{e}^\top \mathrm{diag}(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top)\boldsymbol{e} - \boldsymbol{e}^\top \mathrm{diag}(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)\boldsymbol{e}}{n}\\ =:& \mathrm{I} + \mathrm{II},\end{aligned}$$

where for any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathrm{diag}(\boldsymbol{A})$ corresponds to the diagonal matrix such that all the diagonal elements are equal to the diagonal elements of $\boldsymbol{A}$.

For I, observe that

$$\begin{aligned}\|\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \mathrm{diag}(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top)\|_F^2 \leq \|\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top\|_F^2 &= \mathrm{tr}[\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top]\\ &= \mathrm{tr}[\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top] = n - 2p,\end{aligned}$$

and that

$$\begin{aligned}\|\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k - \mathrm{diag}(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)\|_F^2 \leq \|\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k\|_F^2 &= \mathrm{tr}(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{P}_k^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top)\\ &= \mathrm{tr}(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top) = n - 2p,\end{aligned}$$

We can apply Lemma A7 to show that for any constant $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\mathrm{I}| \le \delta) \to 1. \tag{A2.9}$$

For II, given any fixed $\boldsymbol{P}_j$ and $\boldsymbol{P}_k$, we write $V_i := e_i^2$ and

$$a_{n,i} := \frac{(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)_{i,i}}{n},$$

then we can rewrite II as $\mathrm{II} = \sum_{i=1}^n a_{n,i} V_i$. It is straightforward that for each n, the absolute value all the entires in $a_{n,i}$ are bounded by $\frac{2}{n}$. Thus given any fixed $\boldsymbol{P}_j, \boldsymbol{P}_k$,

$$\sup_{n\ge 1} \sum_{i=1}^n |a_{n,i}| \mathbb{E}[|V_i| \mathbb{1}(|V_i| > a)] \le 2\mathbb{E}[|V_i| \mathbb{1}(|V_i| > a)],$$

which converges to 0 as $a \to \infty$. Moreover,

$$\lim_{n\to\infty} \sum_{i=1}^n |a_{n,i}|^2 \le \lim_{n\to\infty} \frac{4}{n} = 0.$$

This allows us to apply Lemma A11 to get that for any constant $\delta > 0$,

$$\mathbb{P}\left(|\mathrm{II} - \mathbb{E}[\mathrm{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k]| > \delta \mid \boldsymbol{P}_j, \boldsymbol{P}_k\right) \to 0. \tag{A2.10}$$

Thus, it remains to control $\mathbb{E}[\mathrm{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k] = \sum_{i=1}^n a_{n,i}$. We write

$$\boldsymbol{A}_k \boldsymbol{A}_k^\top = \boldsymbol{I} - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top, \tag{A2.11}$$

where $\boldsymbol{A}_k$ is a $n \times (n-2p)$ matrix with orthonormal columns. Since the column space of $\tilde{\boldsymbol{V}}_k$ is at the intersection of $\mathrm{span}(\boldsymbol{X})^\perp$ and $\mathrm{span}(\boldsymbol{P}_k \boldsymbol{X})^\perp$, we have that $\mathrm{span}(\boldsymbol{X})$ must be a subspace of $\mathrm{span}(\boldsymbol{A}_k)$. Hence without loss of generality we can write $\boldsymbol{A}_k := [\boldsymbol{A}_0, \boldsymbol{B}_k]$, where $\boldsymbol{A}_0 \in \mathbb{R}^{n\times p}$ is a matrix with orthonormal columns spanning $\mathrm{span}(\boldsymbol{V}_0)^\perp$. With the above notations, we calculate

$$\begin{aligned}
\mathbb{E}[\mathrm{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k] &= \sum_{i=1}^n a_{n,i} = \frac{1}{n} \mathrm{tr}[\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k] = \frac{1}{n}((n-2p) - \mathrm{tr}[\boldsymbol{A}_k \boldsymbol{A}_k^\top \boldsymbol{P}_k]) \\
&= \frac{1}{n}((n-2p) - \mathrm{tr}[\boldsymbol{A}_0 \boldsymbol{A}_0^\top \boldsymbol{P}_k + \boldsymbol{B}_k \boldsymbol{B}_k^\top \boldsymbol{P}_k]).
\end{aligned}$$

From Assumption 4, we have $\mathrm{tr}[\boldsymbol{A}_0 \boldsymbol{A}_0^\top \boldsymbol{P}_k] \le \sqrt{2p}K$, and using Lemma A15, we have $\mathrm{tr}[\boldsymbol{B}_k \boldsymbol{B}_k^\top \boldsymbol{P}_k] \le \mathrm{tr}[\boldsymbol{B}_k \boldsymbol{B}_k^\top] \le p$, putting together we further have

$$\mathbb{E}[\mathrm{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k] \ge \frac{1}{n}((n-2p) - p - \sqrt{2p}K) \ge \frac{m}{4+m},$$

where the last inequality holds for sufficiently large n. From above and (A2.10), and also our control of the term I in (A2.9), we have that the first statement of (A2.8) holds with probability converging to 1. Using an analogous argument we prove the second statement of (A2.8). In light of this and our analysis of (A2.7), we obtain the desired result. □

A2.3 Proof of Theorem 4

Proof. Without loss of generality, we assume throughout that $\mathbb{P}_e \in \mathcal{D}_1$ and $\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}$. Following analogous argument as in the proof of Theorem 3, we tackle this problem via proving that for any $j, k \in \{1, \ldots, K\}$ and for all $\delta > 0$,

$$\begin{aligned} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2} \geq \delta\right) \to 0; \\ \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2} \geq \delta\right) \to 0; \end{aligned} \tag{A2.12}$$

and that with probability converging to 1,

$$\begin{aligned} \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{\|\boldsymbol{e}\|^2} \geq \frac{1}{5}; \\ \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{\|\boldsymbol{e}\|^2} \geq \frac{1}{5}. \end{aligned} \tag{A2.13}$$

The first claim of (A2.12) directly follows Lemma A6. The second claim of (A2.12) uses Lemma A6 and that $\boldsymbol{P}_k \boldsymbol{\varepsilon} \overset{d}{=} \boldsymbol{\varepsilon}$. In the rest of the proof we focus on proving the first statement of (A2.13), and the second statement can be proven via a similar argument.

