The Annals of Statistics
2025, Vol. 53, No. 2, 724–748
https://doi.org/10.1214/24-AOS2479

RESIDUAL PERMUTATION TEST FOR REGRESSION COEFFICIENT TESTING

BY KAIYUE WEN[1,a], TENGYAO WANG[2,b] AND YUHAO WANG[3,c]

[1]*Department of Computer Science, Stanford University,* [a]*kaiyuew@stanford.edu*
[2]*Department of Statistics, London School of Economics and Political Science,* [b]*t.wang59@lse.ac.uk*
[3]*Institute for Interdisciplinary Information Sciences, Tsinghua University,* [c]*yuhaow@tsinghua.edu.cn*

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-sample valid* size control without requiring the noise to follow strong distributional assumptions. In this paper, we propose a new method, called the *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-sample 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.

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 with $n \geq 2$, 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-sample size 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. Classically, one of the most well-known tests for $b = 0$ in model (1) is the ANOVA test (Fisher (1935)). It is known to be asymptotically valid when p is fixed and $n \to \infty$, and has finite-sample valid size control when the noise possesses spherical symmetry. However, as we will see in Section 3, the finite-sample size of an 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 has finite-sample valid size control under weaker noise assumptions, especially with relatively large p, which is also a topic that has received increasing attention in recent years (see, e.g., Section 2). In particular, we are interested in developing a test with finite-sample valid size control under a fixed design of $\boldsymbol{X}$ and $\boldsymbol{Z}$ by assuming that the noise $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ has *exchangeable components*.

Received November 2022; revised July 2024.
MSC2020 subject classifications. Primary 62J99; secondary 62F03.
Key words and phrases. Permutation test, finite-sample validity, heavy-tail distribution, high-dimensional data.

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

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

We remark that Assumption 1 differs from the common assumption that $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ for fixed design of $\boldsymbol{X}$, $\boldsymbol{Z}$ or that $\mathbb{E}[\boldsymbol{\varepsilon} \mid \boldsymbol{X}, \boldsymbol{Z}] = 0$ (or even the more relaxed assumption $\mathbb{E}[\boldsymbol{\varepsilon}^\top(\boldsymbol{X}, \boldsymbol{Z})] = 0$) when $\boldsymbol{X}$, $\boldsymbol{Z}$ follow a random design, which allows for heteroskedastic noise and typically appears in the regression coefficient tests with asymptotically valid size control. 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-sample size 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-sample 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-sample size validity, we further analyze the statistical power of the proposed test in the regime where p can be up to a constant fraction of n, which we will refer to as the proportional regime in this work, especially when the ε_i's follow a heavy-tailed distribution. It should be noted that just with Assumption 1 we can only guarantee our test to have correct size control, the resulting test may not necessarily have power. Indeed, just with Assumption 1, b may not even be identifiable, so that there is no test that is uniformly valid under the null while still maintaining power against some alternative. For example, in the extreme case where $\boldsymbol{Z} = \boldsymbol{X}\beta^Z$ for some $\beta^Z \in \mathbb{R}^p$, that is, that $\boldsymbol{Z}$ can be expressed as a linear representation of the column vectors of $\boldsymbol{X}$, then b is not an identifiable parameter anymore, that is, no test can have any power. As we can also see in (6), under Assumption 1 and with such choice of $\boldsymbol{Z}$, we can always have $\mathbb{P}(\phi \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$, no matter how large b is. This means that in order for our test to have power, additional assumption on $\boldsymbol{Z}$ must be imposed. 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 proportional regime with heavy-tailed noise. In this paper, we fill this gap by showing that under a suitable modeling assumption of $\boldsymbol{Z}$, 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 regression 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-sample size 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 time, we propose another permutation-based test: naive residual permutation test (naive RPT), which, like ANOVA, also has valid size control under spherically symmetric noise distribution whenever $p < n$. While naive RPT is still not valid for nonspherically 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-sample size 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 $t \in [0, 1]$ and under suitable assumptions of $\boldsymbol{Z}$, 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-sample size 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 $\text{span}(\boldsymbol{A})$ the subspace spanned by the p column vectors of $\boldsymbol{A}$; and we write $\text{span}(\boldsymbol{A})^{\perp}$ as the space that is orthogonal to $\text{span}(\boldsymbol{A})$. Given an n-dimensional vector $\boldsymbol{a}$, we denote by $\text{Proj}_{\boldsymbol{A}}(\boldsymbol{a})$ the projection of $\boldsymbol{a}$ onto the subspace $\text{span}(\boldsymbol{A})$, 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 the variance (ANOVA) test (Fisher (1935)). The 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 (2019)), biology (Lazic (2008)), etc. In the context of single coefficient testing, when $n > p+1$ and $\boldsymbol{\varepsilon}$ follows a spherically symmetric distribution, if $\tilde{\beta} := \arg\min_{\beta} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2$ and $(\hat{\beta}, \hat{b}) := \arg\min_{(\beta,b)} \|\boldsymbol{Y} - \boldsymbol{X}\beta - b\boldsymbol{Z}\|_2^2$, 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, for a general noise distribution of $\boldsymbol{\varepsilon}$, the 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 assumption of ANOVA. In the context of finite-sample size validity, 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-sample size control with only exchangeable noise. However, as mentioned in the Introduction, 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. An alternative test with less restrictive assumptions on dimension p was proposed in D'Haultfœuille and Tuvaandorj (2024), which relaxes the dimensionality assumption by requiring the rows of $\boldsymbol{X}$ to follow a discrete random distribution with a relatively small number of unique values.

Besides finite-sample size validity, a less demanding criteria for the coefficient test is the *asymptotic size validity*. The idea of permutation or randomization has been heavily used to propose an 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, the 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 (2023), to name a few.

More broadly speaking, the regression coefficient test can be viewed as a subdomain of the more general conditional independence testing, that is, 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 tests with valid size control under weak distributional assumptions. In this paper, we show that a linear functional relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is sufficient to have finite-sample size validity with nontrivial power.

