*The Annals of Statistics*
2024, Vol. 52, No. 5, 2059–2080
https://doi.org/10.1214/24-AOS2421


# A CONFORMAL TEST OF LINEAR MODELS VIA PERMUTATION-AUGMENTED REGRESSIONS

BY LEYING GUAN[a]

*Department of Biostatistics, Yale University,* [a]*leying.guan@yale.edu*

Permutation tests are widely recognized as robust alternatives to tests based on normal theory. Random permutation tests have been frequently employed to assess the significance of variables in linear models. Despite their widespread use, existing random permutation tests lack finite-sample and assumption-free guarantees for controlling type I error in partial correlation tests. To address this ongoing challenge, we have developed a conformal test through permutation-augmented regressions, which we refer to as PALMRT. PALMRT not only achieves power competitive with conventional methods but also provides reliable control of type I errors at no more than $2\alpha$, given any targeted level $\alpha$, for arbitrary fixed designs and error distributions. We have confirmed this through extensive simulations.

Compared to the cyclic permutation test (CPT) and residual permutation test (RPT), which also offer theoretical guarantees, PALMRT does not compromise as much on power or set stringent requirements on the sample size, making it suitable for diverse biomedical applications. We further illustrate the differences in a long-Covid study where PALMRT validated key findings previously identified using the t-test after multiple corrections, while both CPT and RPT suffered from a drastic loss of power and failed to identify any discoveries. We endorse PALMRT as a robust and practical hypothesis test in scientific research for its superior error control, power preservation, and simplicity.

## 1. Introduction.

Consider a linear regression model

$$y_i = x_i\beta + \mathbf{z}_i^\top \theta + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where features $x_i \in \mathbb{R}$, $\mathbf{z}_i \in \mathbb{R}^p$ and $\varepsilon_i$ are random errors independent of $(x_i, \mathbf{z}_i)$. Testing whether the coefficient $\beta$ is zero in this linear model, which equates to examining the partial correlation between $x$ and $y$, remains a fundamental statistical query and a prevalent approach in applications like biological signature discovery. For instance, a key question in the recent MY-LC study is whether long-COVID (LC) is associated with specific cell type proportions, after adjusting for age, sex, and Body Mass Index (BMI) [24]:

$$\text{cell type frequency} \sim \text{intercept} + \text{LC} + \text{age} + \text{sex} + \text{BMI} + \text{age} \times \text{BMI} + \text{sex} \times \text{BMI}. \tag{2}$$

While the F/t-test is a standard approach [12–14], it may yield anticonservative p-values under ill-behaved error distributions or limited sample sizes, compromising the reliability of scientific discoveries. Therefore, there is a need for a valid hypothesis test that minimizes assumptions on noise distribution and does not rely on asymptotic theory. Motivated by this, we seek a robust and straightforward testing procedure for partial correlation under the sole assumption of exchangeability.

ASSUMPTION 1.1. The noises $\varepsilon_1, \ldots, \varepsilon_n$ are exchangeable with each other.

Received September 2023; revised June 2024.
*MSC2020 subject classifications.* 62J20, 62H15, 62F35.
*Key words and phrases.* Partial correlation, random permutation, assumption-free, conformal test.

The problem of testing for partial correlation has been intensively studied, with a focus on developing various random permutation methods to increase robustness against diverse noise distributions. Draper and Stoneman [7] introduced a permutation technique for partial correlation testing, referred to as PERMtest, which involves shuffling $x$ alone and disrupts the relationship between $x$ and $Z$. Subsequent methods better accounted for this relationship. A family of methods, such as the Freedman and Lane test (FLtest) [15] and the Kenny test [22], permutes response residuals obtained from a reduced model that regresses $y$ on $Z$. Another type of methods, such as the the Braak test [30], shuffles residuals from a full model regressing $y$ on both $x$ and $Z$. These methods have demonstrated empirically robust type I error control in various benchmark studies when using pivotal test statistics like t-test or F-test statistics [1, 20, 35, 36]. However, finite-sample and assumption-free theoretical guarantees for these random permutation tests remain elusive. This contrasts with permutation tests for simple correlation, which are theoretically sound in terms of type I error control [8, 26, 28]. Indeed, even the widely examined and recommended FLtest can, under specific conditions, yield drastically inflated type I error rates as we demonstrate later.

Recently, Lei and Bickel [25] introduced the cyclic permutation test (CPT), which offers worst-case guarantees for controlling type I error. Given a sample size $n$ and a total feature dimension $p$ for $z$, CPT theoretically ensures type I error control at a target level $\alpha$, provided $n > (\frac{1}{\alpha} - 1)p$. This condition is often challenging to meet in biomedical studies. For instance, in immunology research, sample sizes frequently hover around 100 or a few hundreds, while investigating simultaneously hundreds or even more different biomarkers for their partial correlations with a primary feature of interest. Although applying CPT with a $\alpha = 0.05$ cutoff may suffice with fewer concomitant covariates, the situation complicates when multiple hypothesis corrections are applied, as smaller nominal $p$-value thresholds are needed to maintain a reasonable false discovery rate (FDR). In an independent pursuit of robust partial correlation analysis with increased power and reduced sample size, Wen, Wang and Wang [34] introduces the residual permutation test (RPT), defined under the condition $n > 2p$. This test utilizes a sequence of specially designed permutation matrices, which, in conjunction with the identity matrix, form a group. However, there's a trade-off in the choice of the size parameter $K$: a small permutation set size limits testing for small p-values, while a large $K$ reduces RPT power significantly by design (see Table 1). Obtaining a sequence of well-designed permutation matrices with a reasonable size, such as $(\lceil \frac{1}{\alpha} \rceil - 1)$ as suggested by Wen, Wang and Wang [34], remains challenging for small $\alpha$. Following the proposed RPT Algorithm [34], a nontrivial test requires $\alpha > \frac{1}{n}$.

TABLE 1

*Summary of representative existing permutation methods and PALMRT, on their construction, empirical performance, theoretical guarantee, and dimension constraints for nontrivial rejection under the i.i.d. Gaussian design. See Section 1.1 for more explanations*

| Method | Original Model | Permuted Model | Empirical | Worst-Case | Restriction |
|---|---|---|---|---|---|
| PERMtest | $y \sim x + Z$ | $y \sim x_\pi + Z$ | $\lesssim \alpha$ | $\gg \alpha$ | $n > p$ |
| FLtest | $y \sim x + Z$ | $[(I - H^z)y]_\pi \sim x + Z$ | $\lesssim \alpha$ | $\gg \alpha$ | $n > p$ |
| Kenny test | $y \sim x + Z$ | $[(I - H^z)y]_\pi \sim (I - H^z)x$ | $\lesssim \alpha$ | $\gg \alpha$ | $n > p$ |
| Braak test | $y \sim x + Z$ | $H^{xz}y + [(I - H^{xz})y]_\pi \sim x + Z$ | $\lesssim \alpha$ | $\gg \alpha$ | $n > p + 1$ |
| CPT | $\eta^\top y$ | $\eta^\top y_{(kL+1):(kL+n)}$ | $\approx \alpha$ | $\leq \alpha$ | $\frac{n}{p} > \frac{1}{\alpha} - 1$ |
| RPT | $\min_{\tilde{V} \in \{\tilde{V}_1, \ldots, \tilde{V}_m\}} \lvert x^\top \tilde{V} \tilde{V}^\top y \rvert$ | $\lvert x^\top \tilde{V}_k \tilde{V}_k^\top y_{\pi_k} \rvert$ | $\lesssim \alpha$ | $\leq \alpha$ | $n > \max(2p, \frac{1}{\alpha})$ |
| PALMRT | $y \sim x + Z + Z_\pi$ | $y \sim x_\pi + Z_\pi + Z$ | $\lesssim \alpha$ | $\leq 2\alpha$ | $n > 2p$ |

In this manuscript, we develop PALMRT, which stands for Permutation-Augmented Linear Model Regression Test, as a conformal test to examine the partial correlation relationships in a linear model. Unlike traditional random permutation methods, PALMRT not only empirically controls type I error but also theoretically guarantees a worst-case coverage of $2\alpha$ at any targeted error level $\alpha$. It offers well-calibrated $p$-values for finite-sample type I error control under arbitrary fixed designs or noise distributions.

In our empirical analyses, PALMRT not only maintains empirical Type I error rates below the designated thresholds across diverse simulation settings but also exhibits comparable power to established methods like FLtest when $(n/p)$ is moderately large. It significantly outperforms CPT and RPT in commonly encountered scenarios. Upon re-analyzing the MY-LC study, PALMRT validates the top findings from the original study, while CPT and RPT show reduced discoveries and fail to confirm any main findings after multiple test corrections. Consequently, we recommend adopting PALMRT as the default robust procedure for detecting significant partial correlations in our daily research.

1.1. *Comparison of permutation tests for linear models.* Table 1 summarizes various methods in terms of their construction, empirical performance, theoretical guarantees, and dimension constraints for nontrivial rejection under the i.i.d. Gaussian design. Let $H^z$ and $H^{xz}$ represent the projection matrices onto the column spaces of the concomitant features $Z$ and all features $(x, Z)$, respectively. Given a permutation $\pi$ of $(1, \ldots, n)$, $x_\pi$ and $Z_\pi$ denote the permuted versions of $x$ and $Z$. The intercept term is omitted for brevity in Table 1.