In the rest of the proof we assume throughout that both $\boldsymbol{P}_j$ and $\boldsymbol{P}_k$ are fixed permutation matrices or equivalently being conditioned on. To prove the statement, let $\boldsymbol{e}'$ denote an independent replication of $\boldsymbol{e}$. Recalling the definition of $\boldsymbol{A}_k$ in (A2.11), we have

$$\begin{aligned} &(\boldsymbol{e} - \boldsymbol{e}')^\top (\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)(\boldsymbol{e} - \boldsymbol{e}') \\ &= \|\boldsymbol{e} - \boldsymbol{e}'\|_2^2 - (\boldsymbol{e} - \boldsymbol{e}')^\top \boldsymbol{P}_k (\boldsymbol{e} - \boldsymbol{e}') - (\boldsymbol{e} - \boldsymbol{e}')^\top (\boldsymbol{A_j} \boldsymbol{A_j}^\top - \boldsymbol{A_k} \boldsymbol{A_k}^\top \boldsymbol{P}_k)(\boldsymbol{e} - \boldsymbol{e}') \\ &\geq \|\boldsymbol{e} - \boldsymbol{e}'\|_2^2 - (\boldsymbol{e} - \boldsymbol{e}')^\top \boldsymbol{P}_k (\boldsymbol{e} - \boldsymbol{e}') - (\boldsymbol{e} - \boldsymbol{e}')^\top \left(\boldsymbol{A_j} \boldsymbol{A_j}^\top + \frac{\boldsymbol{A_k} \boldsymbol{A_k}^\top + \boldsymbol{P}_k^\top \boldsymbol{A_k} \boldsymbol{A_k}^\top \boldsymbol{P}_k}{2}\right)(\boldsymbol{e} - \boldsymbol{e}'), \end{aligned} \tag{A2.14}$$

where for the last inequality we apply Cauchy-Schwartz inequality. As $e_i - e_i'$ is symmetric around zero, we have from Lemma 4 that the following event $\mathcal{E}_1$ holds with probability $1 - \frac{10p}{n} \to 1$:

$$\mathcal{E}_1 := \left\{(\boldsymbol{e} - \boldsymbol{e}')^\top (\boldsymbol{A_j} \boldsymbol{A_j}^\top + \frac{\boldsymbol{A_k} \boldsymbol{A_k}^\top + \boldsymbol{P}_k^\top \boldsymbol{A_k} \boldsymbol{A_k}^\top \boldsymbol{P}_k}{2})(\boldsymbol{e} - \boldsymbol{e}') < \frac{1}{5}(\boldsymbol{e} - \boldsymbol{e}')^\top (\boldsymbol{e} - \boldsymbol{e}')\right\}.$$

In addition, as $\|\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k\|_{\text{op}} \leq 2$, we have from Lemma A6 that the following two events $\mathcal{E}_2$ and $\mathcal{E}_3$ hold with probability converging to 1:

$$\begin{aligned} \mathcal{E}_2 &:= \left\{\left|\boldsymbol{e}'^\top \left(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k\right) \boldsymbol{e}\right| < \frac{1}{5}\|\boldsymbol{e}\|_2^2\right\}; \\ \mathcal{E}_3 &:= \left\{\left|\boldsymbol{e}'^\top \left(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k\right) \boldsymbol{e}\right| < \frac{1}{5}\|\boldsymbol{e}\|_2^2\right\}. \end{aligned}$$

Working on the intersection of the three events $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, and applying the decomposition

$$\begin{aligned} (\boldsymbol{e} - \boldsymbol{e}')^\top (\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)(\boldsymbol{e} - \boldsymbol{e}') = \boldsymbol{e}^\top (\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)\boldsymbol{e} + \boldsymbol{e}'^\top (\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)\boldsymbol{e}' \\ - \boldsymbol{e}^\top (\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)\boldsymbol{e}' - \boldsymbol{e}'^\top (\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)\boldsymbol{e}, \end{aligned}$$

we have from (A2.14) that

$$\begin{aligned}
&\boldsymbol{e}^\top(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)\boldsymbol{e} + \boldsymbol{e}'^\top(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)\boldsymbol{e}' \\
&\geq \frac{4}{5}\|\boldsymbol{e}-\boldsymbol{e}'\|_2^2 - (\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') - \frac{1}{5}(\|\boldsymbol{e}\|_2^2 + \|\boldsymbol{e}'\|_2^2) \\
&= \frac{3}{5}(\|\boldsymbol{e}\|_2^2 + \|\boldsymbol{e}'\|_2^2) - (\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') - \frac{8}{5}\boldsymbol{e}^\top\boldsymbol{e}'.
\end{aligned} \tag{A2.15}$$

Define random events

$$\mathcal{E}_4 := \left\{(\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') \leq \frac{1}{5}\|\boldsymbol{e}\|_2^2\right\}, \quad \mathcal{E}_5 := \left\{\boldsymbol{e}^\top\boldsymbol{e}' \leq \frac{1}{8}\|\boldsymbol{e}\|_2^2\right\}.$$

For $\mathcal{E}_4$, we have

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_4^c\right) &\leq \mathbb{P}\left(\mathcal{E}_4^c \;\&\; \|\boldsymbol{e}\|_2^2 \geq n/2\right) + \mathbb{P}\left(\|\boldsymbol{e}\|_2^2 < n/2\right) \\
&= \mathbb{P}\left(\left\{(\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') > \frac{1}{5}\|\boldsymbol{e}\|_2^2\right\} \;\&\; \|\boldsymbol{e}\|_2^2 \geq n/2\right) + \mathbb{P}\left(\|\boldsymbol{e}\|_2^2 < n/2\right) \\
&\leq \mathbb{P}\left(\left\{(\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') > \frac{n}{10}\right\} \;\&\; \|\boldsymbol{e}\|_2^2 \geq n/2\right) + \mathbb{P}\left(\|\boldsymbol{e}\|_2^2 < n/2\right) \\
&\leq \mathbb{P}\left(\left\{(\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') > \frac{n}{10}\right\}\right) + \mathbb{P}\left(\|\boldsymbol{e}\|_2^2 < n/2\right)
\end{aligned}$$

Then using Lemmas A8 and A9, we have that the event $\mathcal{E}_4$ holds with probability converging to 1.

For $\mathcal{E}_5$, using that all the e_ie_i''s are i.i.d. random variables with $\mathbb{E}[|e_ie_i'|] = \mathbb{E}[|e_i|]\mathbb{E}[|e_i'|] < \infty$, we have $\boldsymbol{e}^\top\boldsymbol{e}'/n \to 0$ in probability; thus using a similar argument as $\mathcal{E}_4$, we have $\mathcal{E}_5$ holds with probability converging to 1.

