Residual permutation test for 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 linear models when the dimension of covariates p can be up to a constant fraction of sample size n. In this regime, an important topic is to propose tests with *finite-population valid* size control without requiring the noise to follow strong distributional assumptions. 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 rate-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, Candès and Plan, 2011; Zhang and Zhang, 2014; Barber and Candès, 2015; Chernozhukov et al., 2018; Bradic et al., 2019). In this paper, we consider the setting where 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 $H_1 : b \neq 0$.

Here, 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 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

[*]Equal contribution

test (Fisher, 1935) 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 $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ has *exchangeable components*:

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

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

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 a *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 instance, 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 large dimensions. In this paper, we consider the more challenging question of finite-population Type-I error control in setting where p is allowed to be of the same order of magnitude as n. We propose a *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 the high-dimensional regime of interest, 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, Nichols and Knutsson, 2016; Wang, Peng and Li, 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 regression coefficient testing under this dimensional regime 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 $t \in [0, 1]$, and that $n/p \geq 3 + m$ for some $m > 0$, our proposed test is asymptotically powerful whenever the coefficient b is of order at least $n^{-t/(1+t)}$. In proving this result, a crucial step is to establish a concentration bound for projected length of a random vector with independent heavy-tailed components. This concentration bound may be of independent interest for future research on statistical procedures with heavy-tailed noise, and is stated in Corollary 8. We also studied the minimax rate 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 rate-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, while ANOVA can be shown to have finite-population validity with *spherically symmetric noise*, our simulations show that the it can substantially violate the nominal size control under more general noise distributions. 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 summary, we make the following contributions in this work:

- We propose a new test that has finite population validity with fixed-design linear models and exchangeable noises whenever $p < n/2$.
- 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 rate 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 regression coefficient testing, permutation- and randomization-based tests 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 7, we provide power analysis of RPT and study its minimax rate optimality under some heavy-tailed assumptions. Finally, in Section 8 we provide numerical analysis. In Section 9, we end 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 $|a_n| \le C|b_n|$ 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 hypothesis testing of regression coefficients, permutation- and randomization-based hypothesis tests and heavy-tailed data analysis. In this section, we compare our research to works within each direction.

2.1 Hypothesis testing of regression coefficients

The most classical approach for testing the null hypothesis $b = 0$ is through the analysis of variance (ANOVA) test (Fisher, 1935). 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 when p diverges as a constant fraction of n. However, as we will discuss in Section 3, beyond Gaussianity and some other class of restrictive assumptions on ε, 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. Alternatives with less restrictive assumptions on dimension p were proposed in D'Haultfœuille and Tuvaandorj (2022) and Candes et al. (2018), where the authors proposed "stratified randomization test" and "conditional randomization test", respectively. Different from our test that is fixed design and allows arbitrary $\boldsymbol{X}$, these two works both assume random designs, where the former stipulates that rows of $\boldsymbol{X}$ must follow a discrete random distribution with a relatively small number of unique values and the latter assumes either knowledge or a sufficiently good estimator of the sampling distribution of rows of $\boldsymbol{X}$.

Besides finite-population validity, a less demanding criteria for coefficient test is the *asymptotic validity*. The idea of permutation or randomization have been heavily used to propose asymptotically valid test; see Section 2.2 for more details. In the high-dimensional regime where p is proportional or even much larger than n, debiased / desparsified Lasso was proposed to construct confidence intervals and perform coefficient tests (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014). 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 / desparsified Lasso is guaranteed to establish asymptotically valid p-value and confidence intervals for regression coefficients. We remark that the additional sparsity assumption on the regression coefficients allow for the dimension p to diverge at a much faster rate than n compared to asymptotic regime studied in the current paper. Other follow up studies 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- and randomization-based hypothesis tests

As also mentioned in the introduction section, our new method is based on permutation test (Pitman, 1937a,b, 1938). Application of permutation and related randomization techniques for statistical inference has a long history in statistics and econometrics (Fisher, 1935; Rubin, 1980; Rosenbaum, 1984; Romano, 1990; Kennedy, 1995; Rosenbaum, 2002; Canay, Romano and Shaikh, 2017; Young, 2019). 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.

For the task of regression coefficient testing, Freedman (1981) and Freedman and Lane (1983) proposed tests based on bootstrapped and permuted regression residuals respectively, and proved asymptotic size validity in a fixed dimension. DiCiccio and Romano (2017) considered a permutation test using the studentized partial correlation of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ given $\boldsymbol{X}$ and derived asymptotic size and power of the test in a fixed dimension setting. Toulis (2019) studied a test based on permuting the residuals of $\boldsymbol{Y}$ regression against $(\boldsymbol{Z}, \boldsymbol{X})$. More recently, Lei and Bickel (2021); D'Haultfœuille and Tuvaandorj (2022) used permutation test and its extensions to obtain exact size control for testing a single component or a subvector of regression coefficients.

Other related applications of permutation tests include sharp null hypothesis tests (Caughey, Dafoe and Miratrix, 2017; Caughey et al., 2021), instrumental variable tests (Imbens and Rosenbaum, 2005), and conditional independence tests (Berrett et al., 2020; Kim et al., 2021).

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, Zhou and Fan (2020), such strong tail condition may not be reasonable in real world applications, such as neuroimaging (Eklund, Nichols and Knutsson, 2016), gene expression analysis (Wang, Peng and Li, 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, Cesa-Bianchi and Lugosi, 2013; Lykouris, Mirrokni and Paes Leme, 2018; Lugosi and Mendelson, 2019; Sun, Zhou and Fan, 2020; Fan, Wang and Zhu, 2021). When 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, Li and Wang, 2017; Sun, Zhou and Fan, 2020; Pensia, Jog and Loh, 2020) and covariance matrix analysis (Loh and Tan, 2018; Fan, Wang and Zhu, 2021). The definition of "heavy-tail" may vary across different articles. Among all literature working with heavy-tailed noise, our assumptions are most similar to those in Sun, Zhou and Fan (2020); Bubeck, Cesa-Bianchi and Lugosi (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.

In the context of coefficient testing, few methods have been proposed that can work with heavy-tailed data. We fill this gap by providing statistical power guarantees of our constructed test in the presence of 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. In fact, although not explicitly stated in Fisher (1935), Fisher recognized that ANOVA's validity only requires the noise to be spherically symmetric instead of Gaussian (Stigler, 2016, pp. 163–164). We provide a slight generalization of this result in Lemma 1, which shows that ANOVA is valid when *either* the design *or* the noise is spherically symmetric, in the sense defined below.

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 $\phi_{\mathrm{anova}} \sim F_{1,n-p-1}$.*

For the sake of completeness we provide a proof of Lemma 1 in the Supplementary Material. 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. We consider the sizes of the ANOVA test at nominal levels $\alpha = 0.01, 0.005$. 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 their histogram in Figure 1(a)-(c). 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 increases 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",

				ANOVA		Naive	
n	p	X type	noise type	1%	0.5%	1%	0.5%
300	100	Gaussian	Gaussian	1.01	0.50	1.00	0.49
300	100	Gaussian	t_1	1.81	1.60	1.58	1.16
300	100	Gaussian	t_2	1.53	1.07	1.39	0.89
300	100	t_1	Gaussian	1.01	0.50	1.03	0.49
300	100	t_1	t_1	2.43	2.08	1.58	1.07
300	100	t_1	t_2	1.80	1.30	1.41	0.88
600	100	Gaussian	Gaussian	0.95	0.50	0.96	0.48
600	100	Gaussian	t_1	1.63	1.43	1.28	0.80
600	100	Gaussian	t_2	1.69	1.20	1.28	0.76
600	100	t_1	Gaussian	1.05	0.50	1.02	0.52
600	100	t_1	t_1	1.88	1.66	1.06	0.58
600	100	t_1	t_2	1.74	1.30	1.14	0.63
600	200	Gaussian	Gaussian	1.01	0.49	1.03	0.50
600	200	Gaussian	t_1	1.41	1.22	1.24	0.90
600	200	Gaussian	t_2	1.50	1.04	1.36	0.89
600	200	t_1	Gaussian	1.01	0.49	0.98	0.49
600	200	t_1	t_1	2.02	1.73	1.33	0.86
600	200	t_1	t_2	1.70	1.20	1.34	0.80

Table 1: Percentage of rejection of the ANOVA test and naive residual permutation test, estimated over 100000 Monte Carlo repetitions, for various noise distributions at nominal levels of $\alpha = 1\%$ and $\alpha = 0.5\%$. 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 for all entries are in the range of 0.02% to 0.05%.

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.

It is worth noting that when $\beta = 0$ in (1), we can easily construct a valid permutation test by comparing 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 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.

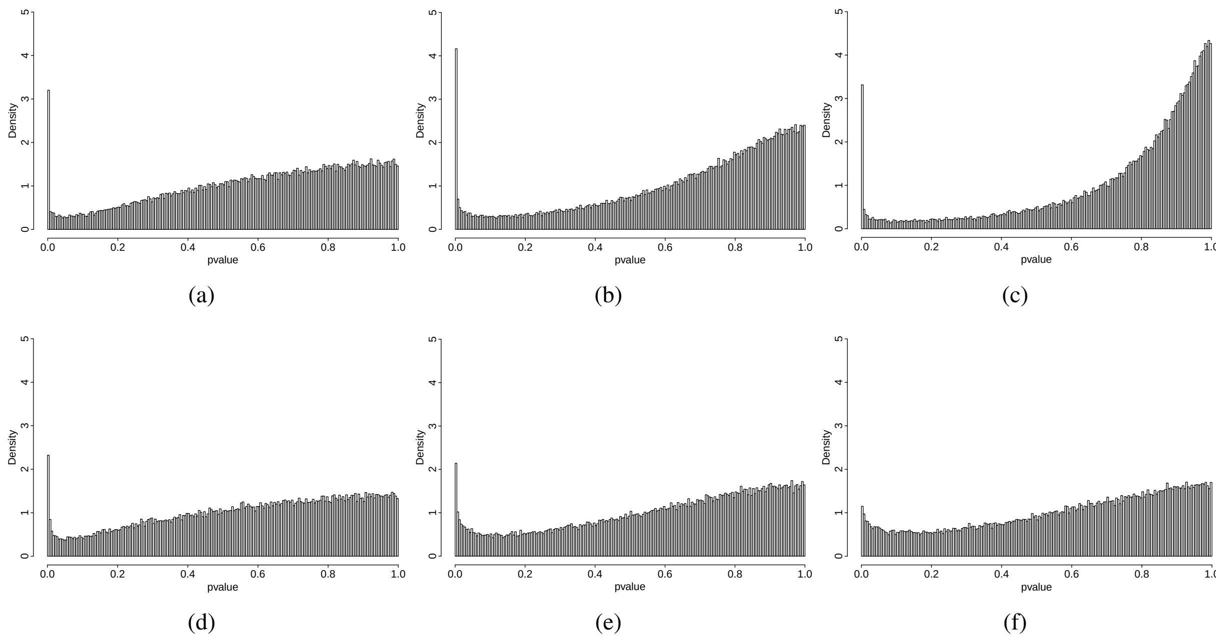


Figure 1: Histogram of p-values under the null for ANOVA test and naive residual permutation test from 100000 Monte-Carol replicates. The first line are the histograms of the ANOVA test under different specifications. Specifically, (a) is the result with Gaussian design, $n = 300, p = 100$ and $\boldsymbol{\varepsilon}$ 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).

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 $\alpha \in (0, 1)$, $\mathbb{P}(\phi_{\text{naive}} \le \alpha) \le \alpha$.

While Lemma 2 is slightly less stringent than Lemma 1, it still requires the spherical symmetry in distributions. 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) . Without the strong Gaussianity or spherically 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 $n/2 \le p < n$.

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 noise. In this section we describe a more refined test using the projected residuals $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\boldsymbol{e}}$, which we call the *residual permutation test* (RPT), and present its finite-population validity guarantee in Theorem 2. For intuition 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 $\boldsymbol{P}_0 = \boldsymbol{I} \in \mathcal{P}$ 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:

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

where T can be any bivariate 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

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

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.

[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 $\mathrm{span}(\boldsymbol{X})^\perp$.

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 $\boldsymbol{V}_0^\top \boldsymbol{\varepsilon}$. Interestingly, this goal can be achieved based on a further 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,

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

In light of this transformation, we have that under H_0, $\boldsymbol{V}_k \hat{\varepsilon} = \boldsymbol{V}_k \boldsymbol{V}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon} = \mathrm{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k \boldsymbol{\varepsilon})$, i.e., a projection of the noise vector $\boldsymbol{P}_k \boldsymbol{\varepsilon}$ onto the space $\mathrm{span}(\boldsymbol{V}_k)$, and equivalently, $\boldsymbol{V}_0 \hat{\varepsilon} = \mathrm{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$. 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,

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

and equivalently $\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\varepsilon} = \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}$ 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,

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

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, which is a key ingredient in the proof of Theorem 2, 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 $K \in \mathbb{N}$. Suppose that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is generated under model (1) where $\boldsymbol{\varepsilon}$ and $\boldsymbol{Z}$ 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 / desparsified 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 Theorem 3, for simplicity we assume that K is a fixed constant. In the Supplementary

Algorithm 1: Permutation set construction

Input: The number of permutation matrices K, the orthonormal matrix $\boldsymbol{V}_0 \in \mathbb{R}^{n\times(n-2p)}$ such that $\boldsymbol{V}_0^\top \boldsymbol{X} = 0$, the maximum 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 function $\sigma_k := \pi^{-1} \circ \tilde{\sigma}_k \circ \pi$, where $\circ$ denotes a composition of two functions and $\tilde{\sigma}_k$ is a permutation function such that

$$\tilde{\sigma}_k(i) := \begin{cases} i + k & \text{if } i \bmod (K+1) \le K + 1 - k \\ i - (K + 1 - k) & \text{otherwise,} \end{cases} \tag{9}$$

    and set $\boldsymbol{P}_k$ as the permutation matrix corresponding to σ_k.

  end

until *(i)* $|\mathrm{tr}[\boldsymbol{V}_0\boldsymbol{V}_0^\top \boldsymbol{P}_k]| \le \sqrt{2}Kp^{1/2}$ *for all* $k = 1, \ldots, K$ *or (ii) the number of iterations has reached its limit* T

Output: Set of permutation matrices $\mathcal{P}_K := \{\boldsymbol{P}_0 := \boldsymbol{I}, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}$ satisfying the criteria (i). When none of the $\mathcal{P}_K$'s comply, report the $\mathcal{P}_K$ with the smallest $\sum_{k=1}^K |\mathrm{tr}[\boldsymbol{V}_0\boldsymbol{V}_0^\top \boldsymbol{P}_k]|$.

Material we further provide an extension of Theorem 3 where we allow K to diverge with n. In particular, we show that for $t < 1$, RPT is still guaranteed to have non-trivial power whenever $K = O(n^{\frac{2t}{1+t}})$.

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.

Theorem 4. *Fix* $K \in \mathbb{N}$. *Suppose that* $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ *is generated under model* (1) *where* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{Z}$ *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, whilst 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 $\tilde{\boldsymbol{V}}_j$ 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 main Theorem 3 in Section 5.2.

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) when $n/p > 3 + m$ for some constant $m > 0$, 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). To understand this algorithm, notice that if we are just interested in Assumption 2, one simple way is to divide the n indices into $m := n/(K+1)$ ordered list of indices and perform cyclic permutation on each sub-list. Specifically, we first denote $S_1, \ldots, S_m$ as m ordered list of indices such that

$$(1, \ldots, n) := (\underbrace{1, \ldots, K+1}_{S_1}, \underbrace{K+2, \ldots, 2(K+1)}_{S_2}, \underbrace{(m-1)(K+1)+1, \ldots, m(K+1)}_{S_m}),$$

Then we define the $\tilde{\boldsymbol{P}}_k$ for $k \geq 1$ (or equivalently its permutation function $\tilde{\sigma}_k$) as that

$$(\tilde{\sigma}_k(1), \ldots, \tilde{\sigma}_k(n)) := (S_1^k, \ldots, S_m^k),$$

where each S_i^k is created via shifting all the elements in S_i by k places. Taking S_1^k for example, it means $S_1^k := (K+2-k, \ldots, K+1, 1, 2, \ldots, K+1-k)$. One can easily verify that the resulting set of permutation matrices $\tilde{\mathcal{P}}_K := \{\boldsymbol{I}, \tilde{\boldsymbol{P}}_1, \ldots, \tilde{\boldsymbol{P}}_K\}$ satisfies Assumption 2 since it is constructed by cyclic permutations[2]. However, since $\tilde{\mathcal{P}}_K$ is blind of $\boldsymbol{X}$, Assumption 4 may not hold. To overcome this challenge, in Algorithm 1 we apply an iterative algorithm where in each round, we set $\sigma_k := \pi^{-1} \circ \tilde{\sigma}_k \circ \pi$ for some random permutation π and loop until it reaches the number of rounds limit or the resulting $\mathcal{P}_K$ satisfies Assumption 4 (Step 6). This allows Algorithm 1 to still preserve Assumption 2, while being more adaptive to $\boldsymbol{X}$. In Proposition 1, we show that after doing T-th round of such iterations, Algorithm 1 is able to deliver a $\mathcal{P}_K$ satisfying the desired properties with probability at least $1 - \frac{1}{K^T}$.

Proposition 1. *Given K, T and assume that $n \geq 2$, 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. The maximum time complexities of Algorithms 1 and 2 are $O(TKn^2)$ and $O(TKn^2 + Kn^3)$ respectively, where T is the maximum number of iterations. The expected time complexities of the two algorithms are instead $O(Kn^2)$ and $O(Kn^3)$, respectively.

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, $|\hat{\boldsymbol{e}}^\top \boldsymbol{V}_0^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}| > |\hat{\boldsymbol{e}}^\top \boldsymbol{V}_0^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \hat{\boldsymbol{\varepsilon}}|$. To achieve this goal, we need to prove that

$$\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} = o_{\mathbb{P}}(1) \tag{10}$$

(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,

$$\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} \gtrsim 1 \quad \text{and} \quad \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} \gtrsim 1. \tag{11}$$

[2]Notice that the "$\tilde{\sigma}_k$" described here is exactly the same as the "$\tilde{\sigma}_k$" in (9)

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 $T \in \mathbb{N}$.