The columns labeled "Original Model" and "Permuted Model" detail the models used to generate the original and permuted test statistics, respectively. Specifically: (1) For CPT, $\eta$ is a length-$n$ vector used to define valid cyclic permutations and generate $m$ cyclic permutation copies. Define $L = \lfloor \frac{n}{m} \rfloor$ as the cyclic step size. CPT requires $\eta^\top Z_{(kL+1):(kL+n)}$ to be a constant vector for all $0 \le k \le m$, where an index greater than $n$ loops back to the beginning (e.g., $x_{kL+n} = x_{kL}$). (2) Let $V \in \mathbb{R}^{n\times(n-p)}$ be an orthonormal matrix orthogonal to the column space of $Z$. RPT generates $m$ permutations $\pi_1, \ldots, \pi_m$ based on a variant of cyclic permutation to satisfy the group requirement, resulting in $V_{\pi_k}$ as the row-permuted version of $V$ for $k = 1, \ldots, m$. RPT then constructs $\tilde{V}_k \in \mathbb{R}^{n\times(n-2p)}$ as the orthonormal matrix belonging to the intersection of the column spaces of $V$ and $V_{\pi_k}$.

The "Empirical" column indicates empirical type I error control from previous studies (the statement on PALMRT is based on our study). The "Worst-case" column indicates the finite-sample and assumption-free theoretical guarantee for a given level $\alpha$, with "$\gg \alpha$" indicating no established guarantee. The "Restriction" column specifies the minimum sample size when a test is defined and can be nontrivial under the i.i.d. Gaussian design. For all methods except CPT and RPT, F-statistics are used for comparison.

## 2. Conformal test via permutation-augmented regressions.

2.1. *Construction of PALMRT.* Let $x \in \mathbb{R}^n$ be the target feature, and $Z$ the observation matrix with rows $\mathbf{z}_1, \ldots, \mathbf{z}_n$. Define $[n]$ as the vector $(1, \ldots, n)$ and $\pi$ as a permutation of $[n]$. Denote $x_\pi$ and $Z_\pi$ as the row-permuted versions of $x$ and $Z$, respectively. PALMRT is a random permutation method for testing partial correlation. For each randomly generated permutation $\pi$ of $[n]$, it constructs "original" and "permuted test" statistics based on a pair of permutation-augmented regressions:

- Original test statistic $T_{\text{original}}$: the F-statistic for significance of $x$ in the model $y \sim x + Z + Z_\pi$:

$$T_{\text{original}} = \big(\|(I - H^{zz_\pi})y\|_2^2 - \|(I - H^{xzz_\pi})y\|_2^2\big)/\|(I - H^{xzz_\pi})y\|_2^2.$$

**Algorithm 1** A conformal test for partial correlation via paired regression

**procedure** PALMRT($y$, $x$, $Z$, $B$) ▷ Test for partial correlation $y \sim x|Z$ using $B$ permutations
  **for** $b = 1, \ldots, B$ **do**
    Generate a random permutation $\pi_b$.
    Construct a pair of statistics as in eq. (3) with $\pi \leftarrow \pi_b$:

$$(T_{0b}, T_{b0}) \leftarrow (T_{\text{original}}, T_{\text{perm}})$$

    Set $\omega_b = \frac{1}{2}\mathbb{1}\{T_{b0} = T_{0b}\} + \mathbb{1}\{T_{b0} > T_{0b}\}$.
  **end for**
  Construct p-value for $H_0 : \beta = 0$ as $p_{\text{val}} \leftarrow \frac{1+\sum_{b=1}^{B} \omega_b}{B+1}$.
  **return** $p_{\text{val}}$
**end procedure**

- Permuted test statistic $T_{\text{perm}}$: the F-statistic for significance of $x_\pi$ in the model $y \sim x_\pi + Z + Z_\pi$:

$$T_{\text{perm}} = \left(\left\|(I - H^{zz_\pi})y\right\|_2^2 - \left\|(I - H^{x_\pi zz_\pi})y\right\|_2^2\right) / \left\|(I - H^{x_\pi zz_\pi})y\right\|_2^2.$$

Here, $H^{(.)}$ denotes the projection matrix onto the column space of its argument (.). For instance, $H^{z_\pi z}$ represents the projection matrix onto the column space of $(Z_\pi, Z)$, $H^{xzz_\pi}$ onto that of $(x, Z, Z_\pi)$, etc. We have more confidence of having nonzero $\beta$ if $T_{\text{original}}$ is larger than $T_{\text{perm}}$. One can easily verify that comparing $T_{\text{original}}$ and $T_{\text{perm}}$ is equivalent to comparing the following simplified expressions on the fitted residuals:

$$T_{\text{original}} = \left\|(I - H^{x_\pi zz_\pi})y\right\|_2^2, \qquad T_{\text{perm}} = \left\|(I - H^{xzz_\pi})y\right\|_2^2. \tag{3}$$

We adopt eq. (3) to construct the original and permuted statistics for any given permutation. Subsequently, we generate $B$ random permutations $\{\pi_b\}_{b=1}^{B}$ of $[n]$ uniformly and compare the associated original and permuted statistics to compute the p-value for testing $H_0 : \beta = 0$. The complete procedure is outlined in Algorithm 1.

2.2. *An example differentiating FLtest and PALMRT.* We offer an illustrative example, see Example 2.1, to demonstrate PALMRT's superior robustness compared to FLtest. Although FLtest has been lauded for its type I error control in prior studies, it may fail under extreme noise and design configurations. In contrast, PALMRT consistently maintains type I error control.

EXAMPLE 2.1. We set $n = 100$ and $p = 1$, and examine a special design where $x = (1, 0, \ldots, 0)^\top$, $z = (0, 1, 0, \ldots, 0)^\top$. We generate the response $y$ under the global null as $y \sim \varepsilon$, where $\varepsilon \sim N(0, I_{n\times n}) + 10^4 \times \text{Multinomial}(1; \frac{1}{n}, \ldots, \frac{1}{n}) \times (-1)^{\text{Bernoulli}(\frac{1}{2})}$, denoted as "multinomial noise." This extreme noise scenario serves as a robustness test. The feature design represents an imbalanced ANOVA setup with many control samples but only one sample per treatment group. Even though exact permutation is feasible by shuffling the zero rows, we employ standard FLtest and PALMRT to evaluate their empirical coverage using 2000 random draws of $y$. Figure 1A shows the miscoverage ratio, defined as

$$\text{miscoverage ratio} = \frac{\text{Empirical miscoverage}}{\text{Targeted miscoverage}}.$$

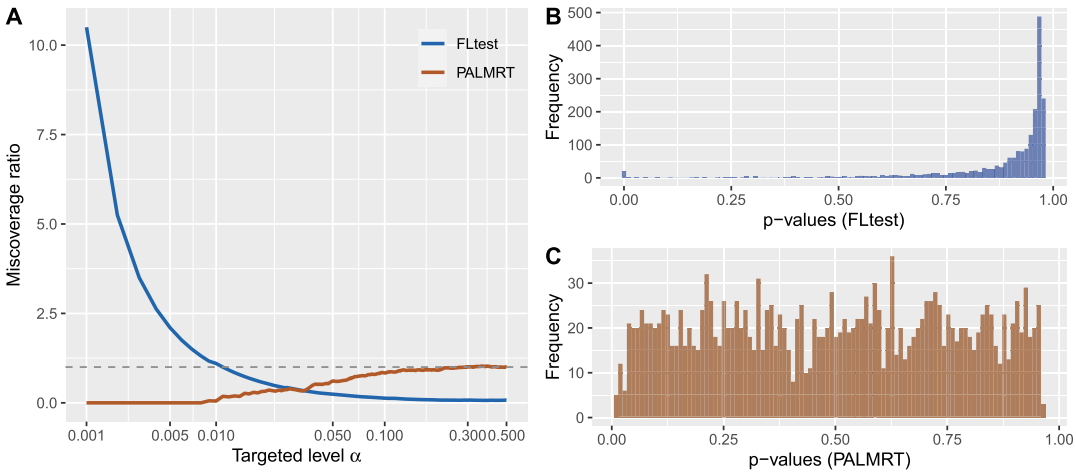


FIG. 1. *Comparison between FLtest and PALMRT on a distinguishing example. Panel A shows the miscoverage ratios using FLtest and PALMRT in Example 2.1 as we range α from* 0.001 *to* 0.5. *Panels B-C display the histograms of p-values using the two tests.*

As the targeted level $\alpha$ becomes small, FLtest encountered excessively inflated type I error – more than 10-fold that of the targeted level when $\alpha = 0.001$. In contrast, PALMRT controls type I error for small $\alpha$. Figures 1B-C are the histograms of p-values using FLtest and PALMRT. The distribution of p-values from FLtest contains a spike close to 0, leading to its type I error inflation.

Example 1 is a distinguishing example where FLtest suffers from being severely anticonservative for small $\alpha$ but PALMRT continues to offer robust type I error control. This robustness of PALMRT is universally true. In the next section, we will show that PALMRT, as well as a family of other permutation tests based on paired constructions, guarantees a maximum type I error of $2\alpha$, irrespective of the noise distribution, design, and sample size.

**3. Type I error control guarantee.** In this section, we establish that PALMRT provides finite-sample type I error guarantees under any design and exchangeable noise. This property is generalized to a family of tests that performs comparisons with paired statistics as realizations of a specially designed bi-variate function on permutation orders.