2.2. *Permutation- and randomization-based hypothesis tests*. As also mentioned in the Introduction section, our new method is based on the 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), Basu (1980), Rosenbaum (1984), Romano (1990), Kennedy (1995), Rosenbaum (2002), Canay, Romano and Shaikh (2017), Young (2019)). The 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 size validity, without any geometric or moment constraints on the underlying distributions.

For the task of regression coefficient testing, Freedman and Lane (1983) proposed tests based on permuted regression residuals, and analyzed its asymptotic size control 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 (2024) used the permutation test and its extensions to obtain exact size control for testing a single component or a subvector of regression coefficients. We note that although all the works mentioned in this paragraph use permutation tests as the basic building block of their tests, their validity guarantees are based on different assumptions. Among them, Toulis (2019), Lei and Bickel (2021), D'Haultfœuille and Tuvaandorj (2024) considered the size validity of their tests under a similar noise invariance assumption to our Assumption 1.

Other related applications of permutation tests include sharp null hypothesis tests (Caughey, Dafoe and Miratrix (2017), Caughey et al. (2023)), instrumental variable tests (Imbens and Rosenbaum (2005)) and conditional independence tests (Berrett et al. (2020), Kim et al. (2022)).

In addition to the above permutation-based testing, the knockoff-based procedure (Candès et al. (2018)) can also be used to perform asymptotically valid coefficient testing in the regression setting. Another related line of works exploit bootstrap or jackknife techniques to provide tests with asymptotic size validity (e.g., Miller (1974), Freedman (1981), Mammen (1993), Chatterjee (1999), Bickel and Sakov (2008)). See the discussions in Lei and Bickel ((2021), Section A in Supplementary Material) for a comprehensive overview of this area.

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 subexponential tail bounds (see, e.g., Wainwright (2019)). However, as also discussed by Sun, Zhou and Fan (2020), such a 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 10 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 have 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 (2024)) 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-sample size 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 an 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 size 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 has valid size 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 $\boldsymbol{Q} \in \mathbb{O}^{n\times n}$, $\boldsymbol{A} \overset{\mathrm{d}}{=} \boldsymbol{QA}$, 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 (Wen, Wang and Wang (2025)). 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 size validity of an 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 an ANOVA test from 100,000 Monte Carlo simulations. We consider the sizes of the ANOVA test at nominal levels $\alpha = 0.01, 0.05$. 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

TABLE 1

Percentage of rejection of the ANOVA test and naive residual permutation test, estimated over 100,000 *Monte Carlo repetitions, for various noise distributions at nominal levels of* $\alpha = 1\%$ *and* $\alpha = 5\%$. *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%

				ANOVA		Naive	
n	p	X type	noise type	1%	5%	1%	5%
300	100	Gaussian	Gaussian	1.01	4.99	1.00	4.96
300	100	Gaussian	t_1	1.81	3.1	1.58	3.38
300	100	Gaussian	t_2	1.53	4.83	1.39	4.83
300	100	t_1	Gaussian	1.01	4.99	1.03	5.03
300	100	t_1	t_1	2.43	3.96	1.58	4.25
300	100	t_1	t_2	1.80	5.03	1.41	5
600	100	Gaussian	Gaussian	0.95	4.9	0.96	4.88
600	100	Gaussian	t_1	1.63	2.45	1.28	3.36
600	100	Gaussian	t_2	1.69	4.61	1.28	4.79
600	100	t_1	Gaussian	1.05	4.86	1.02	4.87
600	100	t_1	t_1	1.88	2.84	1.06	3.84
600	100	t_1	t_2	1.74	4.79	1.14	5.01
600	200	Gaussian	Gaussian	1.01	4.96	1.03	4.93
600	200	Gaussian	t_1	1.41	2.48	1.24	2.82
600	200	Gaussian	t_2	1.50	4.67	1.36	4.72
600	200	t_1	Gaussian	1.01	5.11	0.98	5.09
600	200	t_1	t_1	2.02	3.26	1.33	3.74
600	200	t_1	t_2	1.70	4.64	1.34	4.66

more than twice as large as the nominal level in some 1%-level tests (this issue is more pronounced if we consider a 0.5% test level; see Table A2 in the Supplementary Material). In particular, the performance of noise type t_1 is in general worse than that of t_2. This means that the 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," 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 permutation test with valid size control 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 $\text{span}(\boldsymbol{X})^\perp$, 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 the *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