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

Apply Algorithm 1 with inputs K, T and $\boldsymbol{V}_0$ to generate K permutation matrices $\{\boldsymbol{P}_1, \dots, \boldsymbol{P}_K\}$.

for $k = 1, \dots, K$ **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 $\mathrm{span}(\tilde{\boldsymbol{V}}_k) \subseteq \mathrm{span}(\boldsymbol{V}_0) \cap \mathrm{span}(\boldsymbol{V}_k)$.

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

To prove (10), when $t = 1$, the result is straightforward from Chebyshev's inequality; hence the main challenge is to prove the case $t \in [0, 1)$. In Corollary 8, we establish a more general result, which characterizes the stochastic convergence of $|\boldsymbol{w}^\top \boldsymbol{\varepsilon}|$ where $\boldsymbol{w}$ is an arbitrary deterministic vector and $\boldsymbol{\varepsilon}$ can be heteroscedastic. We refer the readers to Section 6 for its statement as well as the intuitions for its proof.

Thanks to the bounded second order moment of e_i's, the analysis of (11) is simpler. Specially, by using a variant of weak law of large number we develop in this paper to control the weighted sum of e_i^2's and a Chebyshev's inequality to control the sum of cross terms $e_i e_j$'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.

6 Statistical power under broader classes of alternatives

In Theorems 3 and 4, for simplicity of illustration, we consider the class of alternative hypotheses where $\boldsymbol{Z}$ is a linear model and all the noises are i.i.d. In this section, we consider two relaxations of these assumptions. First, we assume that $\boldsymbol{Z}$ follows a linear model with respect to the covariates and all noises are heteroscedastic; second, we allow $\boldsymbol{Z}$ to have some nonlinearity, at the cost of slightly more restrictions on the degree of heteroscedasticity of ε_i's.

Assumption 5. $\boldsymbol{Z}$ *follows the model in* (3); *the random vectors* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{e}$ *are first* n *components of two independent infinite sequences of independent random variables* $\varepsilon_1, \varepsilon_2, \dots$ *and* $e_1, e_2, \dots$, *respectively. Suppose also that*

- *for some universal constants* $C_e, c_e > 0$, *we have* $\mathbb{E}[e_i^2] \le C_e$ *for all* $1 \le i < \infty$, *and*

$$\lim_{a\to\infty} \sup_{n\ge 1} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[e_i^2 \mathbb{1}(e_i^2 \ge a)] = 0 \quad \text{and} \quad \liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[e_i^2] > c_e. \tag{12}$$

- for some fixed $t \in [0,1]$ and some universal constant $C_\varepsilon > 0$, we have $\mathbb{E}[|\varepsilon_i|^{1+t}] \le C_\varepsilon$ for all i and given any fixed $B > 0$,

$$\sum_{i=1}^{\infty} \mathbb{P}\left(|\varepsilon_i|^{1+t} \ge Bi\right) < \infty. \tag{13}$$

Informally speaking, instead of requiring all noises to be i.i.d., Assumption 5 allows noises to be heteroscedastic, under certain restriction on the degree of heteroscedasticity of ε_i's and e_i's. To intuitively understand (13) and the first equation in (12), taking (13) for example, a sufficient condition for it to hold is that there exists a zero-meaned random variable ε_∞ satisfying that $\mathbb{E}[|\varepsilon_\infty|^{1+t}] < \infty$ and that for any $1 \le i < \infty$, $|\varepsilon_i| \preceq_d |\varepsilon_\infty|$, i.e., that $|\varepsilon_i|$ is stochastically dominanted by $|\varepsilon_\infty|$ uniformly for all ε_i's. When such ε_∞ exists, for any $n \ge 1$,

$$\sum_{i=1}^{n} \mathbb{P}\left(|\varepsilon_i|^{1+t} \ge Bi\right) \le \sum_{i=1}^{n} \mathbb{P}\left(|\varepsilon_\infty|^{1+t} \ge Bi\right) \le \int_0^\infty \mathbb{P}\left(\frac{|\varepsilon_\infty|^{1+t}}{B} \ge x\right) dx = \mathbb{E}\left[\frac{|\varepsilon_\infty|^{1+t}}{B}\right] < \infty,$$

which satisfies (13). Analogously, when there exists a zero-meaned random variable e_∞ with $\mathbb{E}[|e_\infty|^2] < \infty$ and $|e_\infty|$ stochastically dominates all $|e_i|$'s, we can also have

$$\sup_{n \ge 1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[e_i^2 \mathbb{1}(e_i^2 \ge a)] \le \mathbb{E}[e_\infty^2 \mathbb{1}(e_\infty^2 \ge a)],$$

which, from dominated convergence theorem, converges to zero as $a \to \infty$. Armed with Assumption 5, we have the following theorem on the power of RPT.

Theorem 5. *Fix $K \in \mathbb{N}$. Assume that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is generated under model (1) with $\boldsymbol{\varepsilon}$ and $\boldsymbol{Z}$ satisfying Assumption 5; $\mathcal{P}_K$ satisfies Assumption 4. In the asymptotic regime where b and p vary with n in a way such that $n > (3C_e/c_e + m)p$ for some constant $m > 0$ and b satisfies (8), we have $\lim_{n\to\infty} \mathbb{P}\left(\phi > \frac{1}{K+1}\right) = 0$.*

In the following, we show that when we are willing to impose slightly more restrictions on the degree of heterogeneity of ε_i's, we can still maintain the $n^{-\frac{t}{1+t}}$ rate even when the expectation of $\boldsymbol{Z}$ cannot be viewed as a linear function of $\boldsymbol{X}$.

Assumption 6. $\boldsymbol{Z}$ is generated according to $\boldsymbol{Z} = \boldsymbol{X}\beta^Z + \boldsymbol{h} + \boldsymbol{e}$, where $\boldsymbol{h}$ is an n-dimensional deterministic vector; $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ follow the same assumptions as the $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ in Assumption 5, with the addition that

$$\lim_{a\to\infty} \sup_{i \ge 1} \mathbb{E}[|\varepsilon_i|^{1+t} \mathbb{1}(|\varepsilon_i|^{1+t} > a)] = 0.$$

In Assumption 6, to alleviate the linearity requirement of $\boldsymbol{Z}$, we introduce an additional uniform constraint concerning the tails of ε_i's. It is worth noting that this new condition is satisfied when there exists a ε_∞ with $\mathbb{E}[|\varepsilon_\infty|^{1+t}] < \infty$ that stochastically dominates all ε_i's. Specifically, when such ε_∞ exists, then

$$\sup_{i \ge 1} \mathbb{E}[|\varepsilon_i|^{1+t} \mathbb{1}(|\varepsilon_i|^{1+t} > a)] \le \mathbb{E}[|\varepsilon_\infty|^{1+t} \mathbb{1}(|\varepsilon_\infty|^{1+t} > a)],$$

which, from dominated convergence theorem, converges to zero as $a \to \infty$.

Theorem 6. *Fix $K \in \mathbb{N}$. Assume that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is generated under model (1) with $\boldsymbol{\varepsilon}$ and $\boldsymbol{Z}$ satisfying Assumption 6; $\mathcal{P}_K$ satisfies Assumption 4. In the asymptotic regime where b, p and $\boldsymbol{h}$ vary with n in a way such that for some constants m, r with $m > 0, r < c_e$, $\limsup_{n\to\infty} \|\boldsymbol{h}\|_2^2/n \le r$, $n > (3C_e/(c_e - r) + m)p$ and b satisfies (8), we have $\lim_{n\to\infty} \mathbb{P}\big(\phi > \frac{1}{K+1}\big) = 0$.*

In fact, when $\boldsymbol{Z}$ and $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K$ are all deterministic and we keep the data generating model of $\boldsymbol{Y}$ as (1), following an analysis analogous to the proof of Theorem 6, we can prove that ϕ is still asymptotically powerful whenever

$$|b| = \Omega(z_n^{-1} n^{-\frac{t}{1+t}}) \text{ if } t < 1 \quad \text{or} \quad |b| = \omega(z_n^{-1} n^{-\frac{1}{2}}) \text{ if } t = 1, \tag{14}$$

where

$$z_n := \left(\frac{\|\boldsymbol{V}_0^\top \boldsymbol{Z}\|_2}{\sqrt{n}} \right)^{-1} \cdot \min_{1 \le j,k \le K} \min_{z \in \{0,1\}} \frac{\boldsymbol{Z}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{Z} + (-1)^z \boldsymbol{Z}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{Z}}{n}. \tag{15}$$

In other words, $\boldsymbol{Z}$ does not necessarily need to be random for RPT to have power. To formally describe the above intuition, we have the following corollary.

Corollary 7. *Fix $K \in \mathbb{N}$. Assume that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is generated under model (1) with $\boldsymbol{\varepsilon}$ as in Assumption 6 and $p < n/2$. $\boldsymbol{Z}, \mathcal{P}_K$ are deterministic such that $\|\boldsymbol{V}_0^\top \boldsymbol{Z}\|_2 > 0$ and $z_n > 0$ uniformly for all $n \ge 3$. In the asymptotic regime where b satisfies (14), we have $\lim_{n\to\infty} \mathbb{P}(\phi > \frac{1}{K+1}) = 0$.*

When $\boldsymbol{Z}$ satisfies the random model as prescribed in Assumption 6 and (n, p) is as in Theorem 6, with probability converging to 1, $z_n \asymp 1$, and the scale delivered by (14) and (8) coincide. In practice, one can choose $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K$ to maximize (15).

In order to prove Theorem 6, one needs to understand the rate of convergence of the term $|\boldsymbol{h}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|$. Based on our analysis of this term in the proof of Theorem 6, it is straightforward to get the following corollary, which characterizes the rate of convergence of $\boldsymbol{w}^\top \boldsymbol{\varepsilon}$ for arbitrary deterministic n-dimensional vector $\boldsymbol{w}$, which we believe is of independent interest:

Corollary 8. *Consider the $\boldsymbol{\varepsilon}$ as in Assumption 6 with $t \in [0, 1)$. Then for any fixed constant $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{\boldsymbol{w} \in \mathcal{S}^{n-1}} \mathbb{P}\left(|\boldsymbol{w}^\top \boldsymbol{\varepsilon}| > \delta n^{\frac{1-t}{2(1+t)}} \right) = 0,$$

where $\mathcal{S}^{n-1} := \{\boldsymbol{w} \in \mathbb{R}^n : \|\boldsymbol{w}\|_2 = 1\}$ is the $(n-1)$-sphere in the n-dimensional Euclidean space.

Informally then, Corollary 8 means that $|\boldsymbol{w}^\top \boldsymbol{\varepsilon}| = o_{\mathbb{P}}(\|\boldsymbol{w}\|_2 n^{\frac{1-t}{2(1+t)}})$ for any choice of n-dimensional unit vector $\boldsymbol{w}$. For example, one can even allow $\max_{1 \le i \le n} |w_i| / \|\boldsymbol{w}\|_2 \asymp 1$. This enables us to prove the rate of convergence of $\boldsymbol{h} \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}$ without any regularity condition on $\boldsymbol{X}$ or $\boldsymbol{h}$.

To prove Corollary 8 (or equivalently to find the rate of convergence of $\boldsymbol{h}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}$), the main challenge is to deal with the heavy-tailedness of ε_i's. We apply a truncation $f_i := \varepsilon_i \mathbb{1}(|\varepsilon_i| \ge Bi)$ and seek to control $\boldsymbol{w}^\top \boldsymbol{f}$ and $\boldsymbol{w}^\top (\boldsymbol{\varepsilon} - \boldsymbol{f})$ separately, where for simplicity we write $\boldsymbol{f} := (f_1, \ldots, f_n)^\top$. We seek to control $\boldsymbol{w}^\top \boldsymbol{f}$ via proving the following two convergence results (see the Supplementary Material for its proof):

- For any fixed $B > 0$, $\sup_{\boldsymbol{w} \in \mathcal{S}^{n-1}} \mathbb{E}[|\boldsymbol{w}^\top (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|^2] = o(n^{\frac{1-t}{1+t}})$;
- As $B \to \infty$, we have that $\sup_{n \ge 1} \|\mathbb{E}[\boldsymbol{f}]\|_2^2 / n^{\frac{1-t}{1+t}} \to 0$ (notice that here $\|\mathbb{E}[\boldsymbol{f}]\|_2^2$ is a function of B).

With the above results, it is straightforward that for any constant $\delta > 0$, by choosing the constant $B_\delta > 0$ sufficiently large, uniformly for all $n \ge 2$,

$$\sup_{\boldsymbol{w} \in \mathcal{S}^{n-1}} |\boldsymbol{w}^\top \mathbb{E}[\boldsymbol{f}_\delta]| \le \|\mathbb{E}[\boldsymbol{f}_\delta]\|_2 \le \frac{\delta}{2} \cdot n^{\frac{1-t}{2(1+t)}},$$

where we rewrite $\boldsymbol{f}$ as $\boldsymbol{f}_\delta$ to emphasize its dependence on B_δ. Moreover, by Chebyshev's inequality, we further have from the above convergence results that as $n \to \infty$,

$$\sup_{\boldsymbol{w} \in \mathcal{S}^{n-1}} \mathbb{P}\left(|\boldsymbol{w}^\top (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])| > \frac{\delta}{2} \cdot n^{\frac{1-t}{2(1+t)}} \right) \to 0. \tag{16}$$

Taking together, we control $\boldsymbol{w}\boldsymbol{f}_\delta$; and our only remaining job is to control the convergence of $\boldsymbol{w}^\top(\boldsymbol{\varepsilon} - \boldsymbol{f}_\delta)$, which we prove by an argument similar in spirit to Borel-Cantelli Lemma (Durrett, 2019).

7 Minimax rate optimality of coefficient tests

In this section, we investigate the minimax rate 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 $\mathbb{P}_\xi$,

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

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 $\eta_1 \le \mathbb{E}[|\xi|^t] \le \eta_2$ 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

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

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 nonasymptotic minimax lower bound for testing $H_0 : b = 0$ against $H_1 : b \ne 0$ in the presence of heavy-tailed noises.

Theorem 9. *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 $\boldsymbol{X}$,*

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

Theorem 9 shows that when entries of $\boldsymbol{\varepsilon}$ have finite $(1+t)$-th moment, the minimax separation rate 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 rate 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 9, 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 10. *Fix $K \in \mathbb{N}$. Assume that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is generated under model (1) with $\boldsymbol{\varepsilon}$ and $\boldsymbol{Z}$ satisfying 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,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}\left(\phi > \frac{1}{K+1}\right) = 0. \tag{17}$$

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

$$\lim_{n\to\infty} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \sup_{\mathbb{P}_e \in \mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}} \mathbb{P}\left(\phi > \frac{1}{K+1}\right) = 0. \tag{18}$$

In Theorem 10, 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 9. 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 rate-optimal hypothesis test in the minimax sense.

8 Numerical studies

8.1 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, the robust permutation test by DiCiccio and Romano (2017) (DR), the residual bootstrap method of Freedman (1981) (RB), the residual permutation approach of Freedman and Lane (1983) (FL), the conditional randomization test (CRT) of Candes et al. (2018), the residual randomization (RR) procedure of Toulis (2019), the desparsified Lasso approach for high-dimensional inference as implemented in the `hdi` R package (HDI) (Dezeure et al., 2015) and the cyclic permutation test of Lei and Bickel (2021) (CPT).

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.

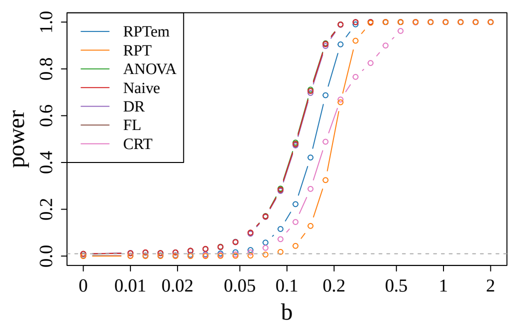
				RPT_{EM}		RPT		DR		FL		CRT	
n	p	$\boldsymbol{X}$	noise	1%	0.5%	1%	0.5%	1%	0.5%	1%	0.5%	1%	0.5%
300	100	$\mathcal{G}$	$\mathcal{G}$	0	0	0	0	0.98	0.5	0.99	0.52	0	0
300	100	$\mathcal{G}$	t_1	0.51	0.12	0.24	0	0.88	0.43	1.28	0.81	1.89	1.66
300	100	$\mathcal{G}$	t_2	0.14	0.02	0.04	0	0.67	0.3	1.23	0.64	0.53	0.37
300	100	t_1	$\mathcal{G}$	0	0	0	0	3.33	2.22	1.01	0.51	0	0
300	100	t_1	t_1	0.01	0	0	0	1.28	0.66	1.21	0.72	0.33	0.29
300	100	t_1	t_2	0	0	0	0	2.54	1.49	1.09	0.55	0	0
600	100	$\mathcal{G}$	$\mathcal{G}$	0.21	0.07	0.01	0	0.95	0.5	0.95	0.47	0	0
600	100	$\mathcal{G}$	t_1	0.73	0.43	0.48	0.28	0.92	0.48	1.09	0.59	1.68	1.49
600	100	$\mathcal{G}$	t_2	0.61	0.33	0.20	0.12	0.68	0.33	1.09	0.58	0.61	0.45
600	100	t_1	$\mathcal{G}$	0.23	0.07	0.01	0	3.95	2.65	0.93	0.47	0	0
600	100	t_1	t_1	0.13	0.03	0	0	1.37	0.72	1.04	0.54	0.25	0.22
600	100	t_1	t_2	0.10	0.03	0	0	3.33	2.04	1.05	0.52	0.01	0
600	200	$\mathcal{G}$	$\mathcal{G}$	0	0	0	0	1.04	0.53	1.02	0.53	0	0
600	200	$\mathcal{G}$	t_1	0.46	0.34	0.26	0.17	0.89	0.44	1.18	0.75	1.5	1.3
600	200	$\mathcal{G}$	t_2	0.12	0.10	0.04	0.03	0.68	0.33	1.2	0.67	0.49	0.34
600	200	t_1	$\mathcal{G}$	0	0	0	0	3.45	2.28	0.98	0.49	0	0
600	200	t_1	t_1	0.01	0	0	0	1.25	0.63	1.13	0.63	0.27	0.23
600	200	t_1	t_2	0	0	0	0	2.71	1.64	1.01	0.51	0	0

Table 2: Percentage of rejections of various tests under the null, estimated over 100000 Monte Carlo repetitions, for various noise distributions at nominal levels of $\alpha = 1\%$ and $\alpha = 0.5\%$. 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 "$\mathcal{G}$" stands for standard normal distribution. Percentage signs are omitted.