Under the null hypothesis $H_0$, PALMRT essentially compares a bi-variate function across two given permutations and their swaps. Specifically, for any two permutations $\pi_1$ and $\pi_2$ of $[n]$, we define a bi-variate function, which also incorporates data $x$, $Z$, and unobserved noise $\varepsilon$ as model parameters, as follows:

$$T^{\mathrm{PALMRT}}(\pi_1, \pi_2; x, Z, \varepsilon) = \left\| \left(I - H^{x_{\pi_2} z_{\pi_2} z_{\pi_1}}\right)\varepsilon \right\|_2^2. \tag{4}$$

Let $\pi_0$ be the identity permutation of $[n]$.

PROPOSITION 3.1. *Under the null hypothesis $H_0$, the statistic pair $(T_{0b}, T_{b0})$ are realizations of $T^{\mathrm{PALMRT}}(.,.; x, Z, \varepsilon)$ evaluated at $(\pi_0, \pi_b)$ and $(\pi_b, \pi_0)$, respectively*:

$$(T_{0b}, T_{b0}) = \left(T^{\mathrm{PALMRT}}(\pi_0, \pi_b), T^{\mathrm{PALMRT}}(\pi_b, \pi_0)\right).$$

Many existing random permutation tests, including FLtest and PERMtest, can be expressed as realizations of such bi-variate functions. For example, construct bi-variate func-

tions $T^{\mathrm{PERM}}(.)$ and $T^{\mathrm{FL}}(.)$ defined below,

$$T^{\mathrm{PERM}}(\pi_1,\pi_2;x,Z,\varepsilon)=\frac{\|(I-H^{z})\varepsilon\|_2^2-\|(I-H^{x_{\pi_1}z})\varepsilon\|_2^2}{\|(I-H^{x_{\pi_1}z})\varepsilon\|_2^2/(n-p-2)},$$

$$T^{\mathrm{FL}}(\pi_1,\pi_2;x,Z,\varepsilon)=\frac{\|(I-H^{z_{\pi_1}})(I-H^{z})\varepsilon\|_2^2-\|(I-H^{x_{\pi_1}z_{\pi_1}})(I-H^{z})\varepsilon\|_2^2}{\|(I-H^{x_{\pi_1}z_{\pi_1}})(I-H^{z})\varepsilon\|_2^2/(n-p-2)}.$$

Let $T_0$ denote the original statistic and $\{T_b\}_{b=1}^{B}$ represent $B$ permuted test statistics. Under $H_0$, it can be verified via direct calculation that $(T_0,T_b)$ are realizations at $(\pi_0,\pi_b)$ and $(\pi_b,\pi_0)$ of $T^{\mathrm{PERM}}(.,.;x,Z,\varepsilon)$ or $T^{\mathrm{FL}}(.,.;x,Z,\varepsilon)$ for PERMtest and FLtest respectively, with the second argument in the bivariate functions being inactive.

What sets $T^{\mathrm{PALMRT}}(\ldots)$ apart and enables its theoretical guarantee? The crucial distinction between $T^{\mathrm{PALMRT}}(\ldots)$ and $T^{\mathrm{FL}}(\ldots)$ or $T^{\mathrm{PERM}}(\ldots)$ lies in the transferability of permutations from the noise parameter $\varepsilon$ to its permutation arguments. This property holds uniquely for $T^{\mathrm{PALMRT}}(\ldots)$ across all noise realizations and designs. Let $\sigma$ be an arbitrary permutation of $[n]$ and $\sigma^{-1}$ its inverse, such that $\sigma\circ\sigma^{-1}=\sigma^{-1}\circ\sigma=\pi_0$ with $\circ$ denoting composition. Then, any permutation of the parameters $\varepsilon$ in $T^{\mathrm{PALMRT}}(\ldots)$ can be expressed equivalently as applying the inverse permutation $\sigma^{-1}$ to the permutation arguments $\pi_1$ and $\pi_2$.

PROPOSITION 3.2. *The application of a permutation $\sigma$ to $\varepsilon$ is equivalent to applying the permutation $\sigma^{-1}$ to $\pi_1,\pi_2$ in $T^{\mathrm{PALMRT}}$:*

$$T^{\mathrm{PALMRT}}(\pi_1,\pi_2;x,Z,\varepsilon_\sigma)=T^{\mathrm{PALMRT}}\big(\pi_1\circ\sigma^{-1},\pi_2\circ\sigma^{-1};x,Z,\varepsilon\big).$$

Proposition 3.2 is derived from simple term rearrangement, and we omit its proof here. This transferability property is pivotal for establishing the type I error guarantees. In fact, for any paired statistics $T_{0b}$, $T_{b0}$ which can be considered as realizations of a bi-variate function $T(.,.;x,Z,\varepsilon)$ at $(\pi_0,\pi_b)$ and $(\pi_b,\pi_0)$ under $H_0$, the resulting p-value from comparing $T_{0b}$ to $T_{b0}$ offers a theoretical guarantee as long as $T(.,.;x,Z,\varepsilon)$ satisfies the transferability Condition 4, as outlined in Theorem 4.1.

CONDITION 4. For any permutations $\pi_1,\pi_2,\sigma$ of $[n]$, the function $T(.,.;x,Z,\varepsilon)$ satisfies

$$T(\pi_1,\pi_2;x,Z,\varepsilon_\sigma)=T\big(\pi_1\circ\sigma^{-1},\pi_2\circ\sigma^{-1};x,Z,\varepsilon\big).$$

THEOREM 4.1. *Let $\pi_1,\ldots,\pi_B$ be $B$ uniformly random permutations of $[n]$, and $T(.,.;x,Z,\varepsilon)$ be a bi-variate function satisfying Condition 4. Under the null hypothesis $H_0$, construct paired statistics $(T_{0b},T_{b0})$ as*

$$T_{0b}=T(\pi_0,\pi_b;x,Z,\varepsilon),\qquad T_{b0}=T(\pi_b,\pi_0;x,Z,\varepsilon).$$

*Substituting these into Algorithm 1, we obtain $\mathbb{P}_{H_0}[p_{val}\le\alpha]<2\alpha$ for all $\alpha>0$, where the probability is marginalized over both noise and permutation randomness.*

REMARK 4.2. Interestingly, the empirical version of RPT, discussed independently in [34], is also a realization of paired constructions described in Theorem 4.1 when $(Z_0,Z_{\pi_b^{-1}})$ is full rank for different permutations $\pi_b$, by setting

$$T(\pi_1,\pi_2;x,Z,\varepsilon)=\big|x_{\pi_2^{-1}}\big(I-H^{z_{\pi_1^{-1}}z_{\pi_2^{-1}}}\big)\varepsilon\big|.$$

This leads to the comparison between $T_{0b} = |x(I - H^{zz_{\pi_b^{-1}}})y|$ and $T_{b0} = |x_{\pi_b^{-1}}(I - H^{zz_{\pi_b^{-1}}})y|$, which can be easily verified. In [34], the authors recognized the power issue with RPT and introduced this empirical version to enhance power. However, it lacks theoretical justification. Here, we demonstrate also that the empirical RPT has a strong theoretical foundation as a special case of Theorem 4.1 and is a version of PALMRT by replacing the pivotal statistics with residuals inner product, hence, the requirement on permutations by RPT is unnecessary in its context.

The worst-case bound established in Theorem 4.1 aligns with the bounds for prediction coverage established for multisplit conformal prediction methods including CV+, Jackknife+, and ensemble conformal predictions [3, 18, 21, 23, 31]. However, our focus diverges significantly as we concentrate on hypothesis testing for the true model parameter $\beta$ rather than an out-of-sample prediction. Traditional exchangeability arguments, applicable when predicting on new samples, are inadequate for assessing the significance of $\beta$. To tackle this, we employ new arguments exploiting the assumption of exchangeable noise. A proof sketch for Theorem 4.1 is provided below.

PROOF SKETCH OF THEOREM 4.1. One key insight is that, when considering both the randomness in the permutations and $\varepsilon$, the $(B+1) \times (B+1)$ matrix $T$, with its $(l,k)$th entry as $T(\pi_l, \pi_k; x, Z, \varepsilon)$, is distributionally equivalent to $\tilde{T}$ whose $(l,k)$th entry is $T(\tilde{\pi}_l, \tilde{\pi}_k; x, Z, \varepsilon)$ with $(\tilde{\pi}_l)_{l=0}^B$ independently and uniformly generated from the permutation space of $[n]$. This equivalence allows us to analyze the p-value from Algorithm 1 by examining corresponding entries in $\tilde{T}$.

Define $f(l, \tilde{T}) = 1 + \sum_{k \neq l}(\mathbb{1}\{\tilde{T}_{kl} > \tilde{T}_{lk}\} + \frac{1}{2}\mathbb{1}\{\tilde{T}_{kl} = \tilde{T}_{lk}\})$. Then, $f(0, \tilde{T})$ corresponds to the numerator when constructing $p_{\text{val}}$, the p-value in Algorithm 1, upon substituting $T$ with $\tilde{T}$. Since $\tilde{\pi}_l$ are i.i.d. generated, we expect $\{f(l, \tilde{T})\}_{l=0}^B$ to be exchangeable for different $l$. Thus, we can bound the probability of $p_{\text{val}} \leq \alpha$ by bounding the size of the index set $\{l : f(l, \tilde{T}) \leq \alpha(B+1)\}$, which can be obtained following similar arguments used in proving prediction interval's coverage for multisplit conformal prediction. □

REMARK 4.3. The analogy between PALMRT and Jackknife$^+$ after constructing the suitable comparison matrix $T$ naturally leads to the question of whether we can draw a similar analogy between RPT and Jackknife$^{\text{mm}}$ (the minmax version) proposed together with Jackknife$^+$ in Barber et al. [3]. The answer to this question is affirmative. While RPT requires that the permutations form a group in Wen, Wang and Wang [34], by combining our analysis framework and construction ideas used in Jackknife$^{\text{mm}}$, we can show that RPT is valid for any uniformly and independently generated $B$ permutations. We refer to the resulting algorithm as $\widetilde{\text{RPT}}$ to distinguish it from RPT. Although $\widetilde{\text{RPT}}$ does not alleviate the over-conservativeness due to considering the worst-case, it does lead to algorithmic simplicity as no special algorithm is needed for generating the permutations. Details of the connection between $\widetilde{\text{RPT}}$ and Jackknife$^{\text{mm}}$ are given in Appendix B of the Supplementary Material [17].