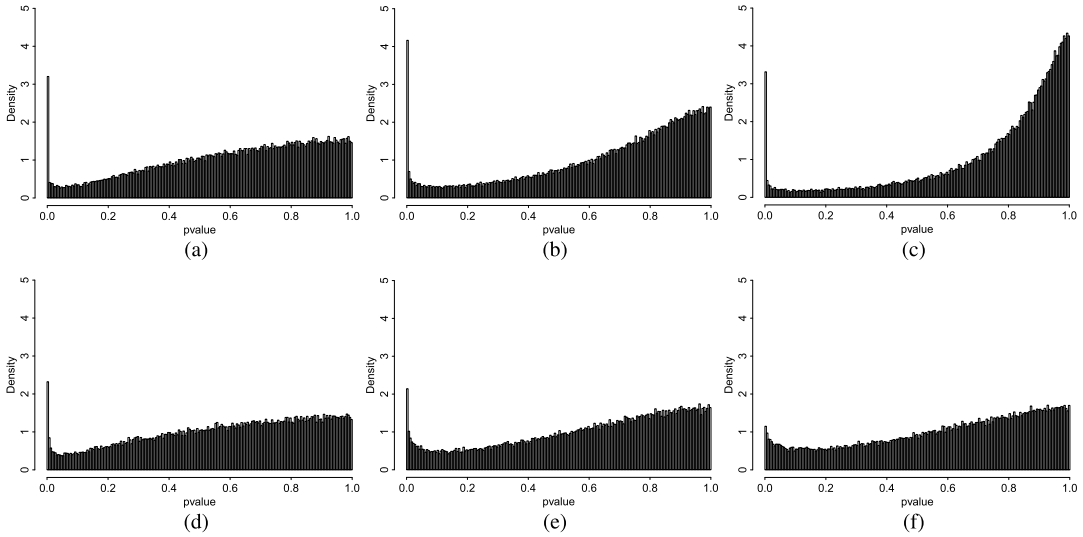


FIG. 1. *Histogram of p-values under the null for ANOVA test and naive residual permutation test from* 100,000 *Monte Carlo replicates. The first line are the histograms of the ANOVA test under different specifications. Specifically,* (*a*) *is the result with Gaussian design,* $n = 300$, $p = 100$ *and* $\boldsymbol{\varepsilon}$ *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 is the histogram for the naive test.* (*e*)–(*f*) *use the same simulation settings as in* (*a*)–(*c*).

$\hat{\boldsymbol{e}} := \boldsymbol{V}_0^\top \hat{\boldsymbol{R}}_e = \boldsymbol{V}_0^\top \boldsymbol{Z}$ as

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

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} has valid Type-I error control.

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 a valid p-value, that is, 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-sample size 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 the 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 preferable alternative to ANOVA in real data analysis when $n/2 \le p < n$.

4. Residual permutation test: Methodology and size validity. In Section 3, we have shown from simulation experiments that a naive permutation test on the residuals, although

more robust than ANOVA, it is still not guaranteed to have finite-sample size 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-sample size validity guarantee in Theorem 2. For intuition behind such construction, we refer the readers to Section 4.1. We will assume throughout this section that the design matrix $(\boldsymbol{X}, \boldsymbol{Z})$ is deterministic.

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 $\operatorname{span}(\boldsymbol{X})^\perp$ 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 $\operatorname{span}(\boldsymbol{V}_0) \cap \operatorname{span}(\boldsymbol{V}_k)$. 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

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

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:

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

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, that is,* $\mathbb{P}(\phi \le \alpha) \le \alpha$ *for all* $\alpha \in [0, 1]$.

We remark that as shown in Theorem 2, an important advantage of RPT is that the result is the finite sample in the sense 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-sample size 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-sample setting with just exchangeable noises. This is because under the null, ϕ_{naive}

[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 $\operatorname{span}(\boldsymbol{V}_0)$ and $\operatorname{span}(\boldsymbol{X})^\perp$ are the same space. Otherwise, $\operatorname{span}(\boldsymbol{V}_0)$ is a subspace of $\operatorname{span}(\boldsymbol{X})^\perp$.

performs permutations on the vector $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_0^\top \boldsymbol{\varepsilon}$ instead of $\boldsymbol{\varepsilon}$ itself. Even if $\boldsymbol{\varepsilon}$ is an exchangeable random vector, $\boldsymbol{V}_0^\top \boldsymbol{\varepsilon}$ may no longer be so, which renders the naive test invalid.

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

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

In light of this transformation, we have that under H_0, $\boldsymbol{V}_k \hat{\boldsymbol{\varepsilon}} = \boldsymbol{V}_k \boldsymbol{V}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon} = \mathrm{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k \boldsymbol{\varepsilon})$, that is, a projection of the noise vector $\boldsymbol{P}_k \boldsymbol{\varepsilon}$ onto the space $\mathrm{span}(\boldsymbol{V}_k)$, and equivalently, $\boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}} = \mathrm{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$. However, this is still not enough, as $\mathrm{Proj}_{\boldsymbol{V}_0}(\boldsymbol{\varepsilon})$ and $\mathrm{Proj}_{\boldsymbol{V}_k}(\boldsymbol{P}_k \boldsymbol{\varepsilon})$ correspond 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 $\mathrm{span}(\tilde{\boldsymbol{V}}_k)$, that is, the intersection of $\mathrm{span}(\boldsymbol{V}_0)$ and $\mathrm{span}(\boldsymbol{V}_k)$. Specifically, using that $\tilde{\boldsymbol{V}}_k$ spans a subspace of $\mathrm{span}(\boldsymbol{V}_k)$, it is straightforward that $\tilde{\boldsymbol{V}}_k^\top = \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \boldsymbol{V}_k^\top$. From this and (7), we have that under H_0,

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

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

In light of the above analysis, we further have that under H_0,

$$a_k := T(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}) = T(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{\varepsilon}),$$

$$b_k := T(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_k \hat{\boldsymbol{\varepsilon}}) = T(\tilde{\boldsymbol{V}}_k^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}, \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}).$$

Writing further that

$$a^* := \min_{\ell \in \{1,\dots,K\}} a_\ell \quad \text{and} \quad b_k^* := \min_{\ell \in \{1,\dots,K\}} T(\tilde{\boldsymbol{V}}_\ell^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}, \tilde{\boldsymbol{V}}_\ell^\top \boldsymbol{P}_k \boldsymbol{\varepsilon}),$$

we can control ϕ as

$$\phi = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\{a^* \le b_k\}\right) \ge \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{1}\{a^* \le b_k^*\}\right), \tag{8}$$

where for the last inequality we use the fact that $b_k^* \le b_k$. Observe that we may also write $a^* = g(\boldsymbol{\varepsilon})$ and $b_k^* = g(\boldsymbol{P}_k \boldsymbol{\varepsilon})$ for $g(\boldsymbol{u}) := \min_{\tilde{\boldsymbol{V}} \in \{\tilde{\boldsymbol{V}}_1,\dots,\tilde{\boldsymbol{V}}_K\}} T(\tilde{\boldsymbol{V}}^\top \boldsymbol{V}_0 \hat{\boldsymbol{\varepsilon}}, \tilde{\boldsymbol{V}}^\top \boldsymbol{u})$, which is a function that depends only on $(\boldsymbol{X}, \boldsymbol{Z}, \mathcal{P}_K)$. This allows us to rewrite the above inequality as

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

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 1 holds, ϕ is a valid p-value.