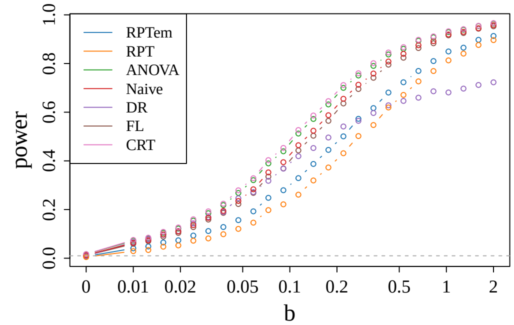
8.2 Numeric analysis of validity under the null

We start by analysing the validity of various tests under the null described in Section 8.1. We estimated the size of RPT, RPT_{EM}, DR, FL, CRT, RB, RR and HDI at nominal levels of 1% and 0.5% for $(n, p) \in \{(300, 100), (600, 100), (600, 200)\}$ (see Table A2 in the Supplementary Material for estimated size at 5% nominal level). RB, RR and HDI displayed more serious violation of the empirical sizes in these simulation settings (see Table A1 in the Supplementary Material). The results for the remaining procedures are summarised in Table 2. Notice that since the p-values of both ANOVA and the naive RPT 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.

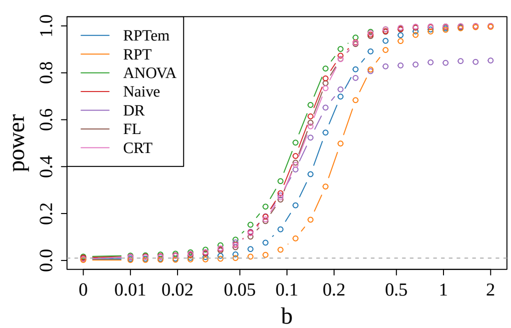
From Table 2, we see that DR has good size control when the design matrix $\boldsymbol{X}$ has Gaussian components and exceeds the nominal size levels when $\boldsymbol{X}$ is generated with t_1 components. FL performed the best when n/p is relatively large, consistent with the asymptotic validity of the test established in Freedman and Lane (1983), though with low n/p ratios and heavy-tailed noise, the empirical sizes can exceed the nominal level. CRT is conservative when components of $\boldsymbol{X}$ and the noise have the same distribution, but can violate the size control when the noise distributions have much heavier tails than that of components of $\boldsymbol{X}$. On the other hand, RPT exhibits valid size controls in all settings, which is consistent with our theoretical findings. More interestingly, the size of RPT_{EM} is also valid across all the simulation settings, even with heavy-tailed noises and heavy-tailed design. In Section 8.3, we further study the empirical power of RPT and RPT_{EM}.



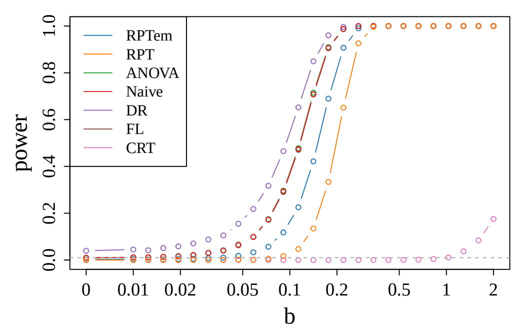
(a) Gaussian design, Gaussian noise



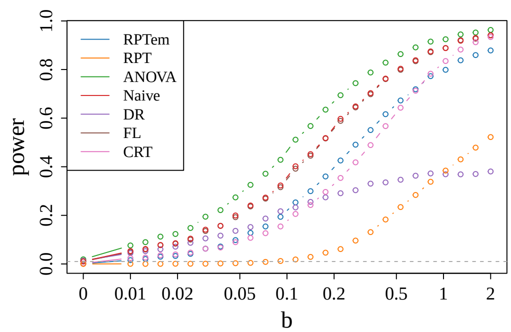
(b) Gaussian design, t_1 noise



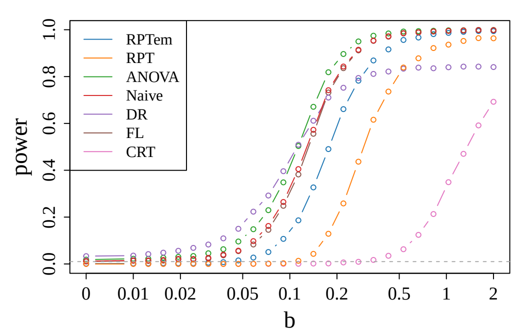
(c) Gaussian design, t_2 noise



(d) t_1 design, Gaussian noise



(e) t_1 design, t_1 noise



(f) t_1 design, t_2 noise

Figure 2: Power (proportion of rejections) with nominal level $\alpha = 0.01$ (represented by the horizontal dashed line) over 10000 replicates for $b = 0$ or on a logarithmic grid between 0.01 and 2. Here $\boldsymbol{X}, \boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ are generated according to various distribution types prescribed in the caption of each figure.

8.3 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 8.1. We set $n = 600$, $p = 100$ and vary the b in (1) for b equals to 0 or one of the 25 different values on an equally spaced logarithmic grid in the range of 0.01 to 2. 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 for RPT_{EM}, RPT, ANOVA, naive RPT, DR, FL and CRT over 10000 repetitions are displayed in Figure 2 (see also Figure A1 in the Supplementary Material for power curves of RB, RR and HDI).

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 ANOVA, the naive RPT and FL. 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, DR, FL and CRT, especially when design and noise are heavy-tailed, 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, RPT is nevertheless still guaranteed to reject the alternative with high probability given a signal size b not too much larger than the competitors. In addition, we observe that DR does not seem to have power converging to 1 as b increases for heavy-tailed noise, while the power of CRT is substantially reduced for heavy-tailed design distributions.

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 8.2, suggests RPT_{EM}, although being lack of theoretical support, can serve as a viable alternative of RPT in empirical analysis. We leave the theoretical investigations of RPT_{EM} as future work.

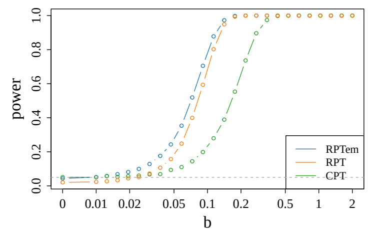
Finally, we compare RPT with the cyclic permutation test (CPT) proposed in Lei and Bickel (2021). As CPT is not well-defined for $n/p < 1/\alpha + 1$, we consider a relatively low dimensional setting where $n = 1000$, $p = 40$ and $\alpha = 0.05$. The data generation mechanism is the same as that described in Section 8.1. Figure 3 shows the power curves of RPT_{EM}, RPT and CPT under various design matrix and noise distributions. We see that all three methods are well-calibrated at 5% level when $b = 0$, with RPT slightly more conservative than CPT and RPT_{EM}. For all the settings, the power of RPT and RPT_{EM} converges to 1 faster than CPT, though CPT has higher rejection rate than RPT as b begins to diverge from zero.

9 Discussion

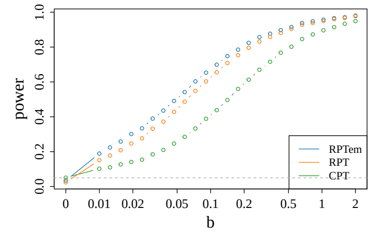
In this paper, we propose a new method for fixed design regression coefficient test with moderately high-dimensional covariates. RPT is a permutation-based approach that exploits the exchangeability of the noise terms to achieve finite-population validity control. Our approach uses the fact that the 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 analysis of ANOVA test, which is of independent interest for practitioners interested in ANOVA.

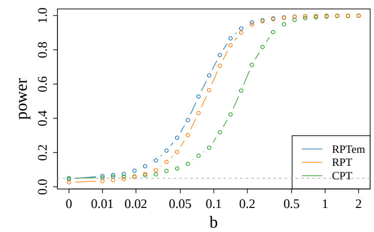
Distribution-free inference and test is an important topic in statistics research. In this paper, permutation test facilitates an important basis for construction of finite-population tests hypothesis tests with



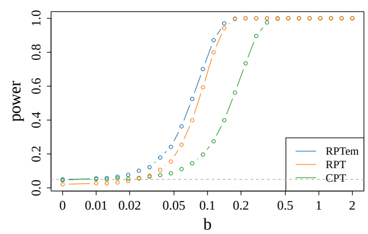
(a) Gaussian design, Gaussian noise



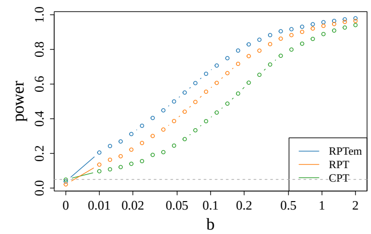
(b) Gaussian design, t_1 noise



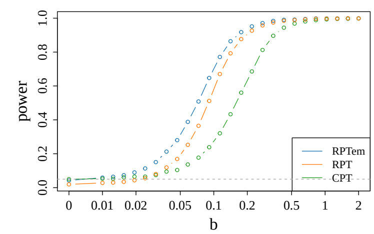
(c) Gaussian design, t_2 noise



(d) t_1 design, Gaussian noise



(e) t_1 design, t_1 noise



(f) t_1 design, t_2 noise

Figure 3: Power (proportion of rejections) with nominal level $\alpha = 0.05$ (represented by the horizontal dashed line) over 10000 replicates for $b = 0$ or on a logarithmic grid between 0.01 and 2. Here $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ are generated according to various distribution types prescribed in the caption of each figure.

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, Robins and Wasserman, 2013; Balasubramanian, Ho and Vovk, 2014; Romano, Patterson and Candes, 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

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

Balasubramanian, V., Ho, S.-S. and Vovk, V. (2014) *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Barber, R. F. and Candès, E. J. (2015) Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, **43**, 2055–2085.

Berrett, T. B., Wang, Y., Barber, R. F. and Samworth, R. J. (2020) The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **82**, 175–197.

Bradic, J., Chernozhukov, V., Newey, W. K. and Zhu, Y. (2019) Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*.

Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013) Bandits with heavy tail. *IEEE Transactions on Information Theory*, **59**, 7711–7717.

Canay, I. A., Romano, J. P. and Shaikh, A. M. (2017) Randomization tests under an approximate symmetry assumption. *Econometrica*, **85**, 1013–1030.

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

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

Caughey, D., Dafoe, A., Li, X. and Miratrix, L. (2021) Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. *arXiv preprint arXiv:2101.09195*.

Caughey, D., Dafoe, A. and Miratrix, L. (2017) Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. *arXiv preprint arXiv:1709.07339*.

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

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

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015) High-dimensional inference: confidence intervals, p-values and R-software hdi. *Statistical science*, 533–558.

D'Haultfœuille, X. and Tuvaandorj, P. (2022) A Robust Permutation Test for Subvector Inference in Linear Regressions. *arXiv preprint arXiv:2205.06713*.

DiCiccio, C. J. and Romano, J. P. (2017) Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, **112**, 1211–1220.

Doane, D. P. and Seward, L. E. (2016) *Applied statistics in business and economics, 5th*. Mcgraw-Hill.

Durrett, R. (2019) *Probability: theory and examples*, vol. 49. Cambridge university press.

Eklund, A., Nichols, T. E. and Knutsson, H. (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, **113**, 7900–7905.

Fan, J., Li, Q. and Wang, Y. (2017) 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**, 247–265.

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

Fisher, R. A. (1935) *The Design of Experiments*. Oliver and Boyd.

Freedman, D. and Lane, D. (1983) A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, **1**, 292–298.

Freedman, D. A. (1981) Bootstrapping regression models. *The Annals of Statistics*, **9**, 1218–1228.

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

Imbens, G. W. and Rosenbaum, P. R. (2005) Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **168**, 109–126.

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

Kennedy, F. E. (1995) Randomization tests in econometrics. *Journal of Business & Economic Statistics*, **13**, 85–94.

Kim, I., Neykov, M., Balakrishnan, S. and Wasserman, L. (2021) Local permutation tests for conditional independence. *arXiv preprint arXiv:2112.11666*.

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

Lei, J., Robins, J. and Wasserman, L. (2013) Distribution-free prediction sets. *Journal of the American Statistical Association*, **108**, 278–287.

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

Loh, P.-L. and Tan, X. L. (2018) High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, **12**, 1429–1467.

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

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

Lykouris, T., Mirrokni, V. and Paes Leme, R. (2018) Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 114–122.

Meinshausen, N. (2015) 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**, 923–945.

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

Pensia, A., Jog, V. and Loh, P.-L. (2020) Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*.

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

Pitman, E. J. (1937b) 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**, 225–232.

Pitman, E. J. (1938) Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, **29**, 322–335.

Romano, J. P. (1990) On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, **85**, 686–692.

Romano, Y., Patterson, E. and Candes, E. (2019) Conformalized quantile regression. *Advances in neural information processing systems*, **32**.

Rosenbaum, P. R. (1984) Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, **79**, 565–574.

Rosenbaum, P. R. (2002) Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, **17**, 286–327.

Rubin, D. B. (1980) Comment on Basu's randomization analysis of experimental data. *Journal of the American Statistical Association*, **75**, 591–593.

Shah, R. D. and Bühlmann, P. (2019) Double-estimation-friendly inference for high-dimensional misspecified models. *arXiv preprint arXiv:1909.10828*.

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

Stigler, S. M. (2016) *The seven pillars of statistical wisdom*. Harvard University Press.

Sun, Q., Zhou, W.-X. and Fan, J. (2020) Adaptive huber regression. *Journal of the American Statistical Association*, **115**, 254–265.

Toulis, P. (2019) Invariant Inference via Residual Randomization. *arXiv preprint arXiv:1908.04218*.

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

Wainwright, M. J. (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Wang, L. (2013) The ℓ_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, **120**, 135–151.

Wang, L., Peng, B. and Li, R. (2015) A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, **110**, 1658–1669.

Young, A. (2019) Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics*, **134**, 557–598.

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

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

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

Appendix A1 provides additional power analysis of RPT when K diverges with n.

Appendix A2 provides proof of all validity statements of the ANOVA test, naive RPT and RPT. It includes the proof of all the theoretical statements in Sections 3 and 4 and also the discussion of the equivalence between (5) and (6).

Appendix A3 provides a preliminary lemma for RPT's power analysis.

Appendix A4 provides proof of the statistical power results of RPT when K is fixed and noises are i.i.d. It includes the proof of the theoretical statements in Section 5.

Appendix A5 provides proof of the rest of the power results of RPT. In includes proof of the theoretical statements in Section 6 and Appendix A1.

Appendix A6 studies the minimax rate optimality of coefficient test with heavy-tailed noises. It includes proof of the theoretical statements in Section 7.

Appendix A7 provides additional numerical analysis.

Notations we define $\|\|_{\text{op}}$ as operator norm, $\|\|_2$ as ℓ_2-norm, $\|\|_F$ as Frobenius norm. We define $a_1/a_2 = \infty$ if $a_1 = a_2 = 0$ or $a_2 = 0$. Without loss of generality, we assume $b > 0$. Let $\mathbf{1}$ be an n dimensional vector with all entries equal to 1.

A1 Statistical power of RPT with a diverging K

In this section, we discuss the power of RPT when we allow K to diverge with n. We have the following theorem:

Theorem A1. *Suppose that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is generated under model (1) where $\boldsymbol{\varepsilon}$ and $\boldsymbol{Z}$ satisfy Assumption 3 and*

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

for some constants $t \in [0, 1]$ and $\kappa \geq 0$. Assume also that $\mathcal{P}_K$ satisfies Assumption 4. In the asymptotic regime where K varies with n with rate

$$K = O(n^{\frac{2\kappa}{2+\kappa}}) \text{ if } \kappa < 2 \quad \text{or} \quad K = o(n) \text{ if } \kappa \geq 2,$$

and p, b vary with n and K such that $n > (3+m)p + \min\{p, \sqrt{2p}K\}$ for some constant $m > 0$ and

$$|b| = \Omega(\sqrt{K} n^{-\frac{t}{1+t}}) \text{ if } t < 1 \quad \text{or} \quad |b| = \omega(\sqrt{K} n^{-\frac{1}{2}}) \text{ if } t = 1, \tag{A1.1}$$

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

When $K = o(n/\sqrt{p})$, we require n/p to be asymptotically larger than 3 to get the desired power. As K gets larger, the threshold becomes 4 instead. When K is a constant, then the power rate in (A1.1) matches the main result (8). As the size of K increases, we need more moments for e_i to maintain the rate (A1.1). In particular, when the fourth order moment exists for e_i, K can be as large as $o(n)$.

In the rest of this section, we discuss the power of RPT when data generating mechanism satisfies Assumption 6 and K is not a constant. Extending the proofs of Theorems 6 and A1, we can conclude that under Assumption 6, with additional constraints that

- uniformly for all i, $\mathbb{E}[|e_i|^{\kappa+2}] \le C_e$ for some constant $C_e > 0$;
- for any fixed constant $B > 0$,

$$\sum_{i=1}^{\infty} \mathbb{P}(|e_i^2 - \mathbb{E}[e_i^2]|^{1+\kappa/2} \ge Bi) < \infty;$$

- $\sup_{i\ge 1} \mathbb{E}[|e_i^2 - \mathbb{E}[e_i^2]|^{1+\kappa/2} \mathbb{1}(|e_i^2 - \mathbb{E}[e_i^2]|^{1+\kappa/2} \ge a)] \to 0$ as $a \to \infty$,

and that $n > (3C_e/(c_e - r) + m)p + C_e/(c_e - r) \cdot \min\{p, \sqrt{2p}K\}$ for some fixed constant $m > 0$, RPT is still asymptotically powerful when b and K scales as in Theorem A1. In other words, even with heteroscedastic noise or nonlinear $\boldsymbol{Z}$, RPT is still guaranteed to be asymptotically powerful with a diverging K.