Theorem 4.1 is the main theoretical result of this work, and we include its full proof in Section 6. Combining Theorem 4.1 with Proposition 3.2, we concludes that Algorithm 1 theoretically controls type I error, albeit with a relaxed upper bound of $2\alpha$.

Of note, unlike existing random permutation tests where the use of pivotal statistics is often crucial, Theorem 4.1 generalizes beyond pivotal statistics, allowing for other nonpivotal paired constructions. For example, $T_{0b}$, $T_{b0}$ could be the absolute value of the regression co-

**Algorithm 2** Exact CI construction for PALMRT

**procedure** CI($y, x, Z, B, \alpha$) ▷ Confidence interval at coverage level $1-\alpha$.
  Calculate the $B \times 4$ matrix $(c_{b1}, c_{b2}, c_{b3}, c_{b4})_{b=1}^{B}$ as defined in Lemma 5.3.
  Calculate $\gamma$, $\{t_l\}_{l=1}^{M}$, $(m_l^s, m_l^u)_{l=1}^{M}$.
  **if** $\gamma < 0$ **then**
    $\mathrm{CI}_\alpha = (-\infty, \infty)$.
  **else**
    Set and record $f_{A_1}(t_1) = \frac{1}{2}(m_1^s - m_1^u)$.
    **for** $l = 2, \ldots, M$ **do**
      Calculate and record $f_{A_1}(t_{l-1}^+)$ and $f_{A_1}(t_l)$ as in eq. (7).
    **end for**
    **if** $\max(f_{A_1}(.)) \le \gamma$ **then**
      $\mathrm{CI}_\alpha = \varnothing$.
    **else**
      $\beta_{\min} = \min\{t_l : f_{A_1}(t_l) \vee f_{A_1}(t_l^+) > \gamma\}$.
      $\beta_{\max} = \max\{t_l : f_{A_1}(t_l) \vee f_{A_1}(t_{l-1}^+) > \gamma\}$.
      $\mathrm{CI}_\alpha = [\beta_{\min}, \beta_{\max}]$.
    **end if**
  **end if**
**end procedure**

efficients of $x$ and $x_\pi$ in the models $y \sim x + Z + Z_{\pi_b}$ and $y \sim x_\pi + Z + Z_{\pi_b}$ respectively. For directional tests, the test statistics may employ either the regression coefficients (for positive effects) or their negations (for negative effects).

**5. An exact confidence interval construction.** In conjunction with the PALMRT $p$-value, a confidence interval for $\beta$ can be constructed by inverting the test. Define $(T_{0b}(\beta), T_{b0}(\beta))$ as the test statistics from replacing $y$ by $(y - x\beta)$ in Algorithm 2 when constructing $(T_{0b}, T_{b0})$. Define $f(\beta)$ as $f(\beta) = \frac{1+\sum_{b=1}^{B}\omega_b(\beta)}{B+1}$, where $\omega_b(\beta) = \mathbb{1}\{T_{0b}(\beta) < T_{b0}(\beta)\} + \frac{1}{2}\mathbb{1}\{T_{0b}(\beta) = T_{b0}(\beta)\}$.

COROLLARY 5.1. *Set* $\mathrm{CI}_\alpha = [\beta_{\min}, \beta_{\max}]$ *where* $\beta_{\min} = \inf\{\beta : f(\beta) > \alpha\}$ *and* $\beta_{\max} = \sup\{\beta : f(\beta) > \alpha\}$. *Then, we have* $\min_\beta \mathbb{P}[\beta \in \mathrm{CI}_\alpha] > 1 - 2\alpha$, *for all* $\alpha > 0$.

REMARK 5.2. The set obtained through directly inverting $\{\beta : f(\beta) > \alpha\}$, is often an interval, but not always guaranteed to be so. By taking the infimum and supremum of this set, we obtain a confidence interval $\mathrm{CI}_\alpha$ with worst-case guarantee at least as strong as direct inversion.

The pertinent question remaining is the efficient computation of the confidence interval $\mathrm{CI}_\alpha$ as delineated in Corollary 5.1. Traditional methods for constructing the confidence interval in permutation tests typically rely on normal theory, Bootstrap [4, 5, 9, 11], or grid search [16]. Here, we provide an exact formulation of $\mathrm{CI}_\alpha$ via examining critical values, derived from pair-wise comparisons at each permutation $\pi_b$. First, we observe that the contribution from the $b^{\text{th}}$ term to the numerator can be explicitly derived.

LEMMA 5.3. *Set* $c_{b1} = \|(I - H^{x_{\pi_b} z_{\pi_b} z})x\|_2^2$, $c_{b2} = x^\top (I - H^{x_{\pi_b} z_{\pi_b} z})y$, $c_{b3} = \|(I - H^{x_{\pi_b} z_{\pi_b} z})y\|_2^2$, *and* $c_{b4} = \|(I - H^{x_\pi z_{\pi_b} z})y\|_2^2$. *Then, we have* $c_{b2}^2 \geq c_{b1}(c_{b3} - c_{b4})$ *and*

$$\omega_b(\beta) = \begin{cases} \frac{1}{2}\mathbb{1}\{c_{b3} = c_{b4}\} + \mathbb{1}\{c_{b3} < c_{b4}\}, & \text{if } c_{b1} = 0, \\ \frac{1}{2}\mathbb{1}\{\beta \in [s_b, u_b]\} + \mathbb{1}\{\beta \in (s_b, u_b)\}, & \text{if } c_{b1} > 0, \end{cases} \tag{5}$$

*where* $s_b = \frac{c_{b2} - \sqrt{c_{b2}^2 - c_{b1}(c_{b3} - c_{b4})}}{c_{b1}}$ *and* $u_b = \frac{c_{b2} + \sqrt{c_{b2}^2 - c_{b1}(c_{b3} - c_{b4})}}{c_{b1}}$.

As a result, we first partition of index set of $\{0, \ldots, B\}$ into three sets $A_1 = \{b : c_{b1} > 0\}$, $A_2 = \{b : c_{b1} = 0, c_{b3} < c_{b4}\}$, and $A_3 = \{b : c_{b1} = 0, c_{b3} = c_{b4}\}$. The value of $\omega_b(\beta)$ remains constant as we vary $\beta$ for $b \in A_2 \cup A_3$. Hence, the requirement of $f(\beta) > \alpha$ is equivalent to imposing a requirement on $f_{A_1}(\beta)$ that captures the contribution of $\omega_b(\beta) \in A_1$, as shown below:

$$f(\beta) > \alpha \quad \Leftrightarrow \quad f_{A_1}(\beta) = \sum_{b \in A_1} \omega_b(\beta) > (B+1)\alpha - 1 - |A_2| - \frac{1}{2}|A_3| := \gamma.$$

It can be shown that the function value of $f_{A_1}(\beta)$ is characterized by comparing $\beta$ to different $s_b$ and $u_b$ values for $b \in A_1$:

$$f_{A_1}(\beta) = \frac{1}{2}\#\{b : s_b \leq \beta\} + \frac{1}{2}\#\{b : s_b < \beta\} - \frac{1}{2}\#\{b : u_b \leq \beta\} - \frac{1}{2}\#\{b : u_b < \beta\}. \tag{6}$$

Let $t_1 < \cdots < t_M$ denote the ordered values of $M$ unique elements in $\bigcup_{b \in A_1}\{s_b, u_b\}$, and let $(m_l^s, m_l^u)$ represent the sizes of $\#\{b : s_b = t_l\}$ and $\#\{b : u_b = t_l\}$, respectively. As we increase $\beta$ in $f_{A_1}(\beta)$, the function value can only changes when we first hit $\{t_l\}_{l=1}^M$, or when $\beta$ slightly increases from these critical values. We represent the concept of increasing slightly from these critical values by $\{t_l^+\}_{l=1}^M$, where $t_l^+$ indicates being infinitesimally larger than $t_l$. Using these new quantities introduced, we can re-express $f_{A_1}(t_1) = \frac{1}{2}(m_1^s - m_1^u)$ and identify induction relations for the function values as we increase $\beta$ to surpass the critical values $t_l$, described as follows:

$$f_{A_1}(t_{l+1}) = f_{A_1}(t_l^+) + \frac{1}{2}(m_{l+1}^s - m_{l+1}^u), \qquad f_{A_1}(t_l^+) = f_{A_1}(t_l) + \frac{1}{2}(m_l^s - m_l^u). \tag{7}$$

We can utilize eq. (7) to efficiently calculate all $f_A(t_l)$ and $f_A(t_l^+)$ in $O(B)$ time, given $\{t_l\}_{l=1}^M$ and $(m_l^s, m_l^u)_{l=1}^M$. Acquisition of the $B \times 4$ matrix $(c_{bl})$, $\{t_l\}_{l=1}^M$ and $(m_l^s, m_l^u)_{l=1}^M$ can be done in $O(Bnp + B\log B)$ time if we record intermediate quantities from Algorithm 1. Consequently, efficient comparisons between $f_A(.)$ with $\gamma$ to determine $\beta_{\min}$ and $\beta_{\max}$ can be achieved. Algorithm 2 presents full details of this implementation, and can provide exact construction of $\mathrm{CI}_\alpha$ as stated in Theorem 5.4.