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

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

We remark that from the above discussion, a more ideal approach would be to compute a^* and b_k^* for $k = 1, \ldots, K$ and construct the test directly using the right-hand side of (8). This test, while almost exact, is unfortunately not directly computable from the data. As a compromise, we replace b_k^* with an upper bound b_k to obtain a feasible test ϕ. As we will see from the numerical simulations, this has resulted in a relatively conservative test.

5. Analysis of statistical power. This section provides power analysis of RPT under mild moment assumptions of noises ε_i and e_i's when 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 $T(x, y) = |\langle x, y \rangle|$. While we continue to assume that the design matrix $\boldsymbol{X}$ is deterministic, $\boldsymbol{Z}$ is assumed to have a random design following specific models in this and the next two sections. 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 $\boldsymbol{e}$.

As mentioned in the Introduction, some assumption on $\boldsymbol{Z}$ is needed for the regression coefficient b to even be identifiable. The structural assumption on $\boldsymbol{Z}$ in (3) is stronger than typically assumed in the regression coefficient testing literature. This is partly due to the fact that previous power results mostly assume a fixed p regime (e.g., Freedman and Lane (1983), DiCiccio and Romano (2017)), or asymptotic regimes where $p = O(n^\gamma)$ for some constant $\gamma < 1$ (e.g., Mammen (1993)); see also the references in the Supplementary Material of Lei and Bickel (2021). On the other hand, when $p \asymp n$, it is not uncommon to see additional structural assumptions on the design matrix. For instance, debiased lasso (Zhang and Zhang (2014)) assumes a nodewise linear regression structure similar to (3) and Lopes (2014) imposed an eigenvalue decay condition on the sample covariance matrix. In addition, the exact form of model (3) is not essential, and is assumed here to simplify our exposition. As we will see later in Section 6, RPT will be asymptotically powerful as long as the quantity defined in (16) is bounded away from zero (Corollary 7). While the modeling assumption in (3) is a sufficient condition for this to hold with asymptotic probability 1, we may relax it to accommodate nonlinear dependence of $\boldsymbol{Z}$ on $\boldsymbol{X}$ and heteroscedastic noise (Theorems 5 and 6). Even if all these models do not work and $\boldsymbol{Z}$ is completely deterministic, Corollary 7 shows that our test is still powerful, provided (16) is large enough, which is an assumption verifiable by practitioners. It would be of interest to propose new tests that have nontrivial power under a nonlinearity assumption weaker than Theorem 6, which we leave for future work.

We also make the 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*

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

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

$$|b| = \Omega\big(n^{-\frac{t}{1+t}}\big) \quad \text{if } t < 1 \quad \text{or} \quad |b| = \omega\big(n^{-\frac{1}{2}}\big) \quad \text{if } t = 1, \tag{9}$$

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

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 works in a proportional regime where the number of covariates can 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 (9) 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 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 nontrivial 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*

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

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 (9), $\lim_{n\to\infty} \mathbb{P}(\phi > \frac{1}{K+1}) = 0$.

The statistical power guarantee in Theorem 3 requires the set of permutations to follow Assumption 4, while the finite-sample size 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 an 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.

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

1 **repeat**

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

3 **for** $k = 1, \dots, K$ **do**

4 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

$$\tilde{\sigma}_k(i) := \begin{cases} i + k & \text{if } i \bmod (K+1) \le K + 1 - k \\ i - (K + 1 - k) & \text{otherwise,} \end{cases} \tag{10}$$

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

5 **end**

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

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

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 (9) 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 force 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 sublist. Specifically, we first denote $S_1, \dots, S_m$ as an m ordered list of indices such that

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

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

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

where each S_i^k is created via shifting all the elements in S_i by k places. Taking S_1^k, for example, means $S_1^k := (K+2-k, \dots, K+1, 1, 2, \dots, K+1-k)$. One can easily verify that the resulting set of permutation matrices $\tilde{\mathcal{P}}_K := \{\boldsymbol{I}, \tilde{\boldsymbol{P}}_1, \dots, \tilde{\boldsymbol{P}}_K\}$ 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 Tth round of such iterations, Algorithm 1 is able to deliver a $\mathcal{P}_K$ satisfying the desired properties with probability at least $1 - \frac{1}{K^T}$.

[2]Notice that the "$\tilde{\sigma}_k$" described here is exactly the same as the "$\tilde{\sigma}_k$" in (10).

Algorithm 2: Residual Permutation Test (RPT)

Input: Design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, additional covariate 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 $\{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}$.

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

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

  Find an orthonormal matrix $\tilde{\boldsymbol{V}}_k \in \mathbb{R}^{n\times(n-2p)}$ such that $\operatorname{span}(\tilde{\boldsymbol{V}}_k) \subseteq \operatorname{span}(\boldsymbol{V}_0) \cap \operatorname{span}(\boldsymbol{V}_k)$.

  Compute

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

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

end

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

PROPOSITION 1. *Given K, T, we have that there exists a $\mathcal{P}_K$ satisfying Assumptions 2 and 4. Moreover, Algorithm 1 always returns a $\mathcal{P}_K$ that satisfies Assumption 2; and with probability at least $1 - \frac{1}{K^T}$, the returned $\mathcal{P}_K$ also satisfies Assumption 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 $O(TKnp)$ and $O(TKnp + Knp^2)$, respectively, where T is the maximum number of iterations. The expected time complexities of the two algorithms are instead $O(Knp)$ and $O(Knp^2)$, respectively. We also remark that due to the construction of the permutation set in Algorithm 1, RPT is inherently a randomized procedure, and unlike permutation tests or bootstrapping, this variability due to randomness cannot be reduced by increasing computational time. It would be of interest to propose a new test that eradicates such reproducibility issue while maintaining all the beneficial features of Algorithm 1, which we leave for future work.