A2 Proof of finite-population validity statements

A2.1 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,

$$\phi_{\text{anova}} = \frac{\|(\boldsymbol{I} - \mathrm{Proj}_{\boldsymbol{X}})(\boldsymbol{\varepsilon})\|_2^2 - \|(\boldsymbol{I} - \mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}})(\boldsymbol{\varepsilon})\|_2^2}{\|(\boldsymbol{I} - \mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}})(\boldsymbol{\varepsilon})\|_2^2/(n-p-1)} = \frac{\|(\mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}} - \mathrm{Proj}_{\boldsymbol{X}})(\boldsymbol{\xi})\|_2^2}{\|(\boldsymbol{I} - \mathrm{Proj}_{\boldsymbol{X},\boldsymbol{Z}})(\boldsymbol{\xi})\|_2^2/(n-p-1)}. \quad \text{(A2.2)}$$

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

By Cochran's theorem, we know that $\phi_{\text{anova}} \sim F_{1,n-p-1}$ 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 (A2.2)), we must have $\phi_{\text{anova}} \sim F_{1,n-p-1}$ for all spherically symmetric $\boldsymbol{\varepsilon}$ as desired.

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

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

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

A2.2 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

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

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}. □

A2.3 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 $\boldsymbol{a} \in \mathbb{R}^n$, $\text{Proj}_{\boldsymbol{U}}(\boldsymbol{a}) := \boldsymbol{UU}^\top\boldsymbol{a}$.

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 straightforward using that

$$\boldsymbol{W}^\top = \boldsymbol{W}^\top\boldsymbol{UU}^\top = \boldsymbol{W}^\top\boldsymbol{VV}^\top$$

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. □

A2.3.1 Proof of Lemma 3

Proof. Let $\xi_0, \ldots, \xi_K \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ 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 $(g(\boldsymbol{P}_{k'}\boldsymbol{\varepsilon}) : k' = 0, \ldots, K)$ in a decreasing order, with random tie-breaking. Also, observe that $R_k \ge \tilde{R}_k$. By Assumptions 1 we have $\boldsymbol{\varepsilon} \stackrel{\text{d}}{=} \boldsymbol{P}_k\boldsymbol{\varepsilon}$ for all k, hence

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

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

$$\mathbb{P}(R_k \le x) = \frac{1}{K+1} \sum_{k'=0}^{K} \mathbb{P}(R_{k'} \le x) \le \frac{1}{K+1} \sum_{k'=0}^{K} \mathbb{P}(\tilde{R}_{k'} \le x). \tag{A2.3}$$

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

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

By taking this back to (A2.3), we may further bound (A2.3) 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. □

A3 A preliminary lemma for power analysis

In this section, our main goal is to prove Lemma A5, which can be used to characterize the stochastic convergence of $\boldsymbol{e}\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{\varepsilon}$ or $\boldsymbol{h}\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{\varepsilon}$ in the proofs of Theorems 3, 4, 6 and A1.

Lemma A3. *Let $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K \in \mathbb{R}^{n\times n}$ be K deterministic matrices that varies with n and satisfies $\|\boldsymbol{M}_k\|_{\text{op}} \le 1$ for all $k = 1, \ldots, K$; let $\boldsymbol{w} := (w_1, \ldots, w_n)^\top$ be an n-dimensional deterministic vector that varies with n. Assume that $\boldsymbol{\varepsilon}$ satisfies Assumption 6 with $t = 1$. Then if $b = \omega\big(\sqrt{K}n^{-1/2}\big)$, we have that for any fixed $\delta, \gamma > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\exists 1 \le k \le K \text{ s.t. } \frac{|\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}|}{b\max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \delta\right) = 0.$$

Proof. We have for any $1 \le k \le K$,

$$\mathbb{E}[|\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}|^2] = \mathbb{E}[\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{M}_k^\top \boldsymbol{w}] \le C_\varepsilon \|\boldsymbol{w}\|_2^2,$$

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

$$\begin{aligned}&\mathbb{P}\left(\exists 1 \le k \le K \text{ s.t. } \frac{|\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}|}{b\max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \delta\right) \le \sum_{k=1}^K \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}|}{b\max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \delta\right)\\
&\le \sum_{k=1}^K \frac{\mathbb{E}[|\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}|^2]}{\delta^2 b^2 \max\{\|\boldsymbol{w}\|_2^4, \gamma^2 n^2\}} \le \frac{C_\varepsilon K}{\delta^2 \gamma b^2 n}.\end{aligned}$$

From above, the desired result follows from $K/b^2 n = o(1)$. □

Lemma A4. *Consider the $\boldsymbol{\varepsilon}$ in Assumption 6 with $t \in [0, 1)$; let $\boldsymbol{w} \in \mathbb{R}^n$ be an n-dimensional vector that varies with n, we have that for any fixed $B > 0$, there exists a sequence of positive real values $c_n \to 0$ that does not depend on $\boldsymbol{w}$ such that*

$$\sum_{i=1}^n \mathbb{E}[w_i^2 \varepsilon_i^2 \mathbb{1}(|\varepsilon_i| \le B i^{\frac{1}{1+t}})] \le c_n \cdot \|\boldsymbol{w}\|_2^2 n^{\frac{1-t}{1+t}}.$$

Moreover,

$$\lim_{B\to\infty} \sup_{n\ge 1} \left\{\sum_{i=1}^n \left(\mathbb{E}[\varepsilon_i \mathbb{1}(|\varepsilon_i| \le B i^{\frac{1}{1+t}})]\right)^2 / n^{\frac{1-t}{1+t}}\right\} = 0.$$

Proof. Without loss of generality we assume throughout this proof that $C_\varepsilon = 1$. To control the first inequality, let $f_i := \varepsilon_i \mathbb{1}(|\varepsilon_i| \le B i^{\frac{1}{1+t}})$, let a_n be a sequence of integers such that as $n \to \infty$, $a_n \to \infty$ and $a_n/n \to 0$, then

$$\begin{aligned}\sum_{i=1}^n w_i^2 \mathbb{E}[f_i^2] &= \sum_{i=1}^{a_n} w_i^2 \mathbb{E}[f_i^2] + \sum_{i=a_n+1}^n w_i^2 \mathbb{E}[f_i^2]\\
&\le \sum_{i=1}^{a_n} w_i^2 \mathbb{E}[\varepsilon_i^2 \mathbb{1}(|\varepsilon_i| \le B i^{\frac{1}{1+t}})] + \sum_{i=a_n+1}^n w_i^2 \mathbb{E}[\varepsilon_i^2 \mathbb{1}(|\varepsilon_i| \le B a_n^{\frac{1}{1+t}})]\\
&\quad + \sum_{i=a_n+1}^n w_i^2 \mathbb{E}[\varepsilon_i^2 \mathbb{1}(B a_n^{\frac{1}{1+t}} < |\varepsilon_i| \le B i^{\frac{1}{1+t}})]\\
&\le \sum_{i=1}^n w_i^2 \mathbb{E}[\varepsilon_i^2 \mathbb{1}(|\varepsilon_i| \le B a_n^{\frac{1}{1+t}})] + \sum_{i=a_n+1}^n w_i^2 \mathbb{E}[\varepsilon_i^2 \mathbb{1}(B a_n^{\frac{1}{1+t}} < |\varepsilon_i| \le B i^{\frac{1}{1+t}})].\end{aligned} \tag{A3.4}$$

For the first term in the above inequality,

$$\sum_{i=1}^{n} w_i^2 \mathbb{E}[\varepsilon_i^2 \mathbb{1}(|\varepsilon_i| \le B a_n^{\frac{1}{1+t}})] = \sum_{i=1}^{n} w_i^2 \mathbb{E}[|\varepsilon_i|^{1+t}|\varepsilon_i|^{1-t}\mathbb{1}(|\varepsilon_i| \le B a_n^{\frac{1}{1+t}})] \le \sum_{i=1}^{n} w_i^2 \mathbb{E}[|\varepsilon_i|^{1+t} B^{1-t} a_n^{\frac{1-t}{1+t}}]$$
$$\le B^{1-t}\|\boldsymbol{w}\|_2^2 \cdot a_n^{\frac{1-t}{1+t}} = c_{1n}\|\boldsymbol{w}\|_2^2 \cdot n^{\frac{1-t}{1+t}},$$

where $c_{1n} := B^{1-t} \cdot a_n^{\frac{1-t}{1+t}} / n^{\frac{1-t}{1+t}} \to 0$ is a sequence of real values that does not depend on $\boldsymbol{w}$.

For the second term on the right hand side of (A3.4), we have

$$\sum_{i=a_n+1}^{n} w_i^2 \mathbb{E}[\varepsilon_i^2 \mathbb{1}(B a_n^{\frac{1}{1+t}} < |\varepsilon_i| \le B i^{\frac{1}{1+t}})] = \sum_{i=a_n+1}^{n} w_i^2 \mathbb{E}[|\varepsilon_i|^{1+t}|\varepsilon_i|^{1-t}\mathbb{1}(B a_n^{\frac{1}{1+t}} < |\varepsilon_i| \le B i^{\frac{1}{1+t}})]$$
$$\le \sum_{i=a_n+1}^{n} w_i^2 \mathbb{E}[|\varepsilon_i|^{1+t}\mathbb{1}(B a_n^{\frac{1}{1+t}} < |\varepsilon_i| \le B i^{\frac{1}{1+t}})] \cdot B^{1-t} n^{\frac{1-t}{1+t}}$$
$$\le B^{1-t}\|\boldsymbol{w}\|_2^2 n^{\frac{1-t}{1+t}} \sup_{i \ge 1} \mathbb{E}[|\varepsilon_i|^{1+t}\mathbb{1}(|\varepsilon_i| > B a_n^{\frac{1}{1+t}})].$$

Recall the restrictions we have about ε_i's in Assumption 6, we have as $n \to \infty$,

$$c_{2n} := B^{1-t} \cdot \sup_{i \ge 1} \mathbb{E}[|\varepsilon_i|^{1+t}\mathbb{1}(|\varepsilon_i| > B a_n^{\frac{1}{1+t}})] \to 0.$$

Notice further that with the above definition, c_{2n} also does not depend on $\boldsymbol{w}$. In light of the above two results, we prove the first inequality where we select $c_n := c_{1n} + c_{2n}$.

For the second inequality, we first prove it when $t \in (0, 1)$. Using that all ε_i's are mean-zeroed, we have for any fixed $B > 0$,

$$\sum_{i=1}^{n} (\mathbb{E}[f_i])^2 = \sum_{i=1}^{n} (\mathbb{E}[\varepsilon_i \mathbb{1}(|\varepsilon_i| > B i^{\frac{1}{1+t}})])^2 \overset{(i)}{\le} \sum_{i=1}^{n} (\mathbb{E}[|\varepsilon_i|^{1+t}])^{\frac{2}{1+t}} \left(\mathbb{E}\left[\left\{\mathbb{1}(|\varepsilon_i| > B i^{\frac{1}{1+t}})\right\}^{\frac{1+t}{t}}\right]\right)^{\frac{2t}{1+t}}$$
$$= \sum_{i=1}^{n} (\mathbb{E}[|\varepsilon_i|^{1+t}])^{\frac{2}{1+t}} (\mathbb{E}[\mathbb{1}(|\varepsilon_i| > B i^{\frac{1}{1+t}})])^{\frac{2t}{1+t}} \le \sum_{i=1}^{n} \mathbb{P}(|\varepsilon_i|^{1+t} > B^{1+t} i)^{\frac{2t}{1+t}}$$
$$\overset{(ii)}{\le} n \left(\frac{\sum_{i=1}^{n} \mathbb{P}(|\varepsilon_i|^{1+t} > B^{1+t} i)}{n}\right)^{\frac{2t}{1+t}} \le n^{\frac{1-t}{1+t}} \left(\sum_{i=1}^{\infty} \mathbb{P}(|\varepsilon_i|^{1+t} > B^{1+t} i)\right)^{\frac{2t}{1+t}}. \tag{A3.5}$$

where (i) uses Hölder's inequality; (ii) uses Jensen's inequality. Now given a fixed $\eta > 0$, from the conditions of ε_i's we must have that there exists a N_η such that $\sum_{i=N_\eta}^{\infty} \mathbb{P}(|\varepsilon_i|^{1+t} > i) \le \eta/2$. Moreover, form Markov's inequality we further have that there exists a B_η such that for any $B \ge B_\eta$, $\sum_{i=1}^{N_\eta} \mathbb{P}(|\varepsilon_i|^{1+t} > Bi) \le \eta/2$. Putting together, we have that for any $B \ge \max\{B_\eta, 1\}$,

$$\sum_{i=1}^{\infty} \mathbb{P}(|\varepsilon_i|^{1+t} > B^{1+t} i) \le \eta.$$

Since the above result holds for arbitrary $\eta > 0$, we have as $B \to \infty$,

$$\sum_{i=1}^{\infty} \mathbb{P}(|\varepsilon_i|^{1+t} > B^{1+t} i) \to 0.$$

In light of above and (A3.5), we prove that the second result in the lemma statement holds with $t \in (0, 1)$.

We now turn to the case with $t = 0$. In this case, we have for any fixed $B > 0$,

$$\sum_{i=1}^{n} (\mathbb{E}[f_i])^2 \leq n \cdot \left(\sup_{i \geq 1} \mathbb{E}[|\varepsilon_i| \mathbb{1}(|\varepsilon_i| > Bi)]\right)^2 \leq n \cdot \left(\sup_{i \geq 1} \mathbb{E}[|\varepsilon_i| \mathbb{1}(|\varepsilon_i| > B)]\right)^2.$$

Using the requirement in Assumption 6 we have $\sup_{i \geq 1} \mathbb{E}[|\varepsilon_i| \mathbb{1}(|\varepsilon_i| > B)]$ converges to zero as $B \to \infty$, which proves the second result in the lemma statement in the $t = 0$ case.

Taking together, we prove the desired result. □

Lemma A5. *Consider the $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K$ and $\boldsymbol{w}$ in Lemma A3; assume that $\boldsymbol{\varepsilon}$ satisfies Assumption 6 with $t \in [0, 1]$ and that b satisfies (A1.1). Then for any fixed $\delta, \gamma > 0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}|}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \delta\right) = 0.$$

Proof. When $t = 1$ the result follows from Lemma A3. In the rest of the proof we assume throughout $t \in [0, 1)$. From the scaling of b we have that there exist $C_b, N_b > 0$ such that for all $n \geq N_b$, $b \geq C_b n^{-\frac{t}{1+t}}$, which yields that for any $B > 0, n \geq N_b$,

$$\frac{\|\boldsymbol{w}\|_2 \sqrt{\sum_{i=1}^{n} \left(\mathbb{E}[\varepsilon_i \mathbb{1}(|\varepsilon_i| \leq B i^{\frac{1}{1+t}})]\right)^2}}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} \leq \frac{\sqrt{\sum_{i=1}^{n} \left(\mathbb{E}[\varepsilon_i \mathbb{1}(|\varepsilon_i| \leq B i^{\frac{1}{1+t}})]\right)^2}}{\sqrt{\gamma} C_b n^{\frac{1-t}{2(1+t)}}}.$$

In light of the above and with Lemma A4, we have that for any fixed $\delta, \gamma > 0$, there exists a constant $B_{\gamma,\delta} > 0$ depending on (γ, δ) such that for n sufficiently large,

$$\frac{\|\boldsymbol{w}\|_2 \sqrt{\sum_{i=1}^{n} \left(\mathbb{E}[\varepsilon_i \mathbb{1}(|\varepsilon_i| \leq B_{\gamma,\delta} i^{\frac{1}{1+t}})]\right)^2}}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} \leq \frac{\delta}{3}. \tag{A3.6}$$

Writing $f_i := \varepsilon_i \mathbb{1}(|\varepsilon_i| \leq B_{\gamma,\delta} i^{\frac{1}{1+t}})$ and $\boldsymbol{f} := (f_1, \ldots, f_n)^\top$, then we have that

$$\begin{aligned} |\boldsymbol{w}^\top \boldsymbol{M}_k \boldsymbol{\varepsilon}| &\leq |\boldsymbol{w}^\top \boldsymbol{M}_k (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])| + |\boldsymbol{w}^\top \boldsymbol{M}_k \mathbb{E}[\boldsymbol{f}]| + |\boldsymbol{w}^\top \boldsymbol{M}_k (\boldsymbol{\varepsilon} - \boldsymbol{f})| \\ &\leq |\boldsymbol{w}^\top \boldsymbol{M}_k (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])| + \|\boldsymbol{w}\|_2 \|\mathbb{E}[\boldsymbol{f}]\|_2 + \|\boldsymbol{w}\|_2 \|\boldsymbol{\varepsilon} - \boldsymbol{f}\|_2. \end{aligned}$$

In light of this decomposition and also (A3.6), we only need to prove that as $n \to \infty$,

$$\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{w}^\top \boldsymbol{M}_k (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \frac{\delta}{3}\right) \to 0 \tag{A3.7}$$

and that

$$\mathbb{P}\left(\frac{\|\boldsymbol{w}\|_2 \|\boldsymbol{\varepsilon} - \boldsymbol{f}\|_2}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \frac{\delta}{3}\right) \leq \mathbb{P}\left(\frac{\|\boldsymbol{\varepsilon} - \boldsymbol{f}\|_2}{b \sqrt{\gamma n}} > \frac{\delta}{3}\right) \to 0. \tag{A3.8}$$

To prove (A3.7), applying Chebyshev's inequality and a union bound, we have

$$\begin{aligned} &\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{w}^\top \boldsymbol{M}_k (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \frac{\delta}{3}\right) \leq \sum_{k=1}^{K} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}_k (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \frac{\delta}{3}\right) \\ &\leq \sum_{k=1}^{K} \frac{9 \mathbb{E}[|\boldsymbol{w}^\top \boldsymbol{M}_k (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|^2]}{\delta^2 b^2 \max\{\|\boldsymbol{w}\|_2^4, \gamma^2 n^2\}}. \end{aligned}$$

Then by applying Lemma A4 where we select $\boldsymbol{w}$ as $\boldsymbol{M}_k^\top \boldsymbol{w}$ and using that all the ε_i's are independent and the basic inequality that $\mathbb{E}[(f_i - \mathbb{E}[f_i])^2] \le \mathbb{E}[f_i^2]$, we have

$$\mathbb{P}\left(\exists 1 \le k \le K \text{ s.t. } \frac{|\boldsymbol{w}^\top \boldsymbol{M}_k(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma n\}} > \frac{\delta}{3}\right) = o\left(\frac{K\|\boldsymbol{w}\|_2^2 n^{\frac{1-t}{1+t}}}{b^2 \max\{\|\boldsymbol{w}\|_2^4, \gamma^2 n^2\}}\right) = o(1),$$

which proves (A3.7).