THEOREM 5.4. *The confidence interval constructed by Algorithm 2 corresponds to the* $\mathrm{CI}_\alpha$ *defined in Lemma 5.1, and guarantees a worst-case coverage of* $(1 - 2\alpha)$ *for a specified mis-coverage level* $\alpha$.

REMARK 5.5. The confidence interval $\mathrm{CI}_\alpha$ can potentially be an empty set ($\varnothing$). Although this does not invalidate our assertion, an empty set offers limited information in practical contexts and might not be desired. In such situations, we may pt to construct the confidence interval using a normal approximation in the regression $y \sim x + Z$ whenever Algorithm 2 produces an empty set as the output for $\mathrm{CI}_\alpha$.

Proofs of Lemma 5.3 and Theorem 5.4 are deferred to Appendix A of the Supplementary Material [17].

**6. Proof of Theorem 4.1.** Before conducting numerical experiments comparing PALMRT and existing methods, we provide the full proof to Theorem 4.1 in this section. We define $\mathcal{S}_n$ as the permutation space of $[n]$, $\mathcal{S}_B$ as the permutation space of $(0, \ldots, B)$, $\mathcal{E} = \{\varepsilon_1, \ldots, \varepsilon_n\}$ as the unordered value set of $(\varepsilon_1, \ldots, \varepsilon_n)$ (duplicates are allowed). Proposition 6.1 is useful for establishing our exchangeability statement. Proposition 6.2 is a minor modification of arguments for bounding "strange set" used in proving multisplitting conformal prediction.

PROPOSITION 6.1. *Let $\sigma$ and $(\pi_b)_{b=1}^B$ be generated independently and uniformly from $\mathcal{S}_n$. Then $\sigma^{-1}$ and $(\pi_b \circ \sigma^{-1})_{b=1}^B$ are also generated independently and uniformly from $\mathcal{S}_n$.*

PROOF. First, it is obvious that $\sigma^{-1}$ is uniformly generated from $\mathcal{S}_n$ since the map $\sigma \mapsto \sigma^{-1}$ is a bijection between $\mathcal{S}_n$ and itself. Hence, by definition, for any $(B+1)$ permutations $\tau_0, \tau_1, \ldots, \tau_B$ of $[n]$, we have

$$\begin{aligned}
&\mathbb{P}\big(\sigma^{-1} = \tau_0, \pi_1 \circ \sigma^{-1} = \tau_1, \ldots, \pi_B \circ \sigma^{-1} = \tau_B\big) \\
&\quad = \mathbb{P}\big(\sigma^{-1} = \tau_0, \pi_1 = \tau_1 \circ \tau_0, \ldots, \pi_B = \tau_B \circ \tau_0\big) \\
&\quad = \mathbb{P}\big(\sigma^{-1} = \tau_0\big)\mathbb{P}(\pi_1 = \tau_1 \circ \tau_0) \ldots \mathbb{P}(\pi_B = \tau_B \circ \tau_0) \\
&\quad = (1/n!)^{B+1},
\end{aligned} \tag{8}$$

where the last two steps have used the fact that $\sigma^{-1}$ and $(\pi_b)_{b=1}^B$ are independent and uniformly from $\mathcal{S}_n$. Eq (8) is the definition for $\sigma^{-1}$ and $(\pi_b \circ \sigma^{-1})_{b=1}^B$ being independent and uniformly generated from $\mathcal{S}_n$. □

PROPOSITION 6.2. *Let $T$ be any $(B+1) \times (B+1)$ matrix, and $W$ the $(B+1) \times (B+1)$ comparison matrix where $W_{lk} = \mathbb{1}\{T_{kl} > T_{lk}\} + \frac{1}{2}\mathbb{1}\{T_{kl} = T_{lk}\}$ for $l \neq k$ and $W_{ll} = 1$. Let $W_{l\cdot}$ be the $l^{\text{th}}$ row sum of $W$. Then, for all $\alpha > 0$, we have*

$$\big|\{l : W_{l\cdot} \leq (B+1)\alpha\}\big| < 2\alpha(B+1).$$

PROOF. Notice that (1) $W_{lk} \geq 0$ for all $l, k$, and (2) $W_{lk} + W_{kl} = 1$ for all $k \neq l$. For any sub-square matrix of $W$ constructed from selecting the same $m$ columns and rows (denoted the index set as $I_m$), we have

$$\sum_{l \in I_m} \sum_{k \in I_m} W_{lk} = \frac{(m+1)m}{2}.$$

Set $S_\alpha = \{l : W_{l\cdot} \leq (B+1)\alpha\}$. Suppose the size of $S_\alpha$ is $s \geq 0$. The corresponding $s$ rows in $W$ must implicate:

$$s\alpha(B+1) \geq \frac{s(s+1)}{2} \quad \Rightarrow \quad 0 \leq s \leq \max\big(0, 2\alpha(B+1) - 1\big) < 2\alpha(B+1).$$

This concludes our proof. □

6.1. *Proof to Theorem 4.1.* On the one hand, under Assumption 1.1 and conditional on the value sets $\mathcal{E} = \{\varepsilon_1, \ldots, \varepsilon_n\}$, generation of $B$ random permutations $(\pi_1, \ldots, \pi_B)$ and new realization of noises $\varepsilon_\sigma$ can be characterized by generating $\sigma, \pi_1, \ldots, \pi_B, \sigma$ independently and uniformly generated from $\mathcal{S}_n$.

On the other hand, by Condition 4 and denoting $\pi_0$ as the identity permutation of $[n]$, we have

$$\begin{aligned}&\big(T(\pi_0,\pi_1;\varepsilon_\sigma),\dots,T(\pi_0,\pi_B;\varepsilon_\sigma),T(\pi_1,\pi_0;\varepsilon_\sigma),\dots,T(\pi_B,\pi_0;\varepsilon_\sigma)\big)\\ &\quad=\big(T\big(\sigma^{-1},\pi_1\circ\sigma^{-1};\varepsilon\big),\dots,T\big(\sigma^{-1},\pi_B\circ\sigma^{-1};\varepsilon\big),\\ &\qquad T\big(\pi_1\circ\sigma^{-1},\sigma^{-1};\varepsilon\big),\dots,T\big(\pi_B\circ\sigma^{-1},\sigma^{-1};\varepsilon\big)\big).\end{aligned}\tag{9}$$

Here, we have dropped the parameters $x$ and $Z$ in $T(.,.;x,Z,\varepsilon)$ for convenience.

Write $\tilde\pi_0=\sigma^{-1},\tilde\pi_1=\pi_1\circ\sigma^{-1},\dots,\tilde\pi_B=\pi_B\circ\sigma^{-1}$, by eq. (9) and Proposition 6.1, the marginalized joint distribution of the B pairs of statistics conditional only on $\mathcal{E}$ can be re-expressed equivalently (in distribution) using $(\tilde\pi_b)_{b=1}^B$:

$$(T_{0b},T_{b0})_{b=1}^B\overset{d}{=}\big(T(\tilde\pi_0,\tilde\pi_b;\varepsilon),T(\tilde\pi_b,\tilde\pi_0;\varepsilon)\big)_{b=1}^B,\tag{10}$$

where $\tilde\pi_0,\dots,\tilde\pi_B$ are independently and uniformly generated from $\mathcal{S}_n$, and $\varepsilon$ can be viewed as fixed. Hence, setting $(\tilde T_{0b},\tilde T_{b0})=(T(\tilde\pi_0,\tilde\pi_b;\varepsilon),T(\tilde\pi_b,\tilde\pi_0;\varepsilon))$, to understand the behavior of the constructed p-value using $(T_{0b},T_{b0})_{b=1}^B$, we can equivalently consider the distribution of $\widetilde{\mathrm{p_{val}}}$:

$$\widetilde{\mathrm{p_{val}}}=\frac{1+\sum_{b=1}^B(\mathbb{1}\{\tilde T_{b0}>\tilde T_{0b}\}+\frac12\mathbb{1}\{\tilde T_{b0}=\tilde T_{0b}\})}{B+1}.$$

Now, we complete the full $\tilde T$ matrix by setting $\tilde T_{kl}=T(\tilde\pi_k,\tilde\pi_l;\varepsilon)$ for all $k,l=0,\dots,B$. We also set the comparison $W$ matrix of $\tilde T$ as described in Proposition 6.2 and set $S_\alpha=\{l:W_{l\cdot}\le\alpha(B+1)\}$. Then, by Proposition 6.2, the size of $S_\alpha$ is upper bounded and $|S_\alpha|<2\alpha(B+1)$.