Now working on the event $\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_5$ (which, as shown above, occurs with probability converging to 1), we have from (A2.15) and also the definitions of $\mathcal{E}_4$ and $\mathcal{E}_5$ that

$$\boldsymbol{e}^\top(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)\boldsymbol{e} + \boldsymbol{e}'^\top(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)\boldsymbol{e}' \geq \frac{1}{5}(\|\boldsymbol{e}\|_2^2 + \|\boldsymbol{e}'\|_2^2).$$

In other words, with probability converging to zero,

$$\underbrace{\boldsymbol{e}^\top\left(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k - \frac{1}{5}\boldsymbol{I}\right)\boldsymbol{e}}_{\mathrm{I}} + \underbrace{\boldsymbol{e}'^\top\left(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k - \frac{1}{5}\boldsymbol{I}\right)\boldsymbol{e}'}_{\mathrm{I}'} < 0.$$

Since I and I′ are two i.i.d. random variables, we have using their independence and identically distributed property that

$$\mathbb{P}(\mathrm{I} < 0) = \sqrt{\mathbb{P}(\mathrm{I} < 0)\mathbb{P}(\mathrm{I}' < 0)} = \sqrt{\mathbb{P}(\mathrm{I} < 0, \mathrm{I}' < 0)} \leq \sqrt{\mathbb{P}(\mathrm{I} + \mathrm{I}' < 0)} \to 0,$$

which proves (A2.13). In light of this and our control of (A2.12), we prove the desired result. □

A2.4 Auxiliary lemmas

Lemma A9. *Let $w_1, \ldots, w_n$ be a sequence of i.i.d. random variables from some distribution $\mathbb{P}_w$. Then if $\mathbb{P}_w \in \mathcal{D}_1 \cup \mathcal{D}_2$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{i=1}^n w_i^2 \geq \frac{1}{2}n\right) = 1.$$

Proof. We first consider $\mathbb{P}_w \in \mathcal{D}_2$. From standard results of weak law of large number, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 \geq \frac{1}{2}\right) = 1.$$

We next consider $\mathbb{P}_w \in \mathcal{D}_1$. First, apparently we have that either $\mathbb{E}[w_1^2] = \infty$ or that $\mathbb{E}[w_1^2]$ is finite and satisfy $\mathbb{E}[w_1^2] \geq (\mathbb{E}[|w_1|])^2 \geq 1$. In either of the two cases, we have that there exists a threshold τ such that $\mathbb{E}[w_1^2 \mathbb{1}(|w_1| \leq \tau)] = 1$.

Let $\tilde{w}_i := w_i \mathbb{1}(|w_1| \leq \tau)$. Then by again standard results of weak law of large number, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{i=1}^{n} \tilde{w}_i^2 \geq \frac{1}{2}n\right) = 1,$$

and the desired result is a direct consequence of that almost surely, $w_i^2 \geq \tilde{w}_i^2$. □

Lemma A10. *Given a constant $B > 0$ and a random variable V with $\mathbb{E}[|V|] < \infty$. Then*

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(|V| \geq Bi) \leq \mathbb{E}\left[\frac{|V|}{B}\right] < \infty.$$

Proof. For any integer n,

$$\sum_{i=1}^{n} \mathbb{P}(|V| \geq Bi) = \sum_{i=1}^{n} \mathbb{P}\left(\frac{|V|}{B} \geq i\right) \leq \int_{x\geq 0} \mathbb{P}\left(\frac{|V|}{B} > x\right) dx = \mathbb{E}\left[\frac{|V|}{B}\right] < \infty.$$

□

Lemma A11. *(Van Thanh [2006, Theorem 3]) Let $a_{n,i}$ $(i, n = 1, \ldots, \infty)$ be a deterministic array with $\lim_{n\to\infty} \sum_{i=1}^{n} |a_{n,i}|^2 = 0$. Let V_i $(i = 1, \ldots, \infty)$ be a sequence of independent random variables satisfying that*

$$\lim_{a\to\infty} \sup_{n\geq 1} \sum_{i=1}^{n} |a_{n,i}| \mathbb{E}[|V_i| \mathbb{1}(|V_i| > a)] = 0.$$

Then

$$\lim_{m\to\infty} \mathbb{E}\left[\left|\sum_{i=1}^{m} a_{m,i} V_i - \mathbb{E}[V_i]\right|\right] = 0.$$

Lemma A12. (Borel-Cantelli Lemma [Durrett, 2019, Theorem 2.3.1]) *Let $\mathcal{E}_1, \ldots$ be a sequence of random events. If $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(\mathcal{E}_i) < \infty$, then*

$$\mathbb{P}(\mathcal{E}_n \text{ i.o.}) = 0,$$

where $\mathcal{E}_n$ i.o. stands for $\mathcal{E}_n$ occurs infinitely often.

Lemma A13. *Let $\boldsymbol{P}$ be a uniformly random permutation matrix. Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a fixed $n \times n$ matrix. Then for any $i = 1, \ldots, n$, $\mathbb{E}[(\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}^\top)_{ii}] = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{M}_{jj}$.*

Proof. Let σ be the random permutation corresponding to $\boldsymbol{P}$, we have

$$\mathbb{E}[(\boldsymbol{P}^\top \boldsymbol{M}\boldsymbol{P})_{i,i}] = \mathbb{E}[\boldsymbol{M}_{\sigma(i),\sigma(i)}] = \frac{1}{n}\sum_j \boldsymbol{M}_{j,j},$$

where the second inequality is due to that $\sigma(i)$ can be viewed as a random variable that samples uniformly at random from the set $\{1, \ldots, n\}$. □

Lemma A14. *Consider the $\boldsymbol{P}$ in Lemma A13. Let $\boldsymbol{f} := (f_1, \ldots, f_n)^\top$ denote a random vector where all the f_i's are zero-mean independent random variables. Then for any $i \neq j$, $\mathbb{E}[(\boldsymbol{P}\boldsymbol{f}\boldsymbol{f}^\top \boldsymbol{P}^\top)_{ij}] = 0$.*

Proof. Let σ be the random permutation corresponding to $\boldsymbol{P}$. Then using that for any $i \neq j$, f_i and f_j are independent,

$$\mathbb{E}[(\boldsymbol{P}\boldsymbol{f}\boldsymbol{f}^\top \boldsymbol{P}^\top)_{ij}] = \mathbb{E}[f_{\sigma(i)} f_{\sigma(j)}] = 0,$$

which proves the desired result. □

Lemma A15. *Consider a symmetric positive semi-definite matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ and a permutation matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$, we have*

$$\mathrm{tr}[\boldsymbol{M}\boldsymbol{P}] \leq \mathrm{tr}[\boldsymbol{M}].$$

Proof. Using the positive semi-definiteness and symmetry of $\boldsymbol{M}$, we have for any i, j (i and j can be equal or unequal),

$$\boldsymbol{M}_{i,j} \leq \frac{\boldsymbol{M}_{i,i} + \boldsymbol{M}_{j,j}}{2}.$$