To prove (A3.8), apparently we already have

$$\sum_{i=1}^\infty \mathbb{P}(f_i \ne \varepsilon_i) = \sum_{i=1}^\infty \mathbb{P}(|\varepsilon_i| > B_{\gamma,\delta} i^{\frac{1}{1+t}}) < \infty.$$

Then given any constant $\eta > 0$, there exists a constant N_η depending only on η such that

$$\mathbb{P}(\exists i > N_\eta \text{ s.t. } f_i \ne \varepsilon_i) \le \sum_{i=N_\eta+1}^\infty \mathbb{P}(f_i \ne \varepsilon_i) \le \frac{\eta}{2}.$$

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

$$\mathbb{P}(\exists \ell \le N_\eta, \text{ s.t. } |\varepsilon_\ell| > M_\eta) \le \frac{\eta}{2}.$$

Writing the event $\mathcal{E}_\eta := \{\forall i > N_\eta, f_i = \varepsilon_i \ \&\ \forall \ell \le N_\eta, |\varepsilon_i| \le M_\eta\}$ then we easily have that under this event, with n sufficiently large, $\frac{\|\boldsymbol{\varepsilon} - \boldsymbol{f}\|_2}{b\sqrt{\gamma n}} \le \frac{\delta}{3}$. In other words, given any fixed $\eta > 0$, with n sufficiently large,

$$\mathbb{P}\left(\frac{\|\boldsymbol{\varepsilon} - \boldsymbol{f}\|_2}{b\sqrt{\gamma n}} > \frac{\delta}{3}\right) \le \mathbb{P}(\mathcal{E}_\eta^c) \le \eta,$$

which proves (A3.8). In light of our control of (A3.7) and (A3.8), we prove the desired result. □

A4 Proof of basic power results

In this section, we prove the power result of RPT in the basic model where $\boldsymbol{Z}$ follows a linear function with respect to $\boldsymbol{X}$ and all noises are i.i.d.

A4.1 Proof of Theorem 3

Lemma A6. *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a matrix with all diagonal entries equal to zero. Then for any $\mathbb{P}_e \in \mathcal{D}_2$, we have for any fixed $\delta > 0$,*

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

Proof. Observe that

$$\mathbb{E}\left[\left(\frac{\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{e}}{n}\right)^2\right] = \mathbb{E}\left[\left(\sum_{i\ne j} \boldsymbol{M}_{i,j}\frac{e_i e_j}{n}\right)^2\right].$$

Using that for any $i \neq j$, $e_i \perp\!\!\!\perp e_j$, we have

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

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

Lemma A7. *For each n, let $V_{n,i}(1 \leq i \leq n)$ be independent random variables. Suppose that there exists a constant $C > 0$ such that for any i, n, $\mathbb{E}[|V_{n,i}|] \leq C$ and that*

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

then $\frac{1}{n}\sum_{i=1}^n (V_{n,i} - \mathbb{E}[V_{n,i}])$ converges in probability to zero.

Proof. We just need to prove that for any fixed $\epsilon_1, \epsilon_2 > 0$, there exists a $N_{\epsilon_1,\epsilon_2}$ such that for any $n \geq N_{\epsilon_1,\epsilon_2}$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (V_{n,i} - \mathbb{E}[V_{n,i}])\right| > \epsilon_1\right) \leq \epsilon_2. \tag{A4.9}$$

Let $\epsilon = \epsilon_1\epsilon_2/3$; then there exists a constant $a_\epsilon > 0$ such that for any $n \geq 1$, $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|V_{n,i}|\mathbb{1}(|V_{n,i}| > a_\epsilon)] \leq \epsilon$.

Write $\bar{V}_{n,i} = V_{n,i}\mathbb{1}(|V_{n,i}| \leq a_\epsilon)$ and $\tilde{V}_{n,i} = V_{n,i}\mathbb{1}(|V_{n,i}| > a_\epsilon)$. Then it holds that

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (V_{n,i} - \mathbb{E}[V_{n,i}])\right|\right] &= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (\bar{V}_{n,i} - \mathbb{E}[\bar{V}_{n,i}]) + \frac{1}{n}\sum_{i=1}^n (\tilde{V}_{n,i} - \mathbb{E}[\tilde{V}_{n,i}])\right|\right] \\
&\leq \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (\bar{V}_{n,i} - \mathbb{E}[\bar{V}_{n,i}])\right|\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |\tilde{V}_{n,i} - \mathbb{E}[\tilde{V}_{n,i}]|\right]
\end{aligned}$$

For the second term on the right hand side of the above inequality, we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |\tilde{V}_{n,i} - \mathbb{E}[\tilde{V}_{n,i}]|\right] \leq \frac{2}{n}\sum_{i=1}^n \mathbb{E}|\tilde{V}_{n,i}| \leq 2\epsilon.$$

For the first term, using the basic inequality that for any random variable V, $\mathbb{E}[|V|] \leq \sqrt{\mathbb{E}[V^2]}$ and the independence of $\bar{V}_{n,i}$'s, we have

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (\bar{V}_{n,i} - \mathbb{E}[\bar{V}_{n,i}])\right|\right] &\leq \left(\frac{1}{n^2}\mathbb{E}\left[\left|\sum_{i=1}^n (\bar{V}_{n,i} - \mathbb{E}[\bar{V}_{n,i}])\right|^2\right]\right)^{1/2} \\
&\leq \left(\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left|\bar{V}_{n,i} - \mathbb{E}[\bar{V}_{n,i}]\right|^2\right)^{1/2} \leq \frac{a_\epsilon}{\sqrt{n}}.
\end{aligned}$$

Hence for any $n \geq a_\epsilon^2/\epsilon^2$,

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (V_{n,i} - \mathbb{E}[V_{n,i}])\right|\right] \leq 3\epsilon.$$

In light of above, and by a Markov's inequality, we can have (A4.9) with $N_{\epsilon_1,\epsilon_2} := a_\epsilon^2/\epsilon^2$, thereby proving the desired result. □

Lemma A8. *The assumption of ε in Theorem 3 is a sufficient condition of that of ε in Assumption 6.*

Proof. We have

$$\sum_{i=1}^{\infty} \mathbb{P}(|\varepsilon_i|^{1+t} \geq Bi) = \sum_{i=1}^{\infty} \mathbb{P}(|\varepsilon_1|^{1+t} \geq Bi) \leq \int_0^{\infty} \mathbb{P}\left(\frac{|\varepsilon_1|^{1+t}}{B} \geq x\right) dx = \mathbb{E}\left[\frac{|\varepsilon_1|^{1+t}}{B}\right] < \infty.$$

Moreover, for any constant $a > 0$,

$$\sup_{i \geq 1} \mathbb{E}[|\varepsilon_i|^{1+t} \mathbb{1}(|\varepsilon_i|^{1+t} > a)] = \mathbb{E}[|\varepsilon_1|^{1+t} \mathbb{1}(|\varepsilon_1|^{1+t} > a)].$$

Since $\mathbb{E}[|\varepsilon_1|^{1+t}]$ is bounded, by dominated convergence theorem, the above quantity converges to zero as $a \to \infty$. □

Proof of Theorem 3. Without loss of generality, we assume throughout this proof that $\mathbb{P}_e \in \mathcal{D}_2$ and $\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}$.

To prove the desired result, it suffices to prove that for all $\delta > 0$,

$$\begin{aligned} &\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}|}{bn} \geq \delta\right) \to 0; \\ &\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}|}{bn} \geq \delta\right) \to 0; \end{aligned} \tag{A4.10}$$

and that with probability converging to 1, for all $1 \leq j, k, \leq K$,

$$\begin{aligned} \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} &\geq \frac{m}{2(4+m)}; \\ \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} &\geq \frac{m}{2(4+m)}. \end{aligned} \tag{A4.11}$$

To prove the first claim of (A4.10), since $\mathbb{P}_e \in \mathcal{D}_2$, we have from the law of large number,

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

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

$$\begin{aligned} &\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}|}{bn} \geq \delta\right) \leq \mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}|}{bn} \geq \delta \mid \mathcal{E}\right) \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \\ &\overset{(i)}{=} \mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}|}{b \max\{\|\boldsymbol{e}\|_2^2, \frac{5}{2}n\}} \geq \frac{2}{5}\delta \mid \mathcal{E}\right) \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \end{aligned}$$

where (i) straightly follows from that we are under $\mathcal{E}$. Then as a direct consequence of Lemma A5 with $\boldsymbol{e}$ as $\boldsymbol{w}$ and $\tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top$ as $\boldsymbol{M}_k$, and also Lemma A8 and that K is a constant, we prove the first claim of (A4.10).

For the second claim of (A4.10), by instead taking $\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k$ as $\boldsymbol{M}_k$, the result follows from the same argument as the first claim of (A4.10).

In the rest of the proof we focus on proving the first statement of (A4.11), and the second statement can be proven via a similar argument. Since we assume K as fixed, it boils down to proving that for any fixed j, k, with probability converging to 1, the first statement of (A4.11) holds. To achieve this goal, we apply the decomposition

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

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

For I, observe that

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

and that

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

we can apply Lemma A6 to show that for any constant $\delta>0$,

$$\lim_{n\to\infty}\mathbb{P}(|\mathrm{I}|\le\delta)\to 1. \tag{A4.12}$$

For II, given any fixed $\boldsymbol{P}_j$ and $\boldsymbol{P}_k$, define $a_{n,i}:=\frac{(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top-\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)_{i,i}}{n}$ and write $V_{n,i}:=na_{n,i}e_i^2$. Then we can rewrite II as $\mathrm{II}=\frac{1}{n}\sum_{i=1}^n V_{n,i}$. Notice that for each n, it holds that $|a_{n,i}|\le 2/n$. From this, we can have that $\mathbb{E}[|V_{n,i}|]\le 2$ uniformly for all i,n and that for any $a>0$,

$$\sup_{n\ge1}\frac{1}{n}\sum_{i=1}^n\mathbb{E}[|V_{n,i}|\mathbb{1}(|V_{n,i}|>a)]\le\mathbb{E}[2e_1^2\mathbb{1}(2e_1^2>a)].$$

Using dominated convergence theorem and that $\mathbb{E}[e_1^2]<\infty$, we have that $\mathbb{E}[2e_1^2\mathbb{1}(2e_1^2>a)]\to 0$ as $a\to\infty$; this allow us to apply Lemma A7 to get that for any constant $\delta>0$,

$$\mathbb{P}\left(|\mathrm{II}-\mathbb{E}[\mathrm{II}\mid\boldsymbol{P}_j,\boldsymbol{P}_k]|>\delta\mid\boldsymbol{P}_j,\boldsymbol{P}_k\right)\to 0. \tag{A4.13}$$

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

$$\boldsymbol{A}_k\boldsymbol{A}_k^\top=\boldsymbol{I}-\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top, \tag{A4.14}$$

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 A13, 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

$$\mathbb{E}[\mathrm{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k] \ge \frac{1}{n}((n-2p) - p - \sqrt{2p}K) \ge \frac{m}{4+m}, \tag{A4.15}$$

where the last inequality holds for sufficiently large n. From above and (A4.13), and also our control of the term I in (A4.12), we have that the first statement of (A4.11) holds with probability converging to 1. Using an analogous argument we prove the second statement of (A4.11). In light of this and our analysis of (A4.10), we obtain the desired result. □

A4.2 Proof of Theorem 4

Lemma A9. *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 $\delta > 0$,*

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

.

Proof. Let σ be the permutation corresponding to $\boldsymbol{P}$. From Lemma A11, we have there exists a partition U_1, U_2, U_3 with $|U_j \cap \sigma(U_j)| = 0$ and that $|U_j| \ge \frac{n}{4} - 1$ for $j = 1, 2, 3$ such that

$$\frac{\boldsymbol{e}^\top\boldsymbol{P}\boldsymbol{e}}{n} = \frac{1}{n}\sum_{j=1}^{3}\sum_{i\in U_j} e_i e_{\sigma(i)}.$$

Then

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

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

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

Let $\tilde{e}_i$ be a sequence of i.i.d. random variables that is independent from $\boldsymbol{w}$ and that $\tilde{e}_i \overset{\mathrm{d}}{=} e_1 e_2$. Then we easily have that $\tilde{e}_i$ 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 $a_n(\delta) := \sup_{m \geq n} \mathbb{P}(|\sum_{i=1}^m \tilde{e}_i/m| > \delta)$,

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

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

$$\sum_{i=1}^{|U_j|} \tilde{e}_i \overset{\mathrm{d}}{=} \sum_{i\in U_j} e_i e_{\sigma(i)}$$

and thus

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

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

Lemma A10. *Assume that $\mathbb{P}_e$ follows a distribution that is symmetric around zero; and let $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix. Then we have that for any $\delta > 0$,*

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

Proof. 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 $\boldsymbol{J}$. Since $\mathbb{P}_e$ is symmetric and all the e_i's are independent, we have that $\boldsymbol{e} \overset{d}{=} \boldsymbol{P}\boldsymbol{J}\boldsymbol{e}$, i.e., they are equal in distribution.

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

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{e}^\top \boldsymbol{U}\boldsymbol{e} \geq \delta\|\boldsymbol{e}\|_2^2\right) &= \mathbb{P}\left(\boldsymbol{e}^\top\boldsymbol{J}^\top\boldsymbol{P}^\top\boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{e} > \delta\boldsymbol{e}^\top\boldsymbol{J}^\top\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{J}\boldsymbol{e}\right)\\ &= \mathbb{P}\left(\boldsymbol{e}^\top\boldsymbol{J}^\top\boldsymbol{P}^\top\boldsymbol{U}\boldsymbol{P}\boldsymbol{J}\boldsymbol{e} > \delta\|\boldsymbol{e}\|_2^2\right). \qquad \text{(A4.16)}\end{aligned}$$

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

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

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 applying Lemma A12, we have

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

From above and Markov's inequality, we have

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

In light of the above equality and (A4.16), we obtain the desired result. □

Lemma A11. *Consider a permutation σ of $\{1,\dots,n\}$ such that for any $i\in\{1,\dots,n\}$, $\sigma(i)\neq i$. Then there exists a partition U_1,U_2,U_3 of the set $\{1,\dots,n\}$ such that*

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

Proof. Let G be a directed graph on vertices $\{1,\dots,n\}$ 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 $i=1,2,3$,

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

and that

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

which proves the desired result. □

Proof of Theorem 4. 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 fixed $\delta>0$,

$$\begin{aligned}\mathbb{P}\left(\exists 1\le k\le K\text{ s.t. }\frac{|\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2}\ge\delta\right)\to0;\\ \mathbb{P}\left(\exists 1\le k\le K\text{ s.t. }\frac{|\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2}\ge\delta\right)\to0;\end{aligned}\tag{A4.17}$$

and that with probability converging to 1, for all j,k,

$$\begin{aligned}\frac{\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{e}-\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{e}}{\|\boldsymbol{e}\|^2}\ge\frac{1}{5};\\ \frac{\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{e}+\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{e}}{\|\boldsymbol{e}\|^2}\ge\frac{1}{5}.\end{aligned}\tag{A4.18}$$

To prove (A4.17), since $\mathbb{P}_e \in \mathcal{D}_1$, we have that there exists some threshold $\tau > 0$ such that $\frac{2}{3} \leq \mathbb{E}[|e_i|^2 \mathbb{1}(|e_i| \leq \tau)] < \infty$, from this and by standard results of weak law of large number, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} |e_i|^2 \mathbb{1}(|e_i| \leq \tau) \geq \frac{1}{2}\right) = 1,$$

which also means that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \geq \frac{1}{2}\right) = 1. \tag{A4.19}$$

Denote the above event by $\mathcal{E}$; then we have

$$\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2} \geq \delta \mid \mathcal{E}\right) = \mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}|}{b \max\{\|\boldsymbol{e}\|_2^2, \frac{1}{2}n\}} \geq \delta \mid \mathcal{E}\right).$$

(A4.17) then follows from exactly the same proof as that of (A4.10).

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. Since we assume K is fixed, we only need to prove that for any fixed j, k, with probability converging to 1, the two inequalities in (A4.18) hold. In the rest of this proof we prove the first inequality of (A4.18), and the second inequality follows from an analogous argument. To prove this, let $\boldsymbol{e}'$ denote an independent replication of $\boldsymbol{e}$. Recall the definition of $\boldsymbol{A}_k$ in (A4.14), we have

$$\begin{aligned}
&(\boldsymbol{e}-\boldsymbol{e}')^\top(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k)(\boldsymbol{e}-\boldsymbol{e}') \\
&= \|\boldsymbol{e}-\boldsymbol{e}'\|_2^2 - (\boldsymbol{e}-\boldsymbol{e}')^\top \boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') - (\boldsymbol{e}-\boldsymbol{e}')^\top(\boldsymbol{A_j}\boldsymbol{A_j}^\top - \boldsymbol{A_k}\boldsymbol{A_k}^\top \boldsymbol{P}_k)(\boldsymbol{e}-\boldsymbol{e}') \\
&\geq \|\boldsymbol{e}-\boldsymbol{e}'\|_2^2 - (\boldsymbol{e}-\boldsymbol{e}')^\top \boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') - (\boldsymbol{e}-\boldsymbol{e}')^\top\left(\boldsymbol{A_j}\boldsymbol{A_j}^\top + \frac{\boldsymbol{A_k}\boldsymbol{A_k}^\top + \boldsymbol{P}_k^\top \boldsymbol{A_k}\boldsymbol{A_k}^\top \boldsymbol{P}_k}{2}\right)(\boldsymbol{e}-\boldsymbol{e}'),
\end{aligned} \tag{A4.20}$$

where for the last inequality we apply Cauchy-Schwartz inequality. As $e_i - e_i'$ is symmetric around zero, we have from Lemma A10 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 A5 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

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

we have from (A4.20) that

$$\begin{aligned} &\boldsymbol{e}^\top(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)\boldsymbol{e} + \boldsymbol{e}'^\top(\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top - \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)\boldsymbol{e}' \\ &\geq \frac{4}{5}\|\boldsymbol{e}-\boldsymbol{e}'\|_2^2 - (\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') - \frac{1}{5}(\|\boldsymbol{e}\|_2^2 + \|\boldsymbol{e}'\|_2^2) \\ &= \frac{3}{5}(\|\boldsymbol{e}\|_2^2 + \|\boldsymbol{e}'\|_2^2) - (\boldsymbol{e}-\boldsymbol{e}')^\top\boldsymbol{P}_k(\boldsymbol{e}-\boldsymbol{e}') - \frac{8}{5}\boldsymbol{e}^\top\boldsymbol{e}'. \end{aligned} \tag{A4.21}$$

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 Lemma A9 and (A4.19), we have that the event $\mathcal{E}_4$ holds with probability converging to 1.