Note that the p-value constructed using $\tilde T$ is

$$\widetilde{\mathrm{p_{val}}}=\frac{1+\sum_{b=1}^B(\mathbb{1}\{\tilde T_{b0}>\tilde T_{0b}\}+\frac12\mathbb{1}\{\tilde T_{b0}=\tilde T_{0b}\})}{B+1}=\frac{W_{0\cdot}}{B+1}.$$

To avoid confusion, we use the lower case $(w_b)_{b=1}^B$ to denote the observed row sums $(W_{b\cdot})_{b=1}^B$ given the current realizations $(\tilde\pi_0,\dots,\tilde\pi_B)=(\tau_0,\dots,\tau_B)$, and $(W_{b\cdot})_{b=1}^B$ be the random variables as we change the permutations $(\tilde\pi_0,\dots,\tilde\pi_B)$. Then, conditional on the unordered set of permutations $\{\tilde\pi_0,\dots,\tilde\pi_B\}=\{\tau_0,\dots,\tau_B\}$, the observed permutations take the form $\tilde\pi_b=\tau_{\zeta_b}$, for $b=0,\dots,B$, where $\zeta$ is a permutation of $(0,\dots,B)$. Since $\tilde\pi_0,\dots,\tilde\pi_B$ are i.i.d generated from $\mathcal{S}_n$, $\zeta$ is uniformly generated from the $\mathcal{S}_B$.

The above results can be used to derive the exchangeability among $W_0,\dots,W_B$ conditional on $\mathcal{E}$ and $\{\tilde\pi_0,\dots,\tilde\pi_B\}=\{\tau_0,\dots,\tau_B\}$. Notice that when $\tilde\pi=\tau_\zeta$ is permuted according to $\zeta$, the row sum $W_{0\cdot}=w_{\zeta_0}$ is permuted accordingly:

$$\begin{aligned}W_0&=1+\sum_{b=1}^B\Big(\mathbb{1}\big\{T(\tau_{\zeta_b},\tau_{\zeta_0};\varepsilon)>T(\tau_{\zeta_0},\tau_{\zeta_b};\varepsilon)\big\}+\frac12\mathbb{1}\big\{T(\tau_{\zeta_b},\tau_{\zeta_0};\varepsilon)=T(\tau_{\zeta_0},\tau_{\zeta_b};\varepsilon)\big\}\Big)\\ &=1+\sum_{b=1}^B\Big(\mathbb{1}\{T_{\zeta_b,\zeta_0}>T_{\zeta_0,\zeta_b}\}+\frac12\mathbb{1}\{T_{\zeta_b,\zeta_0}=T_{\zeta_0,\zeta_b}\}\Big)\\ &=1+\sum_{b\ne\zeta_0}\Big(\mathbb{1}\{T_{b,\zeta_0}>T_{\zeta_0,b}\}+\frac12\mathbb{1}\{T_{b,\zeta_0}=T_{\zeta_0,b}\}\Big)\\ &=w_{\zeta_0}.\end{aligned}$$

Combining the above display with the fact that $\zeta$ is uniform from $\mathcal{S}_B$, we conclude that $\{W_{b\cdot}\}_{b=0}^B$ are exchangeable. Consequently, we have

$$\mathbb{P}(W_0\in S_\alpha)<2\alpha\quad\Rightarrow\quad\mathbb{P}(\widetilde{\mathrm{p_{val}}}\le\alpha)<2\alpha.$$

**7. Numerical experiments.** We evaluate the performance and type I error control of various methods through numerical experiments, setting the sample size at $n = 100$, and varying the dimension of $Z$ in $\{1, 5, 15\}$. We investigate four designs: i.i.d. Gaussian, i.i.d. Cauchy, balanced ANOVA where each feature takes roughly an equal number of nonoverlapping 1's, and a paired design where each feature assumes a value of 1 at two entries and 0 elsewhere, with one unique and one shared one-valued entry across all features. We also consider three noise settings: Gaussian, Cauchy, and multinomial. The paired design and multinomial noise distributions are atypical in real-world applications and are included as challenging test cases for evaluating the robustness of FLtest, which has demonstrated commendable empirical performance under more conventional design structures and noise distributions in existing literature.

We compare our proposed PALMRT against six existing methods:(1) F-test, (2) PERMtest, (3) FLtest, (4) CPT with strong pre-ordering by the genetic algorithm as per Lei and Bickel [25], (5) RPT, and (6) Bias-corrected and Accelerated Bootstrap, previously favored over plain Bootstrap [5, 6, 9, 19]. All comparisons were conducted at the p-value cut-off $\alpha = 0.05$ in the main paper; PALMRT's type I error control is further explored for $\alpha = 0.01$ and $\alpha = 0.001$ in Appendix C of the Supplementary Material [17].

For random permutation tests (PALMRT, FLtest, PERMtest), we employed F-statistics as the test statistics and set the permutation count $B = 2000$. Bias-corrected and Accelerated Bootstrap and CPT can be more time-intensive. We used *bcaboot* R package with 500 bootstraps and 20 jackknife blocks [10]. For CPT, we used the implementation from the authors' GitHub, followed the strong ordering approach with 10,000 optimization steps via genetic algorithms, and set the number of cyclic permutations as 19 (corresponding to $\alpha = 0.05$) [25]. For RPT, we used the implementation provided by the authors and set the the number of permutation as 19 (corresponding to $\alpha = 0.05$) as recommended [34].

We generated 2,000 independent noise instances $\varepsilon$ for each experimental setting and used them to compute empirical type I error, statistical power, and confidence intervals across various signal-to-noise ratios. Sections 7.1 and 7.2 compare type I error and power among F-test, PERMtest, FLtest, CPT, RPT, and PALMRT. Section 7.3 examines CI coverage for $\beta$ and their median lengths from independent runs using normal theory, Bootstrap and Algorithm 2.

7.1. *Type I error control.* To empirically assess type I error control, we simulate the global null distribution with $y = \varepsilon$. Figure 2 displays type I errors from 50 independent repetitions for each setting using F-test, PERMtest, FLtest, CPT, RPT and PALMRT.

In the Gaussian noise or the Gaussian/Cauchy design contexts, all methods effectively control type I errors. However, for ANOVA or Paired designs, Ftest, FLtest, and PERMtest yield inflated type I errors when noise is non-Gaussian – especially pronounced in the Paired design with multinomial errors. This underscores that not only F-test, but also FLtest and PERMtest, lack distribution-free theoretical guarantees and can be anticonservative in finite-sample settings.

In our numerical experiments, only PALMRT, CPT and RPT guarantee worst-case coverage across all experimental settings. Among them, CPT always provides type I error coverage close to $\alpha$, even when a nonconstant $\eta$ does not exist in theory. This is because of CPT's nature, with the latter also due to the provided CPT implementation. For example, under the Gaussian and Cauchy designs, CPT, with high probability, cannot construct a nonconstant $\eta$ that satisfies the required $p \times m$ linear constraints exactly for $p = 15$ and $m = \frac{1}{\alpha} - 1 = 19$. Further examination indicates that the cyclic-invariant constraints of CPT were not perfectly satisfied in the implemented CPT in such settings, thereby permitting nonconstant $\eta$, which led to close-to-random ordering for the test/permuted statistics and trivial tests providing both type I error and power around $\alpha$ regardless of the signal size in our experiments. In contrast,