Let σ be the permutation associated with $\boldsymbol{P}$, we have

$$\mathrm{tr}[\boldsymbol{M}\boldsymbol{P}] = \sum_{i=1}^n \boldsymbol{M}_{i,\sigma(i)} \leq \sum_{i=1}^n \frac{\boldsymbol{M}_{i,i} + \boldsymbol{M}_{\sigma(i),\sigma(i)}}{2} = \mathrm{tr}[\boldsymbol{M}],$$

which proves the desired result. □

Lemma A16. *For any random variable f satisfying $\mathbb{E}[f] \leq 0$, there exists a random variable f' with $\mathbb{E}[f'] = 0$ and that*

$$\mathbb{P}\big(f' = f | f > 0\big) = 1 \quad \& \quad \mathbb{P}\big(f \leq f' \leq 0 | f \leq 0\big) = 1.$$

Proof. If $\mathbb{E}[f] = 0$, then $f' = f$ satisfies all the conditions. Thus we only need to consider the case with $\mathbb{E}[f] < 0$. define $f^+ = f\mathbb{1}(f > 0)$ and $f^- = -f\mathbb{1}(f \leq 0)$. Then $\mathbb{E}[f] = \mathbb{E}[f^+] - \mathbb{E}[f^-]$. As $\mathbb{E}[f] < 0$, we easily have that $\mathbb{E}[f^-] > \mathbb{E}[f^+] \geq 0$.

We now construct a Bernoulli random variable B satisfying that $\mathbb{P}(B = 1) = \frac{\mathbb{E}[f^+]}{\mathbb{E}[f^-]} \in [0, 1)$ and that $B \perp\!\!\!\perp f$. Then apparently, $f' = f^+ - Bf^-$ satisfies all the requirements, as

$$\begin{aligned}
\mathbb{E}[f'] &= \mathbb{E}[f^+] - B\mathbb{E}[f^-] = 0 \\
\mathbb{P}\big(f' = f | f > 0\big) &= \mathbb{P}\big((1-B)f^- = 0 | f^- = 0\big) = 1 \\
\mathbb{P}\big(f \leq f' \leq 0 | f \leq 0\big) &= \mathbb{P}\big((1-B)f^- \geq 0, f^+ = 0 | f^- \geq 0, f^+ = 0\big) = 1.
\end{aligned}$$

□

A2.5 Theoretical analysis of the algorithms

We will first show an lemma.

Lemma A17. *Consider a fixed matrix $\boldsymbol{M} \in R^{n\times n}$ with $n \geq 2$ and a fixed permutation matrix $\boldsymbol{P}_0 \in R^{n\times n}$ satisfying $\mathrm{tr}[\boldsymbol{P}_0] = 0$. Let $\tilde{\boldsymbol{P}} \in R^{n\times n}$ be a uniformly randomly sampled permutation matrix and define $\boldsymbol{P} := \tilde{\boldsymbol{P}}^{-1}\boldsymbol{P}_0\tilde{\boldsymbol{P}}$. Then for any $\delta > 0$, we have that*

$$\mathbb{P}\left(|\mathrm{tr}[\boldsymbol{M}\boldsymbol{P}]| \geq \frac{\sqrt{2\mathrm{tr}[\boldsymbol{M}\boldsymbol{M}^\top]}}{\sqrt{\delta}}\right) \leq \delta.$$

Proof. Let $\tilde{\sigma}$ be the random permutation corresponding to $\tilde{\boldsymbol{P}}$. Then we have that for any $\boldsymbol{P}_{u,v}$, $\boldsymbol{P}_{u,v} = 1$ if and only if $(\boldsymbol{P}_0)_{\tilde{\sigma}(u),\tilde{\sigma}(v)} = 1$. Now that since $\tilde{\sigma}$ is a uniformly random permutation, we have that $(\tilde{\sigma}(u), \tilde{\sigma}(v))$ is a pair that is uniformly at random drawn from the set $\{(i,j) \mid i \neq j \in \{1,\ldots,n\}\}$. From this, we have for any fixed (u,v),

$$\mathbb{P}(\boldsymbol{P}_{u,v} = 1) = \mathbb{P}((\boldsymbol{P}_0)_{\tilde{\sigma}(u),\tilde{\sigma}(v)} = 1) = \frac{n}{n^2-n} = \frac{1}{n-1},$$

and equivalently, $\mathbb{E}[\boldsymbol{P}^2_{u,v}] = \mathbb{E}[\boldsymbol{P}_{u,v}] = \frac{1}{n-1}$.

Notice also that since $\boldsymbol{P}$ is a random permutation matrix, we have that for any fixed u and any fixed $v_1 \neq v_2$, almost surely $\boldsymbol{P}_{u,v_1}\boldsymbol{P}_{u,v_2} = 0$.

Putting together, we have

$$\begin{aligned}\mathbb{E}[\mathrm{tr}[\boldsymbol{M}\boldsymbol{P}]^2] &= \mathbb{E}\left[\left(\sum_u\sum_v \boldsymbol{M}_{u,v}\boldsymbol{P}_{u,v}\right)^2\right] \leq n\sum_u \mathbb{E}\left[\left(\sum_v \boldsymbol{M}_{u,v}\boldsymbol{P}_{u,v}\right)^2\right] \\ &= n\sum_u\sum_v \mathbb{E}[\boldsymbol{M}^2_{u,v}\boldsymbol{P}^2_{u,v}] = \frac{n}{n-1}\sum_u\sum_v \boldsymbol{M}^2_{u,v} = \frac{n}{n-1}\mathrm{tr}[\boldsymbol{M}\boldsymbol{M}^\top] \leq 2\mathrm{tr}[\boldsymbol{M}\boldsymbol{M}^\top].\end{aligned}$$

From above, the desired result follows from Chebyshev's inequality. □

Proof of Proposition 1. Throughout the proof we only consider the case with number of iterations $T = 1$, and the case of $T \geq 2$ can be proven via analogous argument. For any $k_1, k_2 \in \{1,\ldots,K\}$, we have that by setting k_3 as the remainder after dividing $k_1 + k_2$ by $K+1$, we have that $\boldsymbol{P}_{k_3} = \boldsymbol{P}_{k_1}\boldsymbol{P}_{k_2}$. Which proves that the returned $\mathcal{P}_K$ satisfies Assumption 2.

We now define $\boldsymbol{P}_{0k}$ as a permutation matrix such that $(\boldsymbol{P}_{0k})_{u,v} = 1$ if and only if

$$\lceil\frac{u}{K+1}\rceil = \lceil\frac{v}{K+1}\rceil \;\&\; u - v \in \{k, k-(K+1)\};$$

and define $\boldsymbol{P}_\pi$ be the random permutation matrix associated with the permutation π, then we have that almost surely,

$$\boldsymbol{P}_k = \boldsymbol{P}_\pi^{-1}\boldsymbol{P}_{0k}\boldsymbol{P}_\pi.$$

We then have from Lemma A17 and

$$\mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top] = \mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top] = p$$

that for any k,

$$\mathbb{P}\left(|\mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}_k]| \geq \sqrt{2pK}\right) \leq \frac{1}{K^2}.$$

The desired result then follows by applying a union bound for all k.