For $\mathcal{E}_5$, using that all the $e_i e_i'$'s are i.i.d. random variables with $\mathbb{E}[|e_i e_i'|] = \mathbb{E}[|e_i|]\mathbb{E}[|e_i'|] < \infty$, 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 (A4.21) 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 (A4.18). In light of this and our control of (A4.17), we prove the desired result. □

A4.3 Auxiliary lemmas

Lemma A12. *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 $i = 1, \dots, n$, $\mathbb{E}[(\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}^\top)_{ii}] = \frac{1}{n}\sum_{j=1}^n \boldsymbol{M}_{jj}$.*

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 $\{1,\dots,n\}$. □

Lemma A13. *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*

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

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),

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

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

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

which proves the desired result. □

A4.4 Theoretical analysis of the algorithms

We will first show an lemma.

Lemma A14. *Consider a fixed matrix* $\boldsymbol{M} \in R^{n\times n}$ *with* $n \ge 2$ *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*

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

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 $\{(i,j) \mid i \ne j \in \{1,\dots,n\}\}$. 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, $\mathbb{E}[\boldsymbol{P}_{u,v}^2] = \mathbb{E}[\boldsymbol{P}_{u,v}] = \frac{1}{n-1}$.

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

Putting together, we have

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

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 $T \ge 2$ can be proven via analogous argument. Let $\boldsymbol{P}_\pi$ be the random permutation matrix associated with permutation π, and let $\tilde{\boldsymbol{P}}_k$ be the permutation matrix associated with $\tilde{\sigma}_k$, then we have that

$$\boldsymbol{P}_k = \boldsymbol{P}_\pi^{-1}\tilde{\boldsymbol{P}}_k\boldsymbol{P}_\pi,$$

so that for any $k_1, k_2 \in \{1, \dots, K\}$, we have that by setting k_3 as the remainder after dividing $k_1 + k_2$ by $K + 1$,

$$\boldsymbol{P}_{k_1}\boldsymbol{P}_{k_2} = \boldsymbol{P}_\pi^{-1}\tilde{\boldsymbol{P}}_{k_1}\boldsymbol{P}_\pi\boldsymbol{P}_\pi^{-1}\tilde{\boldsymbol{P}}_{k_2}\boldsymbol{P}_\pi = \boldsymbol{P}_\pi^{-1}\tilde{\boldsymbol{P}}_{k_1}\tilde{\boldsymbol{P}}_{k_2}\boldsymbol{P}_\pi = \boldsymbol{P}_\pi^{-1}\tilde{\boldsymbol{P}}_{k_3}\boldsymbol{P}_\pi = \boldsymbol{P}_{k_3}.$$

This proves that the returned $\mathcal{P}_K$ satisfies Assumption 2.

In addition, we have from Lemma A14 and $\mathrm{tr}[\boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{V}_0\boldsymbol{V}_0^\top] = \mathrm{tr}[\boldsymbol{V}_0\boldsymbol{V}_0^\top] = p$ that for any k,

$$\mathbb{P}\left(|\mathrm{tr}[\boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{P}_k]| \ge \sqrt{2p}K\right) \le \frac{1}{K^2}.$$

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. □

A5 Proof of additional power results

A5.1 Proof of Theorem 5

The following is an extension of Marcinkiewicz-Zygmund strong law of large numbers to the sum of non-i.i.d. random variables.

Lemma A15. *Consider the $\boldsymbol{\varepsilon}$ in Assumption 5. If $t < 1$, for any constant $\delta > 0$, it holds that,*

$$\lim_{n\to\infty}\mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 \ge \delta \cdot n^{\frac{2}{1+t}}) = 0.$$

Proof. Without loss of generality, we assume throughout this proof that $C_\varepsilon = 1$. For any constant $\delta, \epsilon > 0$, set $B := \epsilon^{\frac{1}{1-t}} \cdot \delta^{\frac{1}{1-t}}/6^{\frac{1}{1-t}}$. Define $f_i = \varepsilon_i \mathbb{1}(|\varepsilon_i| \le Bi^{\frac{1}{t+1}})$, by Assumption 5, we have that $\sum_{i=1}^\infty \Pr(f_i \ne \varepsilon_i) < \infty$. Thus given $\epsilon > 0$, there exists an integer N_ϵ such that

$$\sum_{i=N_\epsilon}^\infty \mathbb{P}(f_i \ne \varepsilon_i) < \frac{\epsilon}{3}.$$

Moreover, given N_ϵ and ϵ, we have from Markov's inequality that exists a constant $M_\epsilon > 0$ such that

$$\sum_{i=1}^{N_\epsilon} \mathbb{P}(|\varepsilon_i| \geq M_\epsilon) \leq \frac{\epsilon}{3}.$$

Define the two random events

$$\mathcal{E}_1 := \{\forall N_\epsilon \leq i \leq n, f_i = \varepsilon_i\} \quad \& \quad \mathcal{E}_2 := \{\forall i \leq N_\epsilon, |\varepsilon_i| \leq M_\epsilon\}.$$

Now for any $n \geq N_\epsilon$, we then have

$$\begin{aligned}\mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 \geq \delta \cdot n^{\frac{2}{1+t}}) &\leq \mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 \geq \delta \cdot n^{\frac{2}{1+t}} \ \& \ \mathcal{E}_1 \ \& \ \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) \\ &\leq \mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 \geq \delta \cdot n^{\frac{2}{1+t}} \ \& \ \mathcal{E}_1 \ \& \ \mathcal{E}_2) + \frac{2\epsilon}{3}.\end{aligned}$$

Under $\mathcal{E}_1$ & $\mathcal{E}_2$, it holds that $\sum_{i=1}^n \varepsilon_i^2 \leq \sum_{i=1}^n f_i^2 + N_\epsilon M_\epsilon^2$, whence

$$\begin{aligned}\mathbb{P}(\|\varepsilon\|_2^2 \geq \delta \cdot n^{\frac{2}{1+t}} \ \& \ \mathcal{E}_1 \ \& \ \mathcal{E}_2) &\leq \mathbb{P}\left(\sum_{i=1}^n f_i^2 + N_\epsilon M_\epsilon^2 \geq \delta \cdot n^{\frac{2}{1+t}} \ \& \ \mathcal{E}_1 \ \& \ \mathcal{E}_2\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^n f_i^2 + N_\epsilon M_\epsilon^2 \geq \delta \cdot n^{\frac{2}{1+t}}\right).\end{aligned}$$

Now it remains to understand the concentration of $\sum_{i=1}^n f_i^2$. We have

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}[f_i^2] &\leq \sum_{i=1}^n \mathbb{E}[\varepsilon_i^2 \mathbb{1}(|\varepsilon_i| \leq B i^{\frac{1}{1+t}})] \leq \sum_{i=1}^n \mathbb{E}[\varepsilon_i^2 \mathbb{1}(|\varepsilon_i| \leq B n^{\frac{1}{1+t}})] \\ &= \sum_{i=1}^n \mathbb{E}[|\varepsilon_i|^{1+t} |\varepsilon_i|^{1-t} \mathbb{1}(|\varepsilon_i| \leq B n^{\frac{1}{1+t}})] \leq B^{1-t} n^{\frac{1-t}{1+t}} \sum_{i=1}^n \mathbb{E}[|\varepsilon_i|^{1+t} \mathbb{1}(|\varepsilon_i| \leq B n^{\frac{1}{1+t}})] \\ &\leq B^{1-t} n^{\frac{2}{1+t}} = \frac{\epsilon\delta}{6} n^{\frac{2}{1+t}},\end{aligned}$$

where for the last inequality we apply that $\mathbb{E}[|\varepsilon_i|^{1+t}] \leq 1$ for all i. In light of the above and by Markov's inequality, we have for $n \geq \max\{N_\epsilon, (\frac{6N_\epsilon M_\epsilon^2}{\epsilon\delta})^{\frac{1+t}{2}}\}$,

$$\mathbb{P}(\sum_{i=1}^n f_i^2 + N_\epsilon M_\epsilon^2 \geq \delta \cdot n^{\frac{2}{1+t}}) \leq \frac{\sum_{i=1}^n \mathbb{E}[f_i^2] + N_\epsilon M_\epsilon^2}{\delta \cdot n^{\frac{2}{1+t}}} \leq \frac{\epsilon}{3}.$$

Concluding, we have that given any fixed $\epsilon, \delta > 0$, for all $n \geq \max\{N_\epsilon, (\frac{6N_\epsilon M_\epsilon^2}{\epsilon\delta})^{\frac{1+t}{2}}\}$,

$$\mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 \geq \delta \cdot n^{\frac{2}{1+t}}) \leq \epsilon.$$

The proof is then complete. □

Lemma A16. *Consider the $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ with $\|\boldsymbol{M}\|_{\mathrm{op}} \leq 1$ and random vectors $\boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ satisfying Assumption 5. Let $t \in [0, 1]$ be given and assume that b satisfies (8). Then for any fixed $\delta > 0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \boldsymbol{M} \boldsymbol{\varepsilon}|}{bn} > \delta\right) = 0.$$

Proof. When $t = 1$, the result is a direct consequence from Chebyshev's inequality since both e_i's and ε_i's are independent random variables with uniformly bounded second order moments. We now focus on the $t < 1$ case. In this case, we have for any fixed $\delta, \delta' > 0$, by Chebyshev's inequality and union bound,

$$\begin{aligned}
\mathbb{P}\left(\frac{|\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|}{bn} > \delta\right) &\leq \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|}{bn} > \delta \;\&\; \|\boldsymbol{\varepsilon}\|_2^2 \leq \delta' n^{\frac{2}{1+t}}\right) + \mathbb{P}\left(\|\boldsymbol{\varepsilon}\|_2^2 > \delta' n^{\frac{2}{1+t}}\right) \\
&\leq \frac{\mathbb{E}[|\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|^2 \mathbb{1}\{\|\boldsymbol{\varepsilon}\|_2^2 \leq \delta' n^{\frac{2}{1+t}}\}]}{\delta^2 b^2 n^2} + \mathbb{P}\left(\|\boldsymbol{\varepsilon}\|_2^2 > \delta' n^{\frac{2}{1+t}}\right) \\
&\leq \frac{C_e \delta' \cdot n^{\frac{2}{1+t}}}{\delta^2 b^2 n^2} + \mathbb{P}\left(\|\boldsymbol{\varepsilon}\|_2^2 > \delta' n^{\frac{2}{1+t}}\right),
\end{aligned}$$

where $\delta' > 0$ can be an arbitrary constant that does not depend on n. From above, and that

$$\liminf_{n\to\infty} b/n^{-\frac{t}{1+t}} > 0,$$

we have that there exists a constant C such that for any fixed $\delta, \delta' > 0$, with n sufficiently large,

$$\mathbb{P}\left(\frac{|\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|}{bn} > \delta\right) \leq C\delta'/\delta^2 + \mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 > \delta' n^{\frac{2}{1+t}}). \tag{A5.22}$$

In light of the above, we have that given any $\eta > 0$, by choosing δ' as $\delta_\eta := \delta^2\eta/(2C)$,

$$\mathbb{P}\left(\frac{|\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|}{bn} > \delta\right) \leq \frac{\eta}{2} + \mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 > \delta_\epsilon n^{\frac{2}{1+t}}).$$

From above, and by applying Lemma A15 we prove that by choosing $N_\eta > 0$ large enough, for any $n \geq N_\eta$,

$$\mathbb{P}\left(\frac{|\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|}{bn} > \delta\right) \leq \eta,$$

which proves the desired result. □

Proof of Theorem 5. Applying Lemma A16 and that K is fixed, it follows from the proof of Theorem 3 that the only remaining task is to prove that for any fixed $\boldsymbol{P}_j, \boldsymbol{P}_k$, with probability converging to 1,

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

Without loss of generality we just prove the first inequality. We can write

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

The term I can be controlled by a Chebyshev's inequality following the same proof in Lemma A6, since the proof of Lemma A6 can also be generalized to the case where e_i's are heterogeneous but have uniformly bounded variance. For term II, knowing that we are under Assumption 5, it follows from the same lines of proof as the term II in the proof of Theorem 3 that (A4.13) still holds.

Now the only remaining job is to understand $\mathbb{E}[\text{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k]$. Recall the definition of $\boldsymbol{A}_k$; by setting $\boldsymbol{D}_e$ as a diagonal matrix with (i,i)-th entry equal to $\mathbb{E}[e_i^2]$, we have

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

Using that $\liminf_{n\to\infty} \text{tr}[\boldsymbol{D}_e]/n > c_e$, we have that there exists some N such that for all $n \geq N$, $\text{tr}[\boldsymbol{D}_e]/n > c_e$; and therefore following the same proof as (A4.15), we have that with n sufficiently large,

$$\mathbb{E}[\text{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k] \geq c_e - \frac{3C_e p - C_e\sqrt{2p}K}{n} \geq \frac{c_e m}{4C_e/c_e + m}, \tag{A5.23}$$

i.e., that it is bounded by a constant that does not depend on n. Putting together, we prove the desired result. □

A5.2 Proof of Theorem 6

Proof of Theorem 6. We first need to prove that for any fixed $\delta > 0$,

$$\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. }, \frac{|\boldsymbol{h}^\top \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}|}{bn} \geq \delta\right) \to 0;$$
$$\mathbb{P}\left(\exists 1 \leq k \leq K \text{ s.t. } \frac{|\boldsymbol{h}^\top \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k\boldsymbol{\varepsilon}|}{bn} \geq \delta\right) \to 0;$$

and that the same inequalities still hold but with $\boldsymbol{h}$ replaced by $\boldsymbol{e}$. Since $\limsup_{n\to\infty} \|\boldsymbol{h}\|_2^2/n \leq r$, we have that there exists an N such that for all $n \geq N$, $\|\boldsymbol{h}\|_2^2/n \leq 2r$. Then as a direct consequence of Lemma A5 where we select $\boldsymbol{w} = \boldsymbol{h}$, $\boldsymbol{M}_k = \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top$ or $\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k$, the above two inequalities still hold. To prove that the above two inequalities still hold with $\boldsymbol{h}$ replaced by $\boldsymbol{e}$, since K is taken as fixed, we may instead apply Lemma A16 where we select $\boldsymbol{M} = \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top$ or $\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k$.

Following again the proof of Theorem 3, it remains to prove that for any fixed $\boldsymbol{P}_j, \boldsymbol{P}_k$, with probability converging to 1,

$$\frac{(\boldsymbol{e}+\boldsymbol{h})^\top \tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top(\boldsymbol{e}+\boldsymbol{h}) - (\boldsymbol{e}+\boldsymbol{h})^\top \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k(\boldsymbol{e}+\boldsymbol{h})}{n} \geq \frac{(c_e - r)m}{2(4C_e/(c_e-r)+m)};$$
$$\frac{(\boldsymbol{e}+\boldsymbol{h})^\top \tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top(\boldsymbol{e}+\boldsymbol{h}) + (\boldsymbol{e}+\boldsymbol{h})^\top \tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k(\boldsymbol{e}+\boldsymbol{h})}{n} \geq \frac{(c_e - r)m}{2(4C_e/(c_e-r)+m)}.$$

In the rest of the proof we focus on the first inequality, and the second can be proven via an analogous argument. Notice that

$$\begin{aligned}&\frac{(\boldsymbol{e}+\boldsymbol{h})^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top(\boldsymbol{e}+\boldsymbol{h})-(\boldsymbol{e}+\boldsymbol{h})^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k(\boldsymbol{e}+\boldsymbol{h})}{n}\\&=\frac{\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{e}-\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{e}}{n}+\frac{\boldsymbol{h}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{h}-\boldsymbol{h}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{h}}{n}\\&\quad+\frac{\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{h}-\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{h}}{n}+\frac{\boldsymbol{h}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{e}-\boldsymbol{h}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{e}}{n}.\end{aligned}$$

As a direct consequence of Chebyshev's inequality and that $\limsup_{n\to\infty}\|\boldsymbol{h}\|_2^2/n \le r$, we can easily prove that the last two terms on the right hand side of the above inequality converge in probability to zero; for the second inequality, we have that for any fixed constant $\delta > 0$, with sufficiently large n,

$$\frac{\boldsymbol{h}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{h}-\boldsymbol{h}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{h}}{n}\ge -\|\boldsymbol{h}\|_2^2/n \ge -r-\delta.$$

In light of the above and using exactly the same lines of proof as in Theorem 5 to deal with the first term, we have that for any constant $\delta > 0$, as n goes to infinity, then with probability converging to 1,

$$\begin{aligned}&\frac{(\boldsymbol{e}+\boldsymbol{h})^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top(\boldsymbol{e}+\boldsymbol{h})-(\boldsymbol{e}+\boldsymbol{h})^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k(\boldsymbol{e}+\boldsymbol{h})}{n}\\&\ge c_e-\frac{3C_ep-C_e\sqrt{2pK}}{n}-r-\delta\ge\frac{(c_e-r)m}{4C_e/(c_e-r)+m}-\delta;\end{aligned}$$

which proves the desired result since $\delta > 0$ can be chosen arbitrarily small. □

A5.3 Proof of Corollary 7

Proof. Without loss of generality we just consider the case with $t \in [0, 1)$, and the case with $t = 1$ follows from an analogous argument. Let

$$\tilde{z}_n := \min_{1\le j,k\le K}\min_{z\in\{0,1\}}\boldsymbol{Z}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{Z}+(-1)^z\boldsymbol{Z}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{Z}.$$