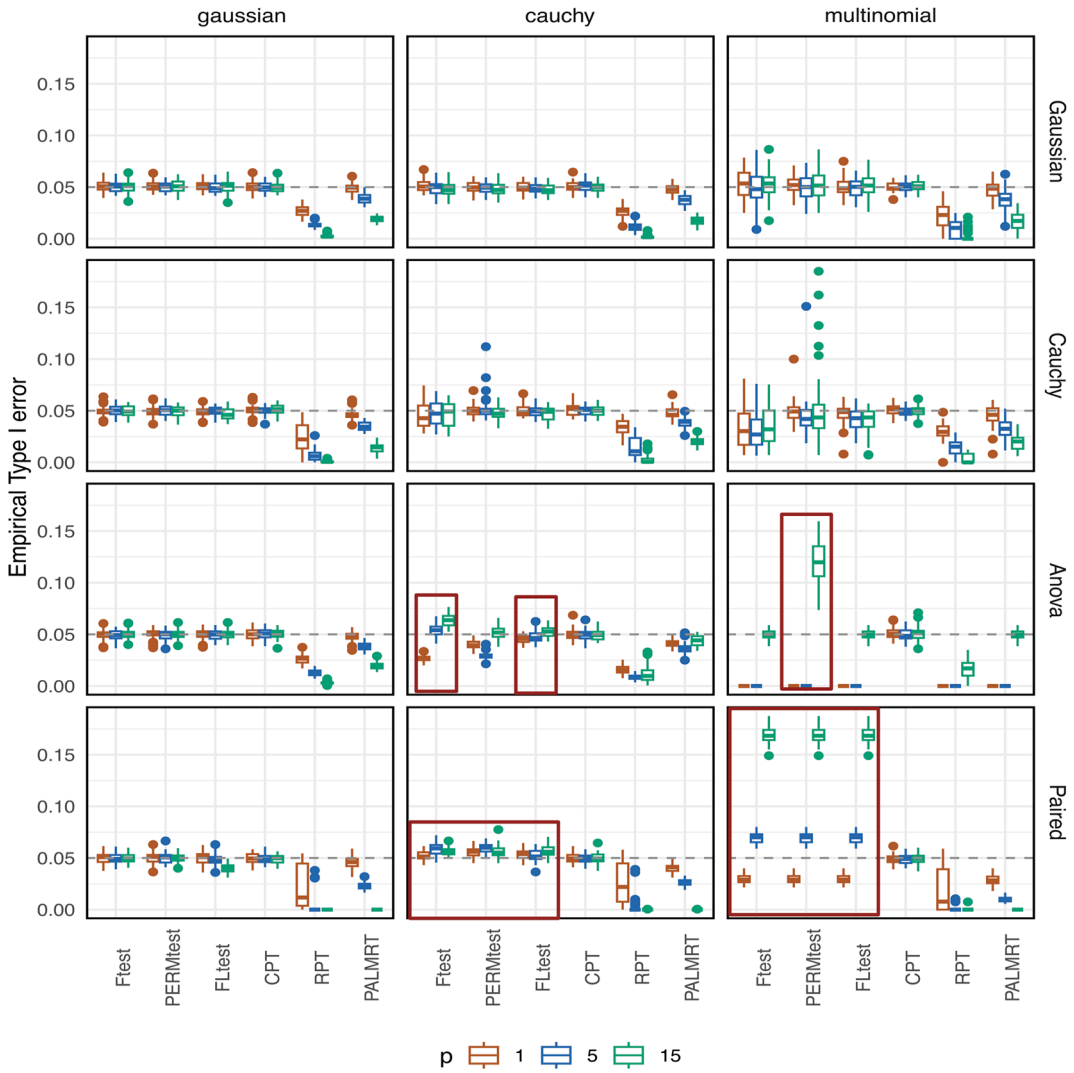


FIG. 2. *Empirical type I error using various methods, organized into boxplots by corresponding designs (row names) and noise distributions (column names). Each boxplot displays the empirical type I error for different methods, separately for different feature dimension p (color). The dashed horizontal line represents the targeted type I error* $\alpha = 0.05$. *Methods under a particular design and noise distribution were circled in red if the empirical type I errors were noticeably higher than* $\alpha$ *for some* $p$.

PALMRT demonstrates empirical type I errors close to or below target levels, while RPT, as noted in its original paper, tends to be conservative, with the degree of conservativeness varying based on design, noise distribution, and dimensionality.

Importantly, the conservative behavior of PALMRT does not sacrifice statistical power when compared to CPT and is generally more powerful compared to RPT, as evidenced by our subsequent power analyses.

7.2. *Power analysis.* In each experiment, we simulate data from the alternative setting by varying the linear coefficient in front of $x$ in eq. (11):

$$y = x\beta + \varepsilon. \tag{11}$$

We set $\beta$ using Monte Carlo simulation to yield F-test powers of approximately 30%, 50%, 70%, 90%, 99%. Figures 3–5 display the median, as well as the 25% and 75% percentiles,

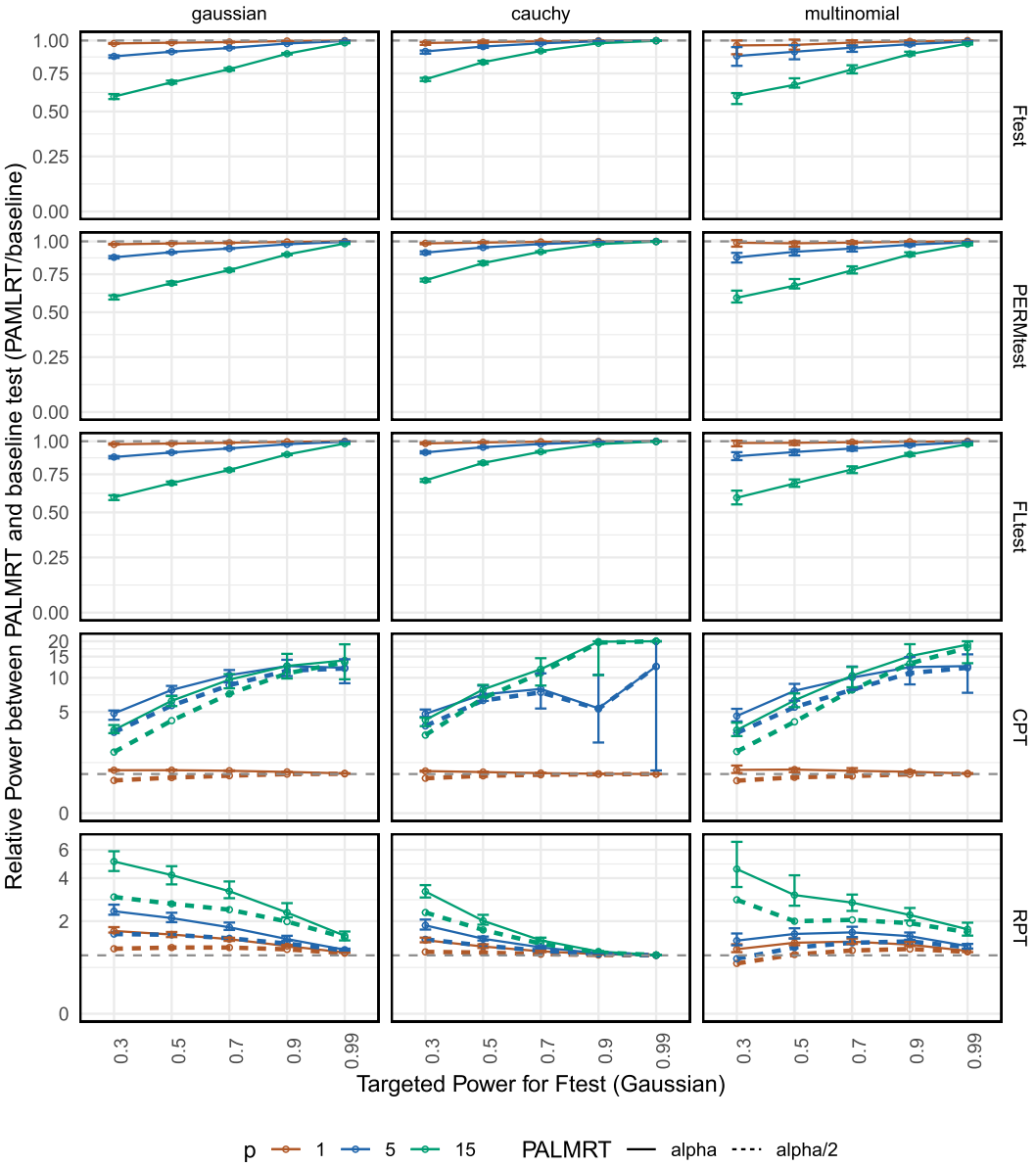


FIG. 3. *Power analysis for Gaussian design is presented, organized by baseline methods (row names) and noise distributions (column names), with different colors indicate varying feature dimensions $p$. Each open circle represents the median ratio between PALMRT and a baseline method, plotted against the targeted F-test absolute power for various signal sizes, with solid lines connecting the dots for overall trend visualization and associated error bars indicating the 25% and 75% quantiles of the ratio's empirical distributions. Additional comparisons between PALMRT$_{\frac{\alpha}{2}}$ and CPT/RPT are also shown in open circles and connected by dashed lines.*

of the relative power comparing PALMRT to F-test, PERMtest, FLtest, CPT and RPT for varying signal strengths, noise distributions and feature dimensions under Gaussian, Cauchy, and ANOVA designs. For visualization purposes, if a median ratio is greater than 20, it is truncated at 20.

Under the Gaussian design, Ftest, PERMtest, and FLtest are most powerful across various noise types and feature dimensions. PALMRT closely rivals them at $p = 1$ and its relative sensitivity to low signal strength decreases with increased feature dimensions. PALMRT achieves higher power than CPT at $p = 1$ and substantially outperforms CPT at $p = 5$ and