Note that since Algorithm 1 returns with non-zero probability, there must exist a $\mathcal{P}_K$ that satisfies both assumptions. □

A3 Theoretical analysis of optimality

A3.1 Proof of Theorem 5

Without loss of generality we consider the scenario where $\beta = \beta^Z = 0$. Let $H_1(\tau)$ be the class of alternatives such that $|b| \geq \tau$, with τ to be specified later. Then using Neyman-Pearson lemma, we have that for any $(\boldsymbol{Z}, \boldsymbol{Y})$ in H_0 and any $(\boldsymbol{Z}', \boldsymbol{Y}')$ in $H_1(\tau)$,

$$\mathcal{R}_{t,\boldsymbol{X}}(\tau) \geq 1 - \mathbb{TV}(\mathbb{P}_{\boldsymbol{Y},\boldsymbol{Z}}, \mathbb{P}_{\boldsymbol{Y}',\boldsymbol{Z}'}).$$

Hence, the problem becomes constructing a $(\boldsymbol{Z}, \boldsymbol{Y})$ and $(\boldsymbol{Z}', \boldsymbol{Y}')$ belonging to H_0 and $H_1(\tau)$ such that their total variation distance is smaller than η.

We can do the following construction. First, we construct Z_i as i.i.d. binary random variables such that $\mathbb{P}(Z_i = n/\gamma) = \gamma/n$ and $\mathbb{P}(Z_i = -(1-\gamma/n)^{-1}) = 1 - \gamma/n$, where $\gamma = -\log(1-\eta)/2$, and without loss of generality, n is sufficiently larger such that $\gamma/n < 1$. Moreover, we construct Z_i' such that for each i, $Z_i = Z_i'$ almost surely.

We then construct $\varepsilon_i, \varepsilon_i'$ as i.i.d. Rademacher random variables that are independent from Z_i, Z_i'; and construct $\tilde{Z}_i$ as i.i.d replicates of Z_i which are independent from other randomness in the problem. Finally let $Y_i = b\tilde{Z}_i + \varepsilon_i$ and $Y_i' = bZ_i' + \varepsilon_i$ where $b = c_\eta n^{-t/(1+t)}$ for some constant $c_\eta > 0$ depending only on η such that $E[|Y_i|^{1+t}] = E[|Y_i'|^{1+t}] = 2$. Then it is straightforward that the distribution of Y_i is in $\mathcal{D}_{1+t}$, so that $(\boldsymbol{Y}, \boldsymbol{Z})$ and $(\boldsymbol{Y}', \boldsymbol{Z}')$ are feasible choices in H_0 and $H_1(\tau)$ respectively with $\tau := c_\eta n^{-t/(1+t)}$.

Using the above construction, we control their total variation distance as

$$\begin{aligned}
\mathbb{TV}(\mathbb{P}_{\boldsymbol{Y},\boldsymbol{Z}}, \mathbb{P}_{\boldsymbol{Y}',\boldsymbol{Z}'}) &= \sup_B \{\mathbb{P}\left((\boldsymbol{Y}, \boldsymbol{Z}) \in B\right) - \mathbb{P}\left((\boldsymbol{Y}', \boldsymbol{Z}') \in B\right)\} \\
&\leq \sup_B \{\mathbb{P}\left((\boldsymbol{Y}, \boldsymbol{Z}) \in B\right) - \mathbb{P}\left((\boldsymbol{Y}', \boldsymbol{Z}') \in B, (\boldsymbol{Y}, \boldsymbol{Z}) \in B\right)\} \\
&\leq \sup_B \mathbb{P}\left((\boldsymbol{Y}, \boldsymbol{Z}) \in B, (\boldsymbol{Y}', \boldsymbol{Z}') \notin B\right) \\
&\leq \mathbb{P}\left(\boldsymbol{Z} \neq \tilde{\boldsymbol{Z}}\right) \leq 1 - (1-\gamma/n)^{2n} \leq 1 - e^{-2\gamma} = \eta.
\end{aligned}$$

A3.2 Preliminary lemmas for Theorem 6

In this section, we invoke the notations introduced at the beginning of Section A2.

Lemma A18. *Consider the $\boldsymbol{M}$ in Lemma A3. If $b \geq \Delta n^{-1/2+\gamma}$ for some $\gamma > 0$, then for any fixed $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_2} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) = 0.$$

Proof. Following the same lines of proof as Lemma A3, we have for any $\delta > 0$,

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top M\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) \le \frac{16}{\delta^2 b^2 n} + \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 < \frac{n}{16}\right).$$

Combining the above with Lemma A24 yields the desired result. □

Lemma A19. *Let $\boldsymbol{g} := (g_1, \ldots, g_n)^\top$ be a n-dimensional vector satisfying $\|\boldsymbol{g}\|_2^2 \le c_g n^{\frac{1-t}{1+t}}$ for some constant $c_g > 0$. Then if $b \ge \Delta n^{-\frac{t}{1+t}+\gamma}$ for some $\gamma > 0$, we have that for any fixed $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) = 0$$

Proof. We have almost surely

$$\frac{|\boldsymbol{w}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} \le \frac{\|\boldsymbol{g}\|_2}{b\|\boldsymbol{w}\|_2},$$

whence

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{g}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) \le \mathbb{P}\left(\frac{\|\boldsymbol{g}\|_2}{b\|\boldsymbol{w}\|_2} > \delta\right).$$

From above and Lemma A24, we obtain our desired result.

□

Lemma A20. *Consider the $\boldsymbol{M}$ in Lemma A3 and let $t \in (0,1), B > 0$ be given. Then if $b \ge \Delta n^{-\frac{t}{1+t}+\gamma}$ for some constants $\gamma, \Delta > 0$, we have that for any fixed $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\ge,-B}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) = 0.$$

Proof. Let t_1 denote a constant in $(-1, t)$ such that

$$\frac{1-t_1}{1+t_1} - \frac{1-t}{1+t} \le \gamma.$$

Let $f_i := \xi_i \mathbb{1}(|\xi_i| \le Bi^{\frac{1}{1+t_1}})$ denote a truncated random variable of ξ_i. ALso, we construct f_i' as in Lemma A16. Moreover, we write $\boldsymbol{f} := (f_1, \ldots, f_n)^\top$ and $\boldsymbol{f}' := (f_1', \ldots, f_n')^\top$. Then following the same derivations as in Lemma A5, we have