It boils down to proving that for any $1 \le k \le K$ and for any fixed $\delta > 0$,

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{\boldsymbol{Z}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{\varepsilon}}{b\tilde{z}_n}>\delta\right)=0, \tag{A5.24}$$

and that the same conclusion holds with $\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top$ replaced by $\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k$. In this proof we just prove the former, and the later follows from an analogous argument. To achieve this goal, write $\boldsymbol{w}_k := \sqrt{n}\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{Z}/\|\boldsymbol{V}_0\boldsymbol{Z}\|_2$. Since $\tilde{\boldsymbol{V}}_k$ spans a subspace of $\boldsymbol{V}_0$, it is straightforward that $\|\boldsymbol{w}_k\|_2 \le \sqrt{n}$; and to prove (A5.24), it suffices to prove that for $|b'| = \Omega(n^{-\frac{t}{1+t}})$, $\lim_{n\to\infty}\mathbb{P}\left(\frac{\boldsymbol{w}_k^\top\boldsymbol{\varepsilon}}{b'n}>\delta\right)=0$, which is a direct consequence of Lemma A5 where we select $K = 1$, $\boldsymbol{w} = \boldsymbol{w}_k$, $\boldsymbol{M}_1 = \boldsymbol{I}$, $b = b'$ and $\gamma = 1$. □

A5.4 Proof of Corollary 8

Proof. From Lemma A4, we have that for any fixed $\delta > 0$, there exists a constant $B_\delta > 0$ such that uniformly for all n,

$$\sqrt{\frac{1}{n^{\frac{1-t}{1+t}}}\sum_{i=1}^{n}\Big(\mathbb{E}[\varepsilon_i \mathbb{1}(|\varepsilon_i| \leq B_\delta i^{\frac{1}{1+t}})]\Big)^2} \leq \frac{\delta}{3}. \tag{A5.25}$$

Now writing $f_i := \varepsilon_i \mathbb{1}(|\varepsilon_i| \leq B_\delta i^{\frac{1}{1+t}})$ and $\boldsymbol{f}$ as in the proof of Lemma A5, we have for any $\boldsymbol{w} \in \mathcal{S}^{n-1}$,

$$|\boldsymbol{w}^\top \boldsymbol{\varepsilon}| \leq |\boldsymbol{w}^\top(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])| + |\boldsymbol{w}^\top \mathbb{E}[\boldsymbol{f}]| + |\boldsymbol{w}^\top(\boldsymbol{\varepsilon} - \boldsymbol{f})| \leq |\boldsymbol{w}^\top(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])| + \|\mathbb{E}[\boldsymbol{f}]\|_2 + \|\boldsymbol{\varepsilon} - \boldsymbol{f}\|_2.$$

In light of the above and (A5.25), we have that to prove the desired statement, we only need to prove that

$$\lim_{n\to\infty}\sup_{\boldsymbol{w}\in\mathcal{S}^{n-1}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|}{n^{\frac{1-t}{2(1+t)}}} > \frac{\delta}{3}\right) = 0 \quad \& \quad \lim_{n\to\infty}\mathbb{P}\left(\frac{\|\boldsymbol{\varepsilon} - \boldsymbol{f}\|_2}{n^{\frac{1-t}{2(1+t)}}} > \frac{\delta}{3}\right) = 0.$$

The second inequality follows from exactly the same lines of proof as in the proof of (A3.8). For the first inequality, we apply Chebyshev's inequality, Lemma A4 and the basic inequality $\mathbb{E}[(f_i - \mathbb{E}[f_i])^2] \leq \mathbb{E}[f_i^2]$ to get that

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{n-1}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|}{n^{\frac{1-t}{2(1+t)}}} > \frac{\delta}{3}\right) \leq \sup_{\boldsymbol{w}\in\mathcal{S}^{n-1}} \frac{9\mathbb{E}[|\boldsymbol{w}^\top(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|^2]}{\delta^2 n^{\frac{1-t}{1+t}}} \leq \frac{9c_n}{\delta^2},$$

which converges to zero as $n \to \infty$. Notice that in the last inequality we use the fact that c_n does not depend on $\boldsymbol{w}$. Putting together we prove the desired result. □

A5.5 Proof of Theorem A1

Since Lemma A5 can also work on the case where K diverges with n, with the help of the proof of Theorem 3, it remains to prove that (A4.11) still holds with a diverging K. Write

$$\mathrm{I}_{1k} := \frac{\boldsymbol{e}^\top(\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top - \mathrm{diag}(\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top))\boldsymbol{e}}{n}, \quad \mathrm{I}_{2k} := \frac{\boldsymbol{e}^\top(\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k - \mathrm{diag}(\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k))\boldsymbol{e}}{n},$$

$$\mathrm{II}_{1k} := \frac{\boldsymbol{e}^\top\mathrm{diag}(\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top)\boldsymbol{e} - \mathbb{E}[e_1^2]\mathrm{tr}[\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top]}{n}, \quad \mathrm{II}_{2k} := \frac{\boldsymbol{e}^\top\mathrm{diag}(\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k)\boldsymbol{e} - \mathbb{E}[e_1^2]\mathrm{tr}[\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k]}{n},$$

and

$$\mathrm{III}_{j,k} := \mathbb{E}[e_1^2]\cdot\frac{\mathrm{tr}[\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top]}{n} - \mathbb{E}[e_1^2]\cdot\frac{\mathrm{tr}[\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k]}{n}.$$

We first consider I_{1k}. From Lemma A6, we have that by Chebyshev's inequality and a union bound, as $n \to \infty$, for any fixed $\delta > 0$,

$$\mathbb{P}(\forall 1 \leq k \leq K, |\mathrm{I}_{1k}| \geq \delta) = O(K/n) = o(1).$$

Using an analogous argument we can prove that the above result still holds with I_{2k}.

Second, we consider II_{1k}. In the following, we control this by applying Lemma A5 where we select $e_i^2 - \mathbb{E}[e_i^2]$ as ε_i, $\mathrm{diag}(\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top)$ as $\boldsymbol{M}_k$, $\boldsymbol{1}$ as $\boldsymbol{w}$, $\kappa/2$ as t and set $\gamma = 1$. We now discuss by cases based on κ.

1. When $0 \le \kappa < 2$, we can set the b in Lemma A5 as $b := \sqrt{K} n^{\frac{-\kappa/2}{1+\kappa/2}}$, then it follows from Lemma A5 that

$$\sup_{1 \le k \le K} \frac{1}{n} \sum_{i=1}^{n} a_{k,i}(e_i^2 - \mathbb{E}[e_i^2]) = o_{\mathbb{P}}(b \cdot \|\mathbf{1}\|_2^2/n) = o_{\mathbb{P}}(\sqrt{K} n^{-\frac{\kappa}{2+\kappa}}) = o_{\mathbb{P}}(1).$$

2. When $\kappa \ge 2$, we can choose the b in Lemma A5 as $b = 1 = \omega(\sqrt{K/n})$ (note that $K = o(n)$ here). Then it follows from Lemma A5 that

$$\sup_{1 \le k \le K} \frac{1}{n} \sum_{i=1}^{n} a_{k,i}(e_i^2 - \mathbb{E}[e_i^2]) = o_{\mathbb{P}}(b \|\mathbf{1}\|_2^2/n) = o_{\mathbb{P}}(1).$$

In light of both cases we have $\sup_{1 \le k \le K} \mathrm{II}_{1k} = o_{\mathbb{P}}(1)$. Analogously we can derive the same bound for II_{2k}.

We thirdly consider $\mathrm{III}_{j,k}$. Using exactly the same analysis as the $\mathbb{E}[\mathrm{II} \mid \boldsymbol{P}_j, \boldsymbol{P}_k]$ in the proof of Theorem 3, we have for any $1 \le j, k \le K$,

$$\mathrm{III}_{j,k} \ge \frac{1}{n}\left(n - 3p - \min\{\sqrt{2p}K, p\}\right) \ge \frac{m}{4+m}.$$

In light of our control of $\mathrm{I}_{1k}, \mathrm{I}_{2k}, \mathrm{II}_{1k}, \mathrm{II}_{2k}$ and $\mathrm{III}_{j,k}$, we prove the desired result.

A6 Proof of minimax rate optimality results

A6.1 Proof of Theorem 9

Without loss of generality we consider the scenario where $\beta = \beta^Z = 0$. Let $H_1(\tau)$ be the class of alternatives such that $|b| \ge \tau$, 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 $H_1(\tau)$,

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

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

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

A6.2 Preliminary lemmas for Theorem 10

We first invoke the following lemma, which is a uniform convergence extension of Lemma A4.

Lemma A17. *Let $t \in (0,1), \gamma > 0$ be given. For any fixed $B > 0$,*

$$\sup_{\boldsymbol{w}\in\mathbb{R}^n_\circ} \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \sum_{i=1}^n \frac{\mathbb{E}[w_i^2\varepsilon_i^2\mathbb{1}(|\varepsilon_i| \le Bi^{\frac{1}{1+t_1}})]}{\|\boldsymbol{w}\|_2^2 n^{\frac{1-t}{1+t}+\gamma}} = o(1).$$

Moreover,

$$\sum_{i=1}^n \left(\mathbb{E}[\varepsilon_i\mathbb{1}(|\varepsilon_i| \le Bi^{\frac{1}{1+t_1}})]\right)^2 / n^{\frac{1-t}{1+t}+\gamma} = o(1),$$

where $\mathbb{R}^n_\circ$ denotes the n-dimensional Euclidean space excluding the original point and t_1 denote a constant in $(-1,t)$ such that

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

Proof. As a direct consequence of Jensen's inequality it is straightforward that

$$\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \mathbb{E}[|\varepsilon_1|^{1+t_1}] \le \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} (\mathbb{E}[|\varepsilon_1|^{1+t}])^{\frac{1+t_1}{1+t}} \le 2^{\frac{1+t_1}{1+t}}. \tag{A6.26}$$

Let $f_i := \varepsilon_i\mathbb{1}(|\varepsilon_i| \le Bi^{\frac{1}{1+t_1}})$. To prove the first statement, following the proof of Lemma A4, we only need to prove that

$$\sup_{\boldsymbol{w}\in\mathbb{R}^n_\circ} \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \sum_{i=1}^n \frac{\mathbb{E}[w_i^2\varepsilon_i^2\mathbb{1}(|\varepsilon_i| \le Ba_n^{\frac{1}{1+t_1}})]}{\|\boldsymbol{w}\|_2^2 n^{\frac{1-t}{1+t}+\gamma}} = o(1) \tag{A6.27}$$

and that

$$\sup_{\boldsymbol{w}\in\mathbb{R}^n_\circ} \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \sum_{i=a_n+1}^n \frac{\mathbb{E}[w_i^2\varepsilon_i^2\mathbb{1}(Ba_n^{\frac{1}{1+t_1}} < |\varepsilon_i| \le Bi^{\frac{1}{1+t_1}})]}{\|\boldsymbol{w}\|_2^2 n^{\frac{1-t}{1+t}+\gamma}} = o(1). \tag{A6.28}$$

For (A6.27), following exactly the same lines of proof as in Lemma A4 but with t replaced by t_1, we can easily have that its left hand side is of order $o(\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \mathbb{E}[|\varepsilon_1|^{1+t_1}])$, which is of order $o(1)$ knowing (A6.26). For (A6.28), following again the proof in Lemma A4 but with t replaced by t_1, we only need to prove that

$$\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \mathbb{E}[|\varepsilon_1|^{1+t_1}\mathbb{1}(|\varepsilon_1| > Ba_n^{\frac{1}{1+t_1}})] = o(1).$$

This follows from the fact that

$$\mathbb{E}[|\varepsilon_1|^{1+t_1}\mathbb{1}(|\varepsilon_1| > Ba_n^{\frac{1}{1+t_1}})] \le \mathbb{E}[|\varepsilon_1|^{1+t}|\varepsilon_1|^{t_1-t}\mathbb{1}(|\varepsilon_1| > Ba_n^{\frac{1}{1+t_1}})]$$
$$\le \frac{\mathbb{E}[|\varepsilon_1|^{1+t}\mathbb{1}(|\varepsilon_1| > Ba_n^{\frac{1}{1+t_1}})]}{B^{t-t_1}a_n^{\frac{t-t_1}{1+t_1}}} \le \mathbb{E}[|\varepsilon_1|^{1+t}]/(B^{t-t_1}a_n^{\frac{t-t_1}{1+t_1}}).$$

Now for the second claim, recall again the proof in Lemma A4, we have

$$\sum_{i=1}^n (\mathbb{E}[f_i])^2 \le 2^{\frac{2}{1+t}} n^{\frac{1-t}{1+t}} \left(\sum_{i=1}^\infty \mathbb{P}(|\varepsilon_i|^{1+t_1} > B^{1+t_1}i)\right)^{\frac{2t}{1+t}},$$

and the desired result follows from that

$$\sum_{i=1}^{\infty} \mathbb{P}(|\varepsilon_i|^{1+t_1} > B^{1+t_1} i) \le \int_0^{\infty} \mathbb{P}\left(\frac{|\varepsilon_i|^{1+t_1}}{B^{1+t_1}} > x\right) dx = \frac{\mathbb{E}[|\varepsilon_i|^{1+t_1}]}{B^{1+t_1}}$$

and (A6.26). □

Armed with the above lemma, we now introduce Lemma A18, which is an extension of Lemma A5.

Lemma A18. *Consider the $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ with $\|\boldsymbol{M}\|_{\text{op}} \le 1$ and let $t \in (0,1)$ 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 $\delta, \gamma' > 0$,*

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathbb{R}^n} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|}{b\max\{\|\boldsymbol{w}\|_2^2, \gamma' n\}} > \delta\right) = 0.$$

Proof. When $t = 1$, the result follows from an argument analogous to the proof of Lemma A3. Therefore we just need to prove that the result holds for $t \in (0, 1)$.

Let f_i be defined as in Lemma A17 for some constant $B > 0$ and write $\boldsymbol{f} := (f_1, \ldots, f_n)^\top$. Then we only need to prove that

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathbb{R}^n} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])|}{b\max\{\|\boldsymbol{w}\|_2^2, \gamma' n\}} > \delta\right) = 0,$$

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathbb{R}^n} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \frac{|\boldsymbol{w}^\top \boldsymbol{M}\mathbb{E}[\boldsymbol{f}]|}{b\max\{\|\boldsymbol{w}\|_2^2, \gamma' n\}} = 0,$$

and that

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathbb{R}^n} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \boldsymbol{M}(\boldsymbol{\varepsilon} - \boldsymbol{f})|}{b\max\{\|\boldsymbol{w}\|_2^2, \gamma' n\}} > \delta\right) = 0.$$

The first inequality follows from exactly the same lines of proof as in Lemma A5, except that we replace Lemma A4 with Lemma A17; the second inequality follows from a Cauchy-Schwartz inequality and Lemma A17. Now we are still left with the task of dealing with the third inequality.

Following the proof of Lemma A5, to prove that the third inequality holds, it remains to prove that given any constant $\eta > 0$, there exists constants N_η, C_η such that the following result hold:

$$\sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\mathcal{E}_\eta^c\right) \le \eta, \quad \text{where} \quad \mathcal{E}_\eta := \{\forall i > N_\eta, f_i = \varepsilon_i, \forall \ell \le N_\eta, |\varepsilon_\ell| \le C_\eta\}. \tag{A6.29}$$

To achieve this goal, observe that for any integer $N > 0$,

$$\sum_{i=N+1}^{\infty} \mathbb{P}(f_i \ne \varepsilon_i) = \sum_{i=N+1}^{\infty} \mathbb{P}(|\varepsilon_i|^{1+t_1} > B^{1+t_1} i) \le \int_N^{\infty} \mathbb{P}(|\varepsilon_1|^{1+t_1} > B^{1+t_1} x) dx$$
$$= \int_0^{\infty} \mathbb{P}\left\{\left(\frac{|\varepsilon_1|^{1+t_1}}{B^{1+t_1}} - N\right) \mathbb{1}(|\varepsilon_1|^{1+t_1} > B^{1+t_1} N) > x\right\} dx$$
$$= \mathbb{E}\left[\left(\frac{|\varepsilon_1|^{1+t_1}}{B^{1+t_1}} - N\right) \mathbb{1}(|\varepsilon_1|^{1+t_1} > B^{1+t_1} N)\right] \le \mathbb{E}\left[\frac{|\varepsilon_1|^{1+t_1}}{B^{1+t_1}} \mathbb{1}(|\varepsilon_1|^{1+t_1} > B^{1+t_1} N)\right],$$

whence

$$\mathbb{E}\left[\frac{|\varepsilon_1|^{1+t_1}}{B^{1+t_1}}\mathbb{1}(|\varepsilon_1|^{1+t_1} > B^{1+t_1}N)\right] = \mathbb{E}\left[\frac{|\varepsilon_1|^{1+t}|\varepsilon_1|^{t_1-t}}{B^{1+t_1}}\mathbb{1}(|\varepsilon_1|^{1+t_1} > B^{1+t_1}N)\right]$$
$$\leq \mathbb{E}\left[\frac{|\varepsilon_1|^{1+t}}{B^{1+t_1}}\mathbb{1}(|\varepsilon_1|^{1+t_1} > B^{1+t_1}N)\right]\cdot B^{t_1-t}N^{\frac{t_1-t}{1+t_1}} \leq 2\frac{N^{-\frac{t-t_1}{1+t_1}}}{B^{1+t}}.$$

From above, by choosing $N_\eta := \left(\frac{4}{\eta B^{1+t}}\right)^{\frac{1+t_1}{t-t_1}}$, we have for any $\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}$,

$$\mathbb{P}(\exists i > N_\eta \text{ s.t. } f_i \neq \varepsilon_i) \leq \sum_{i=N_\eta+1}^{\infty}\mathbb{P}(f_i \neq \xi_i) \leq \frac{\eta}{2}.$$

Further letting $C_\eta := (4N_\eta/\eta)^{\frac{1}{1+t}}$, we have for any $\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}$,

$$\mathbb{P}(\exists \ell \leq N_\eta, \text{ s.t. } |\varepsilon_\ell| > C_\eta) \leq N_\eta\mathbb{P}(|\varepsilon_1| > C_\eta) \leq \frac{2N_\eta}{C_\eta^{1+t}} = \frac{\eta}{2}.$$

In light of the above two inequalities we prove (A6.29), thereby proving the desired result. ☐

Lemma A19. *Let $a_{n,i}$ $(i, n = 1, 2, \ldots)$ 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

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

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

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

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

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

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. □

A6.3 Theoretical analysis of (17)

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

$$\begin{aligned}
&\sup_{\mathbb{P}_e\in\mathcal{D}_{2+\nu}} \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} > \delta\right) \to 0;\\
&\sup_{\mathbb{P}_e\in\mathcal{D}_{2+\nu}} \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}|}{bn} > \delta\right) \to 0;
\end{aligned} \tag{A6.30}$$

and that,

$$\begin{aligned}
&\sup_{\mathbb{P}_e\in\mathcal{D}_{2+\nu}} \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \mathbb{P}\left(\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} < \frac{m}{2(4+m)}\right) \to 0;\\
&\sup_{\mathbb{P}_e\in\mathcal{D}_{2+\nu}} \sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}} \mathbb{P}\left(\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} < \frac{m}{2(4+m)}\right) \to 0.
\end{aligned} \tag{A6.31}$$

To prove the first claim of (A6.30), since $\mathbb{P}_e \in \mathcal{D}_{2+\nu}$, using Lemma A19 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

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

From the above two inequalities, we have the random event $\mathcal{E} := \{\|\boldsymbol{e}\|_2^2 \leq 2^{(3+\nu)/(2+\nu)} n\}$ satisfies that $\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}(\mathcal{E}^c) \to 0$. Therefore, by choosing $\gamma' = 2^{(3+\nu)/(2+\nu)}$ we can control the first inequality of (A6.30) via that

$$\begin{aligned}
\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} > \delta\right) &\leq \sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} > \delta \mid \mathcal{E}\right) + \sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}(\mathcal{E}^c) \\
&\overset{(i)}{=} \sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{b \max\{\|\boldsymbol{e}\|_2^2, \gamma' n\}} > \frac{\delta}{\gamma'} \mid \mathcal{E}\right) + \sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}(\mathcal{E}^c) \\
&\leq \sup_{\boldsymbol{w} \in \mathbb{R}^n} \sup_{\mathbb{P}_\varepsilon \in \mathcal{D}_{1+t}} \mathbb{P}\left(\frac{|\boldsymbol{w}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{b \max\{\|\boldsymbol{w}\|_2^2, \gamma' n\}} > \frac{\delta}{\gamma'}\right) + \sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}(\mathcal{E}^c),
\end{aligned}$$

where for the equality (i) we apply that we are under $\mathcal{E}$. Then as a direct consequence of Lemma A18, we prove the first claim of (A6.30). The second claim of (A6.30) follows from an analogous argument.