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

$$\frac{|\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{\varepsilon}|}{bn} = o_{\mathbb{P}}(1) \tag{11}$$

(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,

$$\frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} - \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} \gtrsim 1 \quad \text{and} \quad \frac{\boldsymbol{e}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{e} + \boldsymbol{e}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{e}}{n} \gtrsim 1. \tag{12}$$

To prove (11), 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 (12) 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,

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

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 zero-mean random variables $\varepsilon_1, \varepsilon_2, \ldots$ and $e_1, e_2, \ldots$, 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

$$\lim_{a \to \infty} \sup_{n \ge 1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[e_i^2 \mathbb{1}(e_i^2 \ge a)\big] = 0 \quad \text{and} \quad \liminf_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[e_i^2\big] > c_e; \tag{13}$$

- 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 $B > 0$,

$$\sum_{i=1}^{\infty} \mathbb{P}\big(|\varepsilon_i|^{1+t} \ge Bi\big) < \infty. \tag{14}$$

Informally speaking, instead of requiring all noises to be i.i.d., Assumption 5 allows noises to be heteroscedastic, under certain restrictions on the degree of heteroscedasticity of ε_i's and e_i's. To intuitively understand (14) and the first equation in (13), taking (14), for example, a sufficient condition for it to hold is that there exists a zero-mean random variable ε_∞ satisfying that $\mathbb{E}[|\varepsilon_\infty|^{1+t}] < \infty$ and that for any $1 \le i < \infty$, $|\varepsilon_i| \preceq_d |\varepsilon_\infty|$, that is, that $|\varepsilon_i|$ is stochastically dominated by $|\varepsilon_\infty|$ uniformly for all ε_i's. When such ε_∞ exists, for any $n \ge 1$,

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

which satisfies (14). Analogously, when there exists a zero-mean random variable e_∞ with $\mathbb{E}[|e_\infty|^2] < \infty$ and $|e_\infty|$ stochastically dominates all $|e_i|$'s, we can also have

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

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

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

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

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

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

which, from the 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 $m > 0$, $r < c_e$, $\limsup_{n\to\infty} \|\boldsymbol{h}\|_2^2/n \leq r$, $n > (3C_e/(c_e - r) + m)p$ and b satisfies (9), we have $\lim_{n\to\infty} \mathbb{P}(\phi > \frac{1}{K+1}) = 0$.*

When $\mathcal{P}_K$ does not satisfy Assumption 4, the same conclusion in Theorem 6 still holds with $n > (4C_e/(c_e - r) + m)p$; see also the analogous comment after Theorem 3. Notice also that 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

$$|b| = \Omega\big(z_n^{-1} n^{-\frac{t}{1+t}}\big) \quad \text{if } t < 1 \quad \text{or} \quad |b| = \omega\big(z_n^{-1} n^{-\frac{1}{2}}\big) \quad \text{if } t = 1, \tag{15}$$

where

$$z_n := \left(\frac{\|\boldsymbol{V}_0^\top \boldsymbol{Z}\|_2}{\sqrt{n}}\right)^{-1} \cdot \min_{1\leq j,k\leq K} \min_{z\in\{0,1\}} \frac{\boldsymbol{Z}^\top \tilde{\boldsymbol{V}}_j \tilde{\boldsymbol{V}}_j^\top \boldsymbol{Z} + (-1)^z \boldsymbol{Z}^\top \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{V}}_k^\top \boldsymbol{P}_k \boldsymbol{Z}}{n}. \tag{16}$$

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 $n \geq 3$. In the asymptotic regime where b satisfies (15), we have $\lim_{n\to\infty} \mathbb{P}(\phi > \frac{1}{K+1}) = 0$.*

An inspection of the proof of Theorem 6 reveals that 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 (15) and (9) coincide. In practice, one can choose $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K$ to maximize (16).

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 $\delta > 0$,*

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

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 the 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 $n \to \infty$,

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{n-1}} \mathbb{P}\left(|\boldsymbol{w}^\top(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])| > \frac{\delta}{2} \cdot n^{\frac{1-t}{2(1+t)}}\right) \to 0. \tag{17}$$

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 the 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

tth-order moment bounded between $[1, 2]$, that is, for some $t > 0$ and some random variable ξ with distribution $\mathbb{P}_\xi$,

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

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

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

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 $\boldsymbol{X}$, β, β^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

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

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 $H_1 : b \neq 0$ 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 $\boldsymbol{X}$,*

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

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*

$$\lim_{n\to\infty}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\sup_{\mathbb{P}_e\in\mathcal{D}_{2+\nu}}\mathbb{P}\bigg(\phi > \frac{1}{K+1}\bigg) = 0. \tag{18}$$

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

$$\lim_{n\to\infty}\sup_{\mathbb{P}_\varepsilon\in\mathcal{D}_{1+t}}\sup_{\mathbb{P}_e\in\mathcal{D}_{1+\nu}\cap\tilde{\mathcal{D}}}\mathbb{P}\bigg(\phi > \frac{1}{K+1}\bigg) = 0. \tag{19}$$

In Theorem 10, the separation rate is slightly worse than (9) 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 data sets. 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 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 5 components sampled uniformly on the sphere $\mathcal{S}^4$ and the rest of the 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 Candès et al. (2018), the residual randomization (RR) procedure of Toulis (2019), the desparsified lasso 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.