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} > \delta\right) \le \underbrace{\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{f}'|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \|\boldsymbol{w}\|_2^2 > \frac{n}{16}\right)}_{=:\mathrm{I}} + \underbrace{\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{f}-\boldsymbol{f}')|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \|\boldsymbol{w}\|_2^2 > \frac{n}{16}\right)}_{=:\mathrm{II}}$$
$$+ \underbrace{\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \|\boldsymbol{w}\|_2^2 > \frac{n}{16}\right)}_{\mathrm{III}} + \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 \le \frac{n}{16}\right).$$

From Lemma A24, we have that $\sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}_{\zeta,v}} \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 \le n/16\right) \to 0$ as $n \to \infty$, which bounds the last term. For I, using analogous analysis as term I in Lemma A5, we have that there exists a universal constant $c > 0$ such that for any feasible choices of $\mathbb{P}_w, \mathbb{P}_\xi$,

$$\mathrm{I} \le \frac{cB^{1-t}}{\delta^2} \frac{n^{\frac{1-t_1}{1+t_1}}}{b^2 n} \le \frac{cB^{1-t}}{\delta^2 \Delta^2} \frac{1}{n^\gamma}.$$

This implies that

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\geq,-B}} \mathrm{I} = 0.$$

We second consider II. Notice that again

$$\mathrm{II} \leq \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{f} - \boldsymbol{f}' - \mathbb{E}[\boldsymbol{f} - \boldsymbol{f}'])|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{6} \mid \|\boldsymbol{w}\|_2^2 > \frac{n}{16} \right) + \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}\mathbb{E}[\boldsymbol{f} - \boldsymbol{f}']|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{6} \mid \|\boldsymbol{w}\|_2^2 > \frac{n}{16} \right)$$
$$=: \mathrm{II}_1 + \mathrm{II}_2.$$

For II_1, as $f_i - f_i'$ is bounded between $[-B, B]$, using the same analysis as in Lemma A18, we have $\sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\geq,-B}} \mathrm{II}_1 \to 0$. For II_2, observe first that using the same analysis as term II_2 in Lemma A5,

$$\|\mathbb{E}[\boldsymbol{f} - \boldsymbol{f}']\|_2^2 \leq n^{\frac{1-t}{1+t}} \left(\frac{\mathbb{E}[\|\xi\|^{1+t_1}]}{B^{1+t_1}}\right)^{\frac{2t}{1+t}}.$$

Then applying Lemma A19 with $\boldsymbol{P}\mathbb{E}[\boldsymbol{f} - \boldsymbol{f}']$ as $\boldsymbol{g}$ and noticing also Lemma A24, we have

$$\sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\geq,-B}} \mathrm{II}_2 \to 0$$

as $n \to \infty$. Putting together yields $\sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\geq,-B}} \mathrm{II} \to 0$.

We third consider III. From Lemma A10, we have that for any $N > 0$,

$$\sum_{i=N}^{\infty} \mathbb{P}(f_i \neq \xi_i) = \sum_{i=N}^{\infty} \mathbb{P}(|\xi_i|^{1+t_1} > B^{1+t_1} i) \leq \mathbb{E}\left[\frac{|\xi_1|^{1+t_1}}{B^{1+t_1}} \mathbb{1}(|\xi_1|^{1+t_1} > B^{1+t_1} N) \right],$$

whence

$$\mathbb{E}\left[\frac{|\xi_1|^{1+t_1}}{B^{1+t_1}} \mathbb{1}(|\xi_1|^{1+t_1} > B^{1+t_1} N) \right] = \mathbb{E}\left[\frac{|\xi_1|^{1+t} |\xi_1|^{t_1 - t}}{B^{1+t_1}} \mathbb{1}(|\xi_1|^{1+t_1} > B^{1+t_1} N) \right]$$
$$\leq \mathbb{E}\left[\frac{|\xi_1|^{1+t}}{B^{1+t_1}} \mathbb{1}(|\xi_1|^{1+t_1} > B^{1+t_1} N) \right] \cdot B^{t_1 - t} N^{\frac{t_1 - t}{1+t_1}} \leq 2 \frac{N^{\frac{t_1 - t}{1+t_1}}}{B^{1+t}}.$$

From above, $\forall \eta > 0$ by choosing $N_\eta = \left(\eta B^{1+t}/6\right)^{\frac{1+t_1}{t_1 - t}}$, we have

$$\mathbb{P}(\exists k > N_\eta \text{ s.t. } f_k \neq \xi_k) \leq \sum_{i=N_\eta}^{\infty} \mathbb{P}(f_i \neq \xi_i) \leq \frac{\eta}{3}.$$

Further letting $C_\eta = (6N_\eta/\eta)^{\frac{1}{1+t}}$, we have

$$\mathbb{P}(\exists \ell \leq N_\eta, \text{ s.t. } |\boldsymbol{\xi}_\ell| > C_\eta) \leq N_\eta \mathbb{P}(|\boldsymbol{\xi}_1| > C_\eta) \leq \frac{2N_\eta}{C_\eta^{1+t}} = \frac{\eta}{3}.$$

Putting together, we have that for the event $\mathcal{E} := \{\forall k > N_\eta, f_k = \xi_k, \forall \ell \le N_\eta, |\xi_i| < C_\eta\}$, $\mathbb{P}(\mathcal{E}^c) \le \frac{2\eta}{3}$, which gives us that

$$\begin{aligned}\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3}\right) &\le \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\ &\le \frac{2\eta}{3} + \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \mathcal{E}\right).\end{aligned}$$

Under the random event $\mathcal{E}$, using Cauchy-Schwartz inequality, we have

$$\begin{aligned}|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})| &\le \|\boldsymbol{w}\|_2 \|\boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})\|_2 \le \|\boldsymbol{w}\|_2 \|\boldsymbol{\xi}-\boldsymbol{f}\|_2 = \|\boldsymbol{w}\|_2 \sqrt{\sum_{i=1}^{N_\eta} \xi_i^2 \mathbb{1}(|\xi_i|^{1+t} > B^{1+t} i)} \\ &\le \|\boldsymbol{w}\|_2 \sqrt{N_\eta C_\eta}.\end{aligned}$$

Putting back yields

$$\mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3}\right) \le \frac{2\eta}{3} + \mathbb{P}\left(\frac{C_\eta \sqrt{N_\eta}\|\boldsymbol{w}\|_2}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3} \mid \mathcal{E}\right) = \frac{2\eta}{3} + \mathbb{P}\left(\frac{C_\eta \sqrt{N_\eta}}{b\|\boldsymbol{w}\|_2} > \frac{\delta}{3}\right),$$

then by lemma A24, we have

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\ge,-B}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3}\right) \\ \le &\frac{2\eta}{3} + \limsup_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\ge,-B}} \mathbb{P}\left(\frac{C_\eta \sqrt{N_\eta}}{b\|\boldsymbol{w}\|_2} > \frac{\delta}{3}\right) = \frac{2\eta}{3}.\end{aligned}$$