In the rest of the proof we focus on proving the first statement of (A6.31), and the second statement can be prove via a similar argument. To prove this statement, we apply again the decomposition

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

where recall that for any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, $\operatorname{diag}(\boldsymbol{A})$ 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 A4.1, we have that for any constant $\delta > 0$,

$$\sup_{\mathbb{P}_e \in \mathcal{D}_{2+\nu}} \mathbb{P}(|\mathrm{I}| < \delta) \leq \frac{2^{(6+\nu)/(2+\nu)}}{n\delta^2}. \tag{A6.32}$$

For II, we apply the same lines of proof as the control of term II in Section A4.1, except that we replace Lemma A7 with Lemma A19. Putting together, we obtain the desired result. □

A6.4 Theoretical analysis of (18)

Lemma A20. *Let $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ be a completely random permutation matrix. We have that for any fixed $\delta > 0$,*

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

.

Proof. Let $e_{1,1}, e_{1,2}, \ldots, e_{1,n}, \ldots$ and $e_{2,1}, e_{2,2}, \ldots, e_{2,n}, \ldots$ be two sequences of i.i.d. random variables from a distribution $\mathbb{P}_e$. Then apparently if $\mathbb{P}_e \in \mathcal{D}_{1+t}$, $1 \leq \mathbb{E}[|e_{1,i} e_{2,i}|^{1+t}] \leq 4$. Then using Lemma A19, we have from Markov's inequality that for any $\delta > 0$,

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

The desired result then follows from the same lines of proof as in Lemma A9.

□

Lemma A21. *We have*

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

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

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

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

Proof of (18). Following analogous argument as in the proof of Theorem 4, we tackle this problem via proving that for any $j,k\in\{1,\dots,K\}$ and for all $\delta>0$,

$$\begin{aligned}&\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\mathbb{P}\left(\frac{|\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2}\geq\delta\right)\to 0;\\&\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\mathbb{P}\left(\frac{|\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2}\geq\delta\right)\to 0;\end{aligned}\tag{A6.33}$$

and that

$$\begin{aligned}&\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\mathbb{P}\left(\frac{\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{e}-\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{e}}{\|\boldsymbol{e}\|^2}<\frac{1}{5}\right)\to 0;\\&\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\mathbb{P}\left(\frac{\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{e}+\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_k\tilde{\boldsymbol{V}}_k^\top\boldsymbol{P}_k\boldsymbol{e}}{\|\boldsymbol{e}\|^2}<\frac{1}{5}\right)\to 0.\end{aligned}\tag{A6.34}$$

For the first claim of (A6.33), writing $\mathcal{E}:=\{\|\boldsymbol{e}\|_2^2\geq n/16\}$, we have with $\gamma':=1/16$,

$$\begin{aligned}\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\mathbb{P}\left(\frac{|\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2}\geq\delta\right)&\leq\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\mathbb{P}\left(\frac{|\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{\varepsilon}|}{b\|\boldsymbol{e}\|_2^2}\geq\delta\mid\mathcal{E}\right)+\sup_{\mathbb{P}_e\in\tilde{\mathcal{D}}}\mathbb{P}(\mathcal{E}^c)\\&=\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\mathbb{P}\left(\frac{|\boldsymbol{e}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{\varepsilon}|}{b\max\{\|\boldsymbol{e}\|_2^2,\gamma'n\}}\geq\delta\mid\mathcal{E}\right)+\sup_{\mathbb{P}_e\in\tilde{\mathcal{D}}}\mathbb{P}(\mathcal{E}^c)\\&\leq\sup_{\boldsymbol{w}\in\mathbb{R}^n}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\mathbb{P}\left(\frac{|\boldsymbol{w}^\top\tilde{\boldsymbol{V}}_j\tilde{\boldsymbol{V}}_j^\top\boldsymbol{\varepsilon}|}{b\max\{\|\boldsymbol{w}\|_2^2,\gamma'n\}}\geq\delta\mid\mathcal{E}\right)+\sup_{\mathbb{P}_e\in\tilde{\mathcal{D}}}\mathbb{P}(\mathcal{E}^c),\end{aligned}$$

which converges to zero knowing that we have Lemmas A18 and A21. The second claim of (A6.33) can be proven via a similar argument.

We now focus on (A6.34), recall the definitions of $\mathcal{E}_1-\mathcal{E}_5$, it remains to prove that

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

				RB		RR		HDI	
n	p	$\boldsymbol{X}$	noise	1%	0.5%	1%	0.5%	1%	0.5%
300	100	$\mathcal{G}$	$\mathcal{G}$	3.36	2.11	3.72	2.41	7.33	5.46
300	100	$\mathcal{G}$	t_1	2.06	1.42	1.97	1.2	1.86	1.64
300	100	$\mathcal{G}$	t_2	3.07	1.84	3.14	1.9	4.32	2.82
300	100	t_1	$\mathcal{G}$	3.64	2.11	3.74	2.34	70.33	69.33
300	100	t_1	t_1	2.05	1.38	2.04	1.31	47.06	44.56
300	100	t_1	t_2	2.74	1.54	2.83	1.65	64.15	61.9
600	100	$\mathcal{G}$	$\mathcal{G}$	1.93	1.08	1.89	1.06	5.36	4.23
600	100	$\mathcal{G}$	t_1	0.99	0.59	1.3	0.73	1.8	1.6
600	100	$\mathcal{G}$	t_2	1.73	0.93	1.65	0.92	4.38	3.12
600	100	t_1	$\mathcal{G}$	1.93	1.06	1.91	1.01	77.05	76
600	100	t_1	t_1	1.37	0.84	1.29	0.74	51.2	48.5
600	100	t_1	t_2	1.66	0.91	1.58	0.83	71.18	69.46
600	200	$\mathcal{G}$	$\mathcal{G}$	3.55	2.11	3.51	2.26	6.7	5.3
600	200	$\mathcal{G}$	t_1	1.46	0.93	1.75	1.04	1.95	1.75
600	200	$\mathcal{G}$	t_2	2.53	1.34	2.96	1.77	4.33	2.58
600	200	t_1	$\mathcal{G}$	3.92	2.59	3.74	2.34	78.18	76.43
600	200	t_1	t_1	1.68	1.12	1.72	1.03	56.9	54.83
600	200	t_1	t_2	2.55	1.49	2.46	1.36	74.05	72.05

Table A1: Percentage of rejections of various tests under the null, estimated over 100000 Monte Carlo repetitions, for various noise distributions at nominal levels of $\alpha = 1\%$ and $\alpha = 0.5\%$. This table supplements Table 2 in the main text and the same data generation mechanism is used. Percentage signs are omitted.

We can control $\mathcal{E}_1 - \mathcal{E}_5$ following the same lines of proof as in the proof of those events in Section A4.2, except that for $\mathcal{E}_2$ and $\mathcal{E}_3$ we replace Lemma A5 by Lemma A18; for $\mathcal{E}_4$, we replace Lemma A9 and (A4.19) by Lemmas A20 and A21 respectively; and for $\mathcal{E}_5$, we additionally control the uniform convergence of $|\boldsymbol{e}^\top \boldsymbol{e}'|/n$ with Lemma A19.

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

A7 Additional numerical comparisons

We report here additional simulations for sizes of the residual bootstrap (RB) procedure (Freedman, 1981), residual randomization (RR) procedure (Toulis, 2019) and the desparsified Lasso coefficient test as implemented in the `hdi` R package (HDI) (Dezeure et al., 2015) at nominal levels of 1% and 0.5%. As can be seen from Table A1, all the methods are above the nominal size level in the majority of the simulation settings considered here, especially when the design or the noise is heavy-tailed.

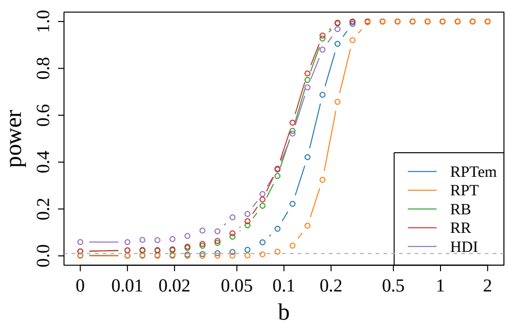
We have focused primarily on size controls at $\alpha = 1\%$ and $\alpha = 0.5\%$ in the main text. This is partly due to the fact that the size invalidity of many procedure are most obvious at small nominal levels. For instance, the distribution of the p-values of ANOVA under the null shown in Figure 1 shows a single large spike near 0. Moreover, in many applications, coefficient tests are conducted multiple times, which necessitates consideration of test validity at small nominal levels due to multiple testing corrections. Nonetheless, for completeness, we include in Table A2 the estimated sizes of all tests considered in our numerical simulations at the 5% nominal level. We observe that at this nominal level, in addition to RPT_{em} and RPT, ANOVA,

n	p	$\boldsymbol{X}$	noise	RPT_{em}	RPT	ANOVA	Naive	DR	FL	CRT	RB	RR	HDI
300	100	$\mathcal{G}$	$\mathcal{G}$	0.1	0	5	5	5	5	0.1	11	11.1	17.0
300	100	$\mathcal{G}$	t_1	1.1	0.5	3.1	3.4	4.7	3.8	3	6.3	6.6	3.3
300	100	$\mathcal{G}$	t_2	0.5	0.1	4.8	4.8	3.9	4.9	1.9	9.7	9.9	10.9
300	100	t_1	$\mathcal{G}$	0.1	0	5	5	9	5	0	11.4	11.2	76.1
300	100	t_1	t_1	0.2	0	4	4.2	5.7	4.4	0.6	6.9	6.9	55.6
300	100	t_1	t_2	0.1	0	5	5	8.5	5	0	9.9	9.8	71.7
600	100	$\mathcal{G}$	$\mathcal{G}$	1.8	0.1	4.9	4.9	4.9	4.9	0	7.5	7.4	13.9
600	100	$\mathcal{G}$	t_1	2.3	1.2	2.5	3.4	4.8	3.8	2.5	4.1	4.6	2.6
600	100	$\mathcal{G}$	t_2	2.3	0.5	4.6	4.8	4	4.9	1.9	6.7	6.6	10.8
600	100	t_1	$\mathcal{G}$	1.7	0.1	4.9	4.9	9.5	4.9	0	7.5	7.4	82.2
600	100	t_1	t_1	1.5	0	2.8	3.8	5.8	4.1	0.4	5.3	5	60.9
600	100	t_1	t_2	1.5	0	4.8	5	9.2	5.1	0	7	6.8	77.8
600	200	$\mathcal{G}$	$\mathcal{G}$	0.1	0	5	4.9	4.9	4.9	0	10.9	11	17.7
600	200	$\mathcal{G}$	t_1	1	0.5	2.5	2.8	4.8	3.4	2.4	5.8	6.1	2.6
600	200	$\mathcal{G}$	t_2	0.5	0.1	4.7	4.7	4.1	4.8	1.9	9	9.8	12.2
600	200	t_1	$\mathcal{G}$	0.1	0	5.1	5.1	9.1	5.1	0	11.4	11.3	82.8
600	200	t_1	t_1	0.3	0	3.3	3.7	5.6	4.1	0.5	6.2	6.2	63.4
600	200	t_1	t_2	0.1	0	4.6	4.7	8.7	4.8	0	9.6	9.4	80.6

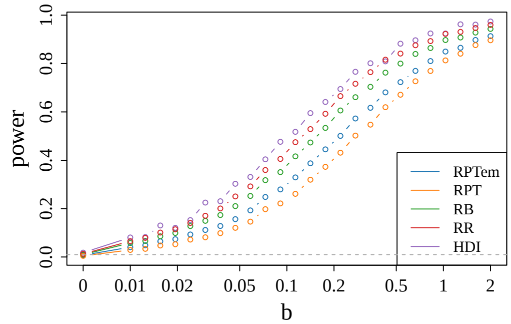
Table A2: Percentage of rejections of various tests under the null, estimated over 100000 Monte Carlo repetitions, for various noise distributions at nominal levels of $\alpha = 5\%$. The data generation mechanism is the same as in Table 2. Percentage signs are omitted.

naive RPT, FL, CRT also show valid size control. This is inline with our observation in Figure 1, where the violation of uniformity of p-value null distributions from ANOVA and naive RPT is mostly manifested through a large spike near 0, which would be smoothed out at higher nominal levels. However, we would like to point out that the task of single coefficient testing is typically carried out multiple times in many applications, and in view of the multiple testing correction needed, it is the size validity at small p-values that are more relevant for practitioners.

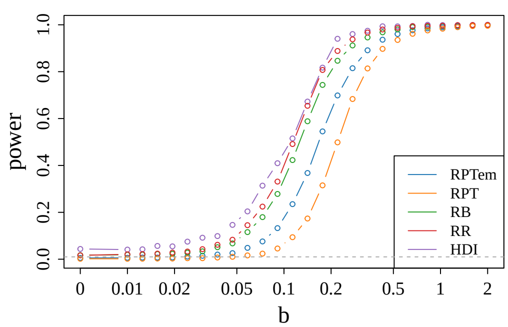
Finally, we report in Figure A1 a power comparison of RPT with the additional methods mentioned in Table A1. RB, RR and HDI procedures exhibit better power than RPT and RPT_{EM} in most of the simulation setups. However, this should be viewed in the context of their above nominal size under the null as reported in Table A1.



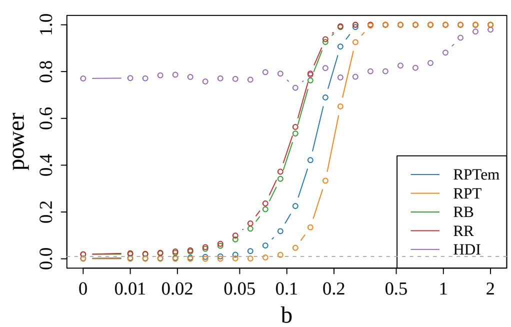
(a) Gaussian design, Gaussian noise



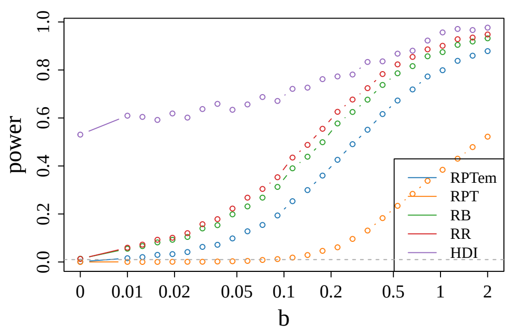
(b) Gaussian design, t_1 noise



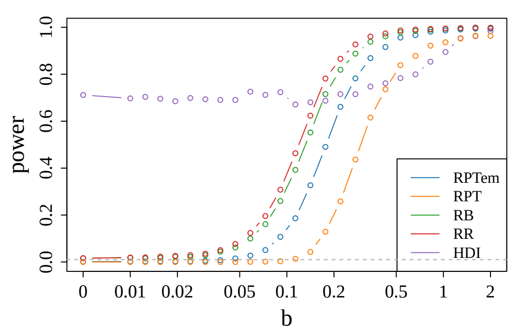
(c) Gaussian design, t_2 noise



(d) t_1 design, Gaussian noise



(e) t_1 design, t_1 noise



(f) t_1 design, t_2 noise

Figure A1: Power (proportion of rejections) with nominal level $\alpha = 0.01$ (represented by the horizontal dashed line) over 10000 replicates for $b = 0$ or on a logarithmic grid between 0.01 and 2. Here $\boldsymbol{X}, \boldsymbol{\varepsilon}$ and $\boldsymbol{e}$ are generated according to various distribution types prescribed in the caption of each figure.