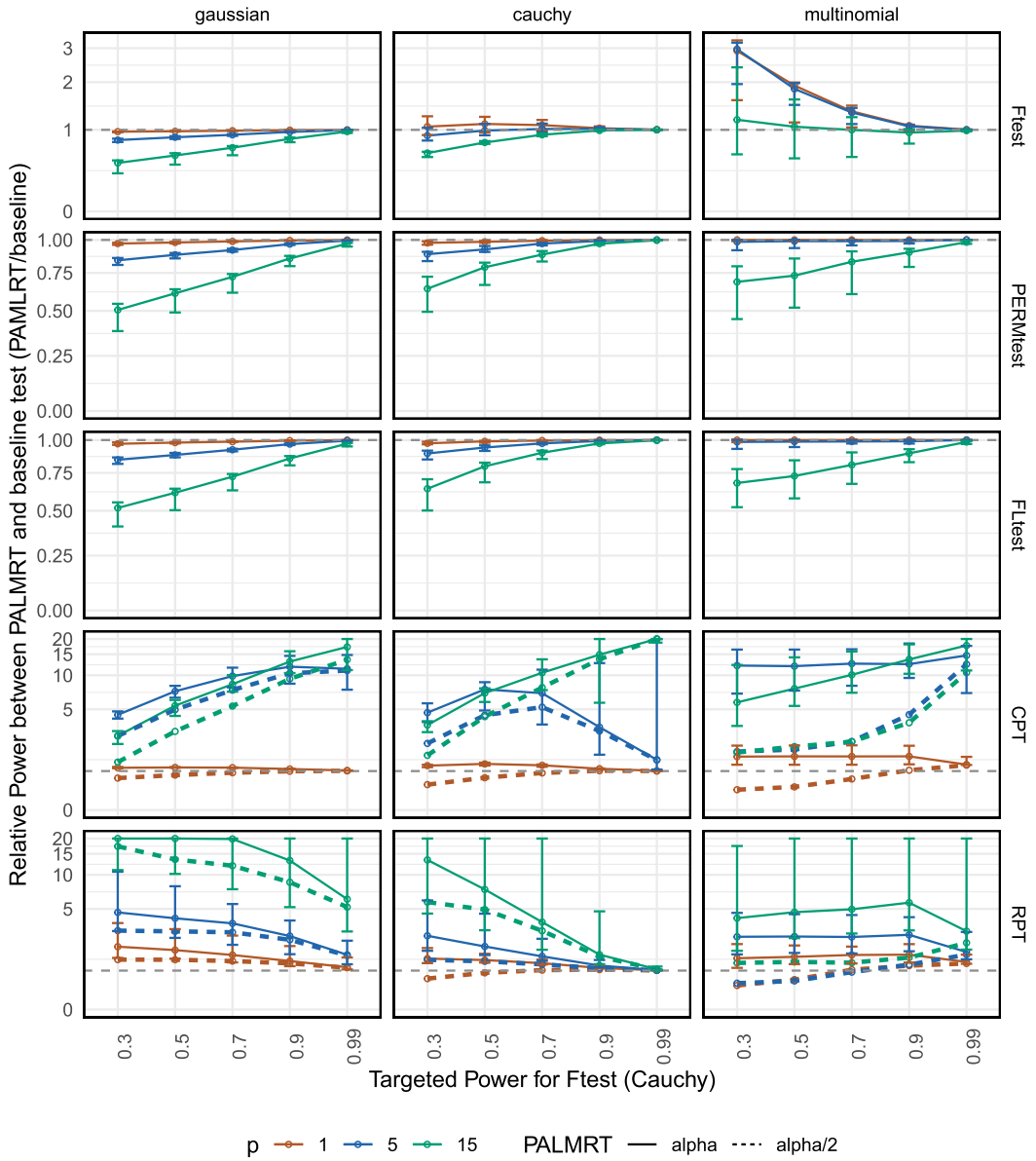


FIG. 4. *Power analysis for Cauchy design is presented, organized by baseline methods (row names) and noise distributions (column names), with different colors indicate varying feature dimensions $p$. Each open circle represents the median ratio between PALMRT and a baseline method, plotted against the targeted F-test absolute power for various signal sizes, with solid lines connecting the dots for overall trend visualization and associated error bars indicating the 25% and 75% quantiles of the ratio's empirical distributions. Additional comparisons between PALMRT$_{\frac{\alpha}{2}}$ and CPT/RPT are also shown in open circles and connected by dashed lines.*

$p = 15$, despite being more empirically conservative for controlling the type I error. The gap between CPT and PALMRT does not disappear for $p = 5$ and $p = 15$ as we increase the signal strength. This is very different from the group bound method [27] which is also conservative but is powerless compared to CPT in a wide range of signal-to-noise ratios when examined in Lei and Bickel [25]. PALMRT also shows much higher power compared to RPT with different concomitant feature dimension $p$, especially among the regime with a low signal-to-noise ratio.

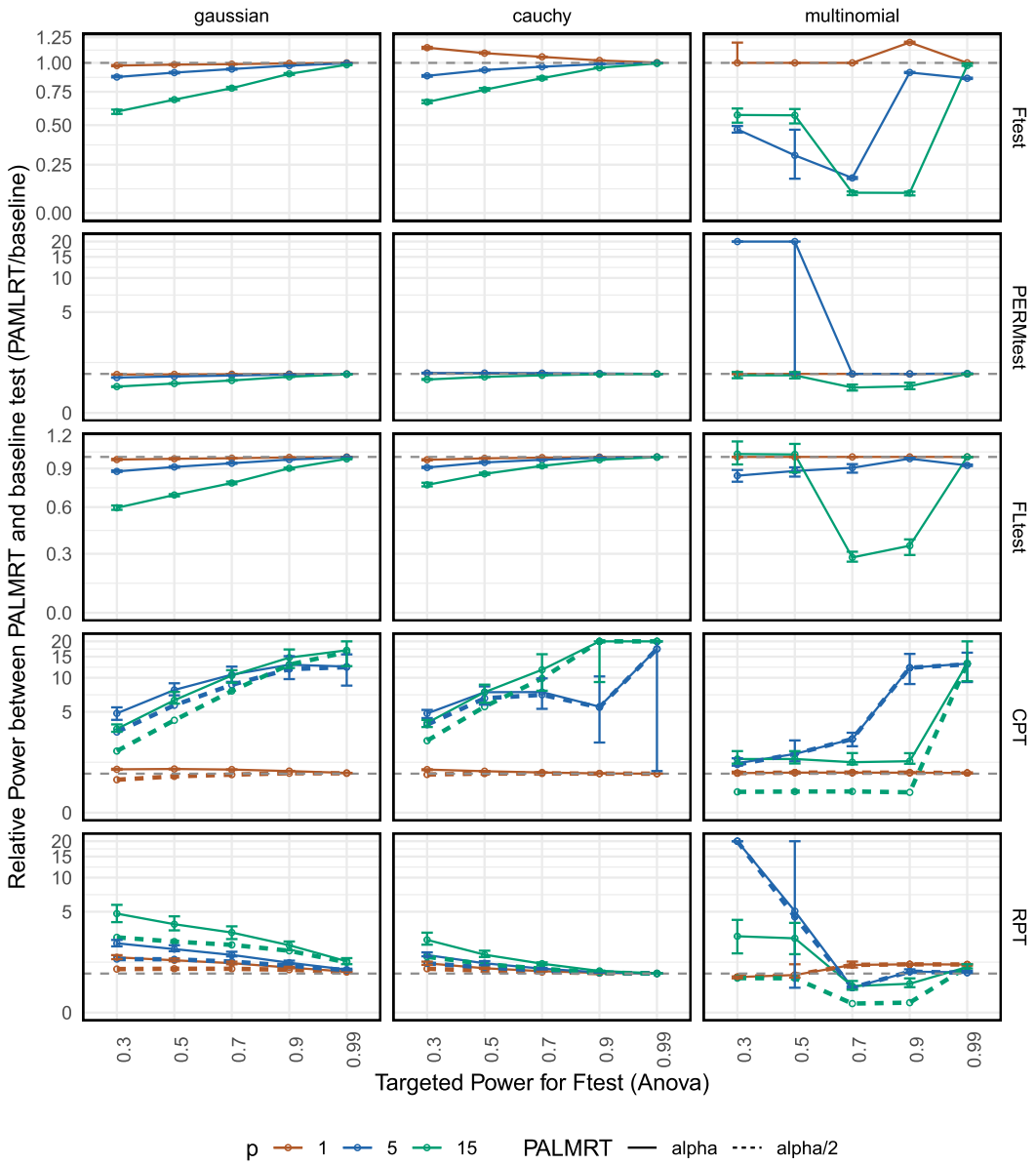


FIG. 5. *Power analysis for Anova design is presented, organized by baseline methods (row names) and noise distributions (column names), with different colors indicate varying feature dimensions p. Each open circle represents the median ratio between PALMRT and a baseline method, plotted against the targeted F-test absolute power for various signal sizes, with solid lines connecting the dots for overall trend visualization and associated error bars indicating the 25% and 75% quantiles of the ratio's empirical distributions. Additional comparisons between PALMRT$_{\frac{\alpha}{2}}$ and CPT/RPT are also shown in open circles and connected by dashed lines.*

In the Cauchy design, the relative sensitivities of PALMRT trend similarly to the Gaussian setting. However, F-test loses power with heavy-tailed noise distributions like Cauchy or multinomial, and random permutation tests can outperform F-test at low signal strengths. In the ANOVA design, results align with the Cauchy design when the noise is Gaussian or Cauchy. When the noise is multinomial, although the overall pattern is difficult to characterize, CPT and RPT are both significantly worse than PALMRT. Note that due to the extreme behavior of multinomial loss, when coupled with the ANOVA design, there is a huge discrepancy between the achieved power and the targeted power, even for FLtest, which remained

low for $p \in \{5, 15\}$ until the targeted F-test power reached 0.9 (see Appendix C of the Supplementary Material [17]).

Overall, when $(n/p)$ is large, PALMRT and FLtest perform similarly, but their relative diverges as $(n/p)$ decreases. PALMRT has consistently higher power than CPT in the Gaussian, Cauchy, and ANOVA designs, despite becoming more conservative as $p$ increases, and is also more powerful than RPT. Recall that PALMRT was more conservative than CPT and achieved lower empirical type I errors. The high power and low type I error of PALMRT relative to CPT might seem contradictory at first. However, CPT and PALMRT are very different procedures, and there is no guarantee that CPT has higher power than PALMRT when CPT's empirical type I error is worse. For instance, in cases where an exact nonconstant $\eta$ does not exist, CPT reduces to a trivial test while PALMRT and RPT still have nontrivial power. CPT also tended to have lower power in other settings when $p$ is moderately large compared to $n$. While rigorous reasoning for this is a separate question for further pursuit, we believe that this could be related to the restrictive space of test/permuted statistics of CPT, which takes the simple form $\eta_\pi^\top y$, linear with respect to permutation, whereas PALMRT and RPT allow nonlinear transformations on the permutation operations.

As CPT and RPT provide worst-case theoretical guarantees at $\alpha$ while PALMRT provides such a guarantee at $2\alpha$, we additionally include the comparison between PALMRT with a