As η is arbitrary, we have as $n \to \infty$,

$$\sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\ge,-B}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{P}(\boldsymbol{\xi}-\boldsymbol{f})|}{b\|\boldsymbol{w}\|_2^2} > \frac{\delta}{3}\right) \to 0$$

In light of our control of I – III, our desired result follows. □

Lemma A21. *Consider the $\boldsymbol{M}$ in Lemma A3; let $t \in (0, 1]$ be given and assume that $b \ge n^{-\frac{t}{1+t}+\gamma}$. Then for any fixed $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\xi \in \mathcal{D}_{1+t} \cap \tilde{\mathcal{D}}_{\ge,-B}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\xi}|}{b\|\boldsymbol{w}\|_2^2} \le \delta\right) = 1.$$

Proof. When $t = 1$, the result follows from Lemma A18. Otherwise, we apply the decomposition

$$\xi_i = (\xi_i \mathbb{1}(\xi_i \ge 0) - \mathbb{E}[\xi_i \mathbb{1}(\xi_i \ge 0)]) - ((-\xi_i)\mathbb{1}(\xi_i < 0) - \mathbb{E}[(-\xi_i)\mathbb{1}(\xi_i < 0)]) =: \xi_{1,i} - \xi_{2,i};$$

and define $\boldsymbol{\xi}_1 := (\xi_{1,1}, \ldots, \xi_{1,n})^\top$, $\boldsymbol{\xi}_2 := (\xi_{2,1}, \ldots, \xi_{2,n})^\top$. Then our desired result follows by applying Lemma A20 but with $\boldsymbol{\xi}$ replaced by $\boldsymbol{\xi}_1$ or $\boldsymbol{\xi}_2$ and taking a union bound. □

Lemma A22. *Let $a_{n,i}$ $(i, n = 1, \ldots, \infty)$ be a deterministic array with $\sum_{i=1}^{n} |a_{n,i}|^2 \leq \frac{4}{n}$. Let V_i $(i = 1, \ldots, \infty)$ be a sequence of independent random variables obeying the law $\mathbb{P}_{V_i}$. Then for any $\gamma > 0$,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_{V_1},\ldots,\mathbb{P}_{V_n} \in D_{1+\gamma}} \mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i - \mathbb{E}[V_i])\right|\right] = 0.$$

Proof. Let $a = n^{\frac{1}{2(\gamma+1)}}$; define

$$V_i' = V_i \mathbb{1}(|V_i| > a), \quad V_i'' = V_i \mathbb{1}(|V_i| \leq a).$$

We first have

$$\begin{aligned}
\sum_{i=1}^{n} |a_{n,i}| \mathbb{E}[|V_i'|] &= \sum_{i=1}^{n} |a_{n,i}| \mathbb{E}[|V_i| \mathbb{1}(|V_i| > a)] = \sum_{i=1}^{n} |a_{n,i}| \mathbb{E}[|V_i|^{1+\gamma} |V_i|^{-\gamma} \mathbb{1}(|V_i| > a)] \\
&\leq a^{-\gamma} \sum_{i=1}^{n} |a_{n,i}| \mathbb{E}[|V_i|^{1+\gamma} \mathbb{1}(|V_i| > a)] \leq 2a^{-\gamma} \sum_{i=1}^{n} |a_{n,i}| \\
&\overset{(i)}{\leq} 2a^{-\gamma} n^{1/2} \left(\sum_{i=1}^{n} |a_{n,i}|^2\right)^{1/2} \leq 4a^{-\gamma}.
\end{aligned}$$

where (i) uses Cauchy-Schwartz inequality. From above, we have

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i - \mathbb{E}[V_i])\right|\right] &= \mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i' - \mathbb{E}[V_i'] + V_i'' - \mathbb{E}[V_i''])\right|\right] \\
&\leq \mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i'' - \mathbb{E}[V_i''])\right|\right] + 2\sum_{i=1}^{n} |a_{n,i}| \mathbb{E}[|V_i'|] \\
&\leq \mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i'' - \mathbb{E}[V_i''])\right|\right] + 8a^{-\gamma}
\end{aligned}$$

To deal with the first summand on the right hand side of the above inequality, we apply Hölder's inequality to get that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i'' - \mathbb{E}[V_i''])\right|\right] \leq \left[\mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i'' - \mathbb{E}[V_i''])\right|^2\right]\right]^{1/2} \\
&= \left[\mathbb{E}\left[\sum_{i=1}^{n} a_{n,i}^2 (V_i'' - \mathbb{E}[V_i''])^2\right]\right]^{1/2} \leq \left[\mathbb{E}\left[\sum_{i=1}^{n} a_{n,i}^2 4a^2\right]\right]^{1/2} = 2a\left[\sum_{i=1}^{n} a_{n,i}^2\right]^{1/2} \leq \frac{4a}{\sqrt{n}}
\end{aligned}$$

Putting together, we have

$$\sup_{\mathbb{P}_{V_1},\ldots,\mathbb{P}_{V_n} \in D_{1+\gamma}} \mathbb{E}\left[\left|\sum_{i=1}^{n} a_{n,i}(V_i - \mathbb{E}[V_i])\right|\right] \leq 8a^{-\gamma} + \frac{4a}{\sqrt{n}} \leq 12n^{-\frac{\gamma}{2(\gamma+1)}},$$

which gives us the desired result. □

Lemma A23. *Consider the $\boldsymbol{P}$ as in Lemma A8. We have that for any fixed $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \in \mathcal{D}_{1+\nu}} \mathbb{P}(|\boldsymbol{w}^\top \boldsymbol{P}\boldsymbol{w}|/n > \delta) = 0.$$

.

Proof. Let $w_{1,1}, w_{1,2}, \ldots, w_{1,n}, \ldots$ and $w_{2,1}, w_{2,2}, \ldots, w_{2,n}, \ldots$ be two sequences of i.i.d. random variables from a distribution $\mathbb{P}_w$. Then apparently if $\mathbb{P}_w \in \mathcal{D}_{1+t}$, $1 \le \mathbb{E}[|w_{1,i}w_{2,i}|^{1+t}] \le 4$. Then using Lemma A22, we have from Markov's inequality that for any $\delta > 0$,

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \mathcal{D}_{1+t}} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n w_{1,i}w_{2,i}\right| > \delta\right) = 0.$$

The desired result then follows from the same lines of proof as in Lemma A8.