8.2. *Numeric analysis of validity under the null.* We start by analyzing 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 5% for $(n, p) \in \{(300, 100), (600, 100), (600, 200)\}$ (see Table A2 in the Supplementary Material for the estimated size at the 0.5% 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 summarized 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 size 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

TABLE 2

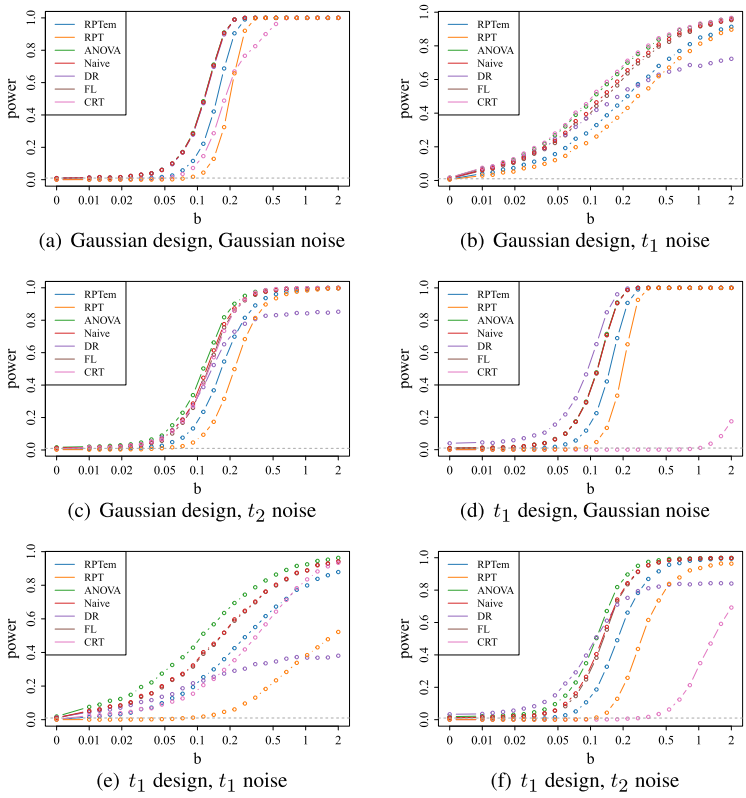
Percentage of rejections of various tests under the null, estimated over 100,000 *Monte Carlo repetitions, for various noise distributions at nominal levels of* $\alpha = 1\%$ *and* $\alpha = 5\%$. *Data are generated from the model in* (1) *and* (3) *with* $b = 0$. $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{\epsilon}$ *are generated according to the various distribution types prescribed in the table. Here, "*$\mathcal{G}$*" stands for standard normal distribution. Percentage signs are omitted*

				RPT_{EM}		RPT		DR		FL		CRT	
n	p	$\boldsymbol{X}$	Noise	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
300	100	$\mathcal{G}$	$\mathcal{G}$	0	0.08	0	0	0.98	5.04	0.99	5.02	0	0.05
300	100	$\mathcal{G}$	t_1	0.51	1.12	0.24	0.5	0.88	4.66	1.28	3.8	1.89	3.03
300	100	$\mathcal{G}$	t_2	0.14	0.45	0.04	0.09	0.67	3.88	1.23	4.91	0.53	1.86
300	100	t_1	$\mathcal{G}$	0	0.09	0	0	3.33	9.02	1.01	4.99	0	0
300	100	t_1	t_1	0.01	0.23	0	0	1.28	5.67	1.21	4.35	0.33	0.59
300	100	t_1	t_2	0	0.09	0	0	2.54	8.45	1.09	5	0	0.01
600	100	$\mathcal{G}$	$\mathcal{G}$	0.21	1.77	0.01	0.06	0.95	4.91	0.95	4.91	0	0.04
600	100	$\mathcal{G}$	t_1	0.73	2.31	0.48	1.24	0.92	4.77	1.09	3.82	1.68	2.47
600	100	$\mathcal{G}$	t_2	0.61	2.29	0.2	0.55	0.68	4.04	1.09	4.87	0.61	1.94
600	100	t_1	$\mathcal{G}$	0.23	1.72	0.01	0.08	3.95	9.52	0.93	4.92	0	0
600	100	t_1	t_1	0.13	1.49	0	0.01	1.37	5.76	1.04	4.15	0.25	0.42
600	100	t_1	t_2	0.1	1.54	0	0.02	3.33	9.25	1.05	5.05	0.01	0.01
600	200	$\mathcal{G}$	$\mathcal{G}$	0	0.12	0	0	1.04	4.94	1.02	4.94	0	0.04
600	200	$\mathcal{G}$	t_1	0.46	1.02	0.26	0.51	0.89	4.77	1.18	3.41	1.5	2.37
600	200	$\mathcal{G}$	t_2	0.12	0.5	0.04	0.1	0.68	4.08	1.2	4.82	0.49	1.94
600	200	t_1	$\mathcal{G}$	0	0.12	0	0	3.45	9.07	0.98	5.11	0	0
600	200	t_1	t_1	0.01	0.26	0	0	1.25	5.61	1.13	4.12	0.27	0.49
600	200	t_1	t_2	0	0.1	0	0	2.71	8.74	1.01	4.75	0	0.01

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, though sometimes conservative, 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}.

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 10,000 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 finite-sample size validity with only exchangeable noise in the proportional regime. Moreover, RPT is nevertheless still guaranteed to reject the alternative with high probability given a signal size b not too much



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

(c) Gaussian design, t_2 noise (d) t_1 design, Gaussian noise

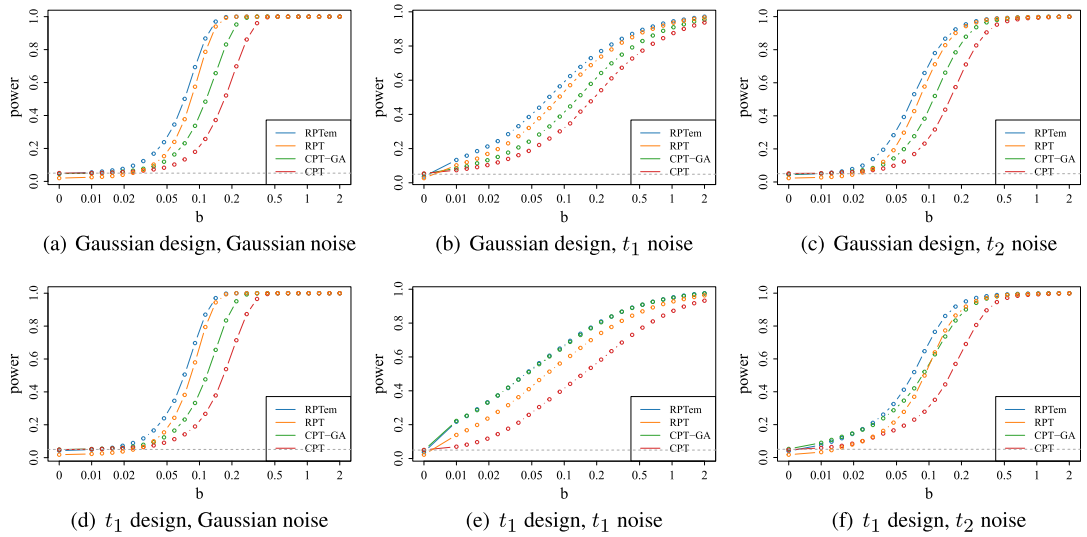
(e) t_1 design, t_1 noise (f) t_1 design, t_2 noise