□

Lemma A24. *We have*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_w \in \tilde{\mathcal{D}}} \mathbb{P}\left(\|\boldsymbol{w}\|_2^2 < \frac{n}{16}\right) = 0.$$

Proof. Let $\tilde{w}_i := \frac{1}{2}\mathbb{1}\left(|w_i| \ge \frac{1}{2}\right)$, then $E[\tilde{w}_i^2] \ge \frac{1}{8}$. By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\sum_{i=1}^n (\tilde{w}_i^2 - E[\tilde{w}_i^2])\right| \ge \frac{n}{16}\right) \le \exp\left(-\frac{n}{128}\right).$$

In light of the above inequality and that almost surely, $|w_i| \ge |\tilde{w}_i|$, we obtain the desired result. □

A3.3 Theoretical analysis of (15)

Proof. Following the proof of Theorem 3, we only need to show that for any fixed j, k, for all $\delta > 0$,

$$\begin{aligned}
&\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu} \cap \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} > \delta\right) \to 0; \\
&\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu} \cap \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}|}{bn} > \delta\right) \to 0;
\end{aligned} \tag{A3.16}$$

and that,

$$\begin{aligned}
&\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu} \cap \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} < \frac{m}{2(4+m)}\right) \to 0; \\
&\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu} \cap \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} < \frac{m}{2(4+m)}\right) \to 0.
\end{aligned} \tag{A3.17}$$

To prove the first claim of (A3.16), since $\mathbb{P}_e \in \mathcal{D}_{2+\nu}$, using Lemma A22 yields

$$\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{E}\left[\left|\frac{1}{n}\|\boldsymbol{e}\|_2^2 - \mathbb{E}[\boldsymbol{e}_1^2]\right|\right] \to 0,$$

whence by Markov's inequality,

$$\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}\left(\|\boldsymbol{e}\|_2^2 > 2\mathbb{E}[\boldsymbol{e}_1^2]n\right) \to 0.$$

For $\mathbb{P}_e \in \mathcal{D}_{2+\nu}$, using Hölder's inequality, we have

$$\mathbb{E}[\boldsymbol{e}_1^2] \le (\mathbb{E}[|\boldsymbol{e}_1|^{2+\nu}])^{2/(2+\nu)} \le 2^{2/(2+\nu)}.$$

From the above two inequalities, we have the random event $\mathcal{E} := \{\|\boldsymbol{e}\|_2^2 \le 2^{(3+\nu)/(2+\nu)}n\}$ satisfies that $\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}(\mathcal{E}^c) \to 0$. Therefore, we can control the first inequality of (A3.16) via that

$$\begin{aligned}
\mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} \ge \delta\right) &\le \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} \ge \delta \mid \mathcal{E}\right)\mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \\
&\overset{(i)}{\le} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{2b\|\boldsymbol{e}\|_2^2} \ge \delta \mid \mathcal{E}\right)\mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \le \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{2b\|\boldsymbol{e}\|_2^2} \ge \delta\right) + \mathbb{P}(\mathcal{E}^c),
\end{aligned}$$

where for the inequality (i) we apply that we are under $\mathcal{E}$. Then as a direct consequence of Lemma A21, we prove the first claim of (A3.16). For the second claim of (A3.16), using that $\boldsymbol{P}_k\boldsymbol{\varepsilon} \overset{d}{=} \boldsymbol{\varepsilon}$, The result follows the same argument as the first claim of (A3.16).

In the rest of the proof we focus on proving the first statement of (A3.17), and the second statement can be prove via a similar argument. To prove this statement, we apply again the decomposition

$$\begin{aligned}
\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} &= \frac{\boldsymbol{e}^\top (\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top - \mathrm{diag}(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top))\boldsymbol{e} - \boldsymbol{e}^\top (\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k - \mathrm{diag}(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k))\boldsymbol{e}}{n} \\
&\quad + \frac{\boldsymbol{e}^\top \mathrm{diag}(\tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top)\boldsymbol{e} - \boldsymbol{e}^\top \mathrm{diag}(\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)\boldsymbol{e}}{n} \\
&=: \mathrm{I} + \mathrm{II},
\end{aligned}$$

where recall that for any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathrm{diag}(\boldsymbol{A})$ corresponds to the diagonal matrix such that all the diagonal elements are equal to the diagonal elements of $\boldsymbol{A}$.

For I, using the same lines of proof as the term I in Section A2.2, we have that for any constant $\delta > 0$,

$$\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}(|\mathrm{I}| < \delta) \le \frac{2^{(6+\nu)/(2+\nu)}}{n\delta^2}. \tag{A3.18}$$

For II, we apply the same lines of proof as the control of term II in Section A2.2, except that we replace Lemma A11 with Lemma A22. Putting together, we obtain the desired result. □

A3.4 Theoretical analysis of (16)

Proof. Following analogous argument as in the proof of Theorem 4, we tackle this problem via proving that for any $j, k \in \{1, \ldots, K\}$ and for all $\delta > 0$,

$$\begin{aligned}
&\sup_{\mathbb{P}_e \in \mathcal{D}_{1+\nu} \cap \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2} \ge \delta\right) \to 0; \\
&\sup_{\mathbb{P}_e \in \mathcal{D}_{1+\nu} \cap \tilde{\mathcal{D}}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2} \ge \delta\right) \to 0;
\end{aligned} \tag{A3.19}$$

and that

$$\begin{aligned}\sup_{\mathbb{P}_e \in \mathcal{D}_{1+\nu} \cap \tilde{\mathcal{D}}} \mathbb{P}\left(\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{\|\boldsymbol{e}\|^2} < \frac{1}{5} \right) \to 0; \\ \sup_{\mathbb{P}_e \in \mathcal{D}_{1+\nu} \cap \tilde{\mathcal{D}}} \mathbb{P}\left(\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{\|\boldsymbol{e}\|^2} < \frac{1}{5} \right) \to 0.\end{aligned} \tag{A3.20}$$

The first claim of (A3.19) directly follows Lemma A21. The second claim of (A3.19) uses Lemma A21 and that $\boldsymbol{P}_k \boldsymbol{\varepsilon} \overset{d}{=} \boldsymbol{\varepsilon}$. To prove (A3.20), recall the definition of $\mathcal{E}_1 - \mathcal{E}_5$, it remains to prove that

$$\lim_{n \to \infty} \sup_{\mathbb{P}_e \in \mathcal{D}_{1+\nu} \cap \tilde{\mathcal{D}}} \mathbb{P}(\mathcal{E}_1^c \cup \cdots \cup \mathcal{E}_5^c) = 0.$$

We can control $\mathcal{E}_1 - \mathcal{E}_5$ following the same lines of proof as in the proof of those events in Section A2.3, except that for $\mathcal{E}_2$ and $\mathcal{E}_3$ we replace Lemma A6 by Lemma A21; for $\mathcal{E}_4$, we replace Lemmas A8 and A9 by Lemmas A23 and A24 respectively; and for $\mathcal{E}_5$, we additionally control the uniform convergence of $|\boldsymbol{e}^\top \boldsymbol{e}'|/n$ with Lemma A22.

In light of our control of all the random events, the desired result follows. □