FIG. 2. *Power (proportion of rejections) with nominal level* $\alpha = 0.01$ *(represented by the horizontal dashed line) over* 10,000 *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.*

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 size 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 proposed in Lei and Bickel (2021). As the cyclic permutation test is not well-defined for $n/p < 1/\alpha + 1$, we consider a relatively low-dimensional setting where $n = 1000$, $p = 40$ and $\alpha = 0.05$. We consider the test with both the default variable ordering (CPT) and a preordering computed using a genetic algorithm (CPT-GA). Due to computational limitations, the genetic algorithm is computed with only 1000 random initializations. The data generation mechanism is the same as that described in Section 8.1, except that to adapt the high computational cost of CPT-GA, for each specification we first generate 10 repetitions of $(\boldsymbol{X}, \boldsymbol{Z})$, then for each $(\boldsymbol{X}, \boldsymbol{Z})$, we generate 1000 repetitions of $\boldsymbol{\varepsilon}$, summing up to 10,000 repetitions. Figure 3 shows the power curves of RPT_{EM}, RPT, CPT-GA and CPT under various design matrix and noise distributions. We see that all four methods are well-calibrated at 5% level when $b = 0$, with RPT slightly more conservative than the other three approaches. 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. For CPT-GA, the performance of RPT and CPT-GA are comparable; and CPT-GA can significantly outperform RPT when both the design and noise



(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

FIG. 3. *Power* (*proportion of rejections*) *with nominal level* $\alpha = 0.05$ (*represented by the horizontal dashed line*) *over* 10,000 *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*.

are heavy-tailed. Moreover, we have found that genetic algorithm can significantly increase the power of CPT, which is consistent with the observation from Lei and Bickel (2021). From Lei and Bickel (2021), it is expected that once we increase the number of random initializations for genetic algorithm from 1000 to the recommended setting of 10,000, CPT-GA can become more powerful.

9. Discussion. In this paper, we propose a new method for the fixed-design regression coefficient test when the number of covariates p can be as large as a fraction of the sample size n. RPT is a permutation-based approach that exploits the exchangeability of the noise terms to achieve finite-sample size control. Our approach uses the fact that the empirical residuals of the classical OLS fit is equivalent to the projection of the noise vector onto an subspace orthogonal to the design to construct a test with valid size 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 byproduct, we propose RPT_{EM} and demonstrate its size validity and power via numerical experiments. It would be of interest to understand the theoretical properties of RPT_{EM} in a future study.

In the regime where $n/2 \le p < n$, we propose the naive RPT, and prove its finite-sample size 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 the ANOVA test, which is of independent interest for practitioners interested in ANOVA.

In this paper, permutation test facilitates an important basis for construction of our test. This sheds light on extending permutation tests to solve other problems in modern statistics, which we leave as future work. In addition, permutation tests and its related 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.

Acknowledgments. KW and TW have contributed equally. Part of this work was done while KW was at the Institute for Interdisciplinary Information Sciences, Tsinghua University. The authors would like to thank the Editor, the Associate Editor and two anonymous referees for helpful comments that improved the paper. YW is also affiliated with Shanghai Qi Zhi Institute. Correspondence should be addressed to Yuhao Wang.

Funding. The research of TW was supported by EPSRC Grant EP/T02772X/2.

The research of YW was supported by the grants of National Natural Science Foundation of China (NSFC) 12201341, National Key R & D Program (2022YFA1008100), and Shanghai Qi Zhi Institute Innovation Program SQZ202304.

SUPPLEMENTARY MATERIAL

Supplementary Material to "Residual permutation test for regression coefficient testing" (DOI: 10.1214/24-AOS2479SUPP; .pdf). The Supplementary Material contains additional power analysis of RPT when K diverges with n. The proofs of all the theoretical statements appeared in the paper and additional simulation studies.

REFERENCES

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877 https://doi.org/10.1214/11-AOS910

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. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

BASU, D. (1980). Randomization analysis of experimental data: The Fisher randomization test. *J. Amer. Statist. Assoc.* **75** 575–595. MR0590687

BERRETT, T. B., WANG, Y., BARBER, R. F. and SAMWORTH, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 175–197. MR4060981

BICKEL, P. J. and SAKOV, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statist. Sinica* **18** 967–985. MR2440400

BRADIC, J., CHERNOZHUKOV, V., NEWEY, W. K. and ZHU, Y. (2019). Minimax semiparametric learning with approximate sparsity. Preprint. Available at arXiv:1912.12213.

BUBECK, S., CESA-BIANCHI, N. and LUGOSI, G. (2013). Bandits with heavy tail. *IEEE Trans. Inf. Theory* **59** 7711–7717. MR3124669 https://doi.org/10.1109/TIT.2013.2277869

CANAY, I. A., ROMANO, J. P. and SHAIKH, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica* **85** 1013–1030. MR3664187 https://doi.org/10.3982/ECTA13081

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

CAUGHEY, D., DAFOE, A. and MIRATRIX, L. (2017). Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. Preprint. Available at arXiv:1709.07339.

CAUGHEY, D., DAFOE, A., LI, X. and MIRATRIX, L. (2023). Randomisation inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 1471–1491. MR4718545 https://doi.org/10.1093/jrsssb/qkad080

CHATTERJEE, S. B. (1999). *Generalised Bootstrap Techniques*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Indian Statistical Institute—Kolkata. MR4380488

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. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CONT, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **1** 223.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

D'HAULTFŒUILLE, X. and TUVAANDORJ, P. (2024). A robust permutation test for subvector inference in linear regressions. *Quant. Econ.* **15** 27–87. MR4703622

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527

DICICCIO, C. J. and ROMANO, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *J. Amer. Statist. Assoc.* **112** 1211–1220. MR3735371 https://doi.org/10.1080/01621459.2016.1202117

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

DURRETT, R. (2019). *Probability—Theory and Examples*. *Cambridge Series in Statistical and Probabilistic Mathematics* **49**. Cambridge Univ. Press, Cambridge. MR3930614 https://doi.org/10.1017/9781108591034

EKLUND, A., NICHOLS, T. E. and KNUTSSON, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. USA* **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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972 https://doi.org/10.1111/rssb.12166

FAN, J., WANG, W. and ZHU, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Ann. Statist.* **49** 1239–1266. MR4298863 https://doi.org/10.1214/20-aos1980

FREEDMAN, D. and LANE, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econom. Statist.* **1** 292–298.

FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228. MR0630104

HARTIGAN, J. A. (1970). Exact confidence intervals in regression problems with independent symmetric errors. *Ann. Math. Stat.* **41** 1992–1998. MR0267701 https://doi.org/10.1214/aoms/1177696700

IMBENS, G. W. and ROSENBAUM, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *J. Roy. Statist. Soc. Ser. A* **168** 109–126. MR2113230 https://doi.org/10.1111/j.1467-985X.2004.00339.x

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

KENNEDY, P. E. (1995). Randomization tests in econometrics. *J. Bus. Econom. Statist.* **13** 85–94. MR1323048 https://doi.org/10.2307/1392523

KIM, I., NEYKOV, M., BALAKRISHNAN, S. and WASSERMAN, L. (2022). Local permutation tests for conditional independence. *Ann. Statist.* **50** 3388–3414. MR4524501 https://doi.org/10.1214/22-aos2233

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

LEI, J., ROBINS, J. and WASSERMAN, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108** 278–287. MR3174619 https://doi.org/10.1080/01621459.2012.751873

LEI, L. and BICKEL, P. J. (2021). An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika* **108** 397–412. MR4259139 https://doi.org/10.1093/biomet/asaa079

LOH, P.-L. and TAN, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electron. J. Stat.* **12** 1429–1467. MR3804842 https://doi.org/10.1214/18-EJS1427

LOPES, M. (2014). A residual bootstrap for high-dimensional regression with near low-rank designs. *Adv. Neural Inf. Process. Syst.* **27**.

LUGOSI, G. and MENDELSON, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190. MR4017683 https://doi.org/10.1007/s10208-019-09427-x

LUGOSI, G. and MENDELSON, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** 393–410. MR4206683 https://doi.org/10.1214/20-AOS1961

LYKOURIS, T., MIRROKNI, V. and PAES LEME, R. (2018). Stochastic bandits robust to adversarial corruptions. In *STOC'*18*—Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 114–122. ACM, New York. MR3826238 https://doi.org/10.1145/3188745.3188918

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176 https://doi.org/10.1214/aos/1176349025

MEINSHAUSEN, N. (2015). Group bound: Confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 923–945. MR3414134 https://doi.org/10.1111/rssb.12094

MILLER, R. G. JR. (1974). An unbalanced jackknife. *Ann. Statist.* **2** 880–891. MR0356353

PAOLELLA, M. S. (2019). *Linear Models and Time-Series Analysis*: *Regression*, *ANOVA*, *ARMA and GARCH*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3839332

PENSIA, A., JOG, V. and LOH, P.-L. (2024). Robust regression with covariate filtering: Heavy tails and adversarial contamination. *J. Amer. Statist. Assoc.* 1–12.

PITMAN, E. J. (1937a). Significance tests which may be applied to samples from any populations. *Suppl. J. R. Stat. Soc.* **4** 119–130.

PITMAN, E. J. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Suppl. J. R. Stat. Soc.* **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. *J. Amer. Statist. Assoc.* **85** 686–692. MR1138350

ROMANO, Y., PATTERSON, E. and CANDES, E. (2019). Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **32**.

ROSENBAUM, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574. MR0763575

ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487 https://doi.org/10.1214/ss/1042727942

SHAH, R. D. and BÜHLMANN, P. (2023). Double-estimation-friendly inference for high-dimensional misspecified models. *Statist. Sci.* **38** 68–91. MR4535395 https://doi.org/10.1214/22-sts850

SHAH, R. D. and PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** 1514–1538. MR4124333 https://doi.org/10.1214/19-AOS1857

STIGLER, S. M. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard Univ. Press, Cambridge, MA. MR3585675 https://doi.org/10.4159/9780674970199

SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265. MR4078461 https://doi.org/10.1080/01621459.2018.1543124

TOULIS, P. (2019). Invariant inference via residual randomization. Preprint. Available at 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. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

WANG, L. (2013). The L_1 penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.* **120** 135–151. MR3072722 https://doi.org/10.1016/j.jmva.2013.04.001

WANG, L., PENG, B. and LI, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.* **110** 1658–1669. MR3449062 https://doi.org/10.1080/01621459.2014.988215

WEN, K., WANG, T. and WANG, Y. (2025). Supplement to "Residual permutation test for regression coefficient testing." https://doi.org/10.1214/24-AOS2479SUPP

YOUNG, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Q. J. Econ.* **134** 557–598.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHU, Y. and BRADIC, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* **113** 1583–1600. MR3902231 https://doi.org/10.1080/01621459.2017.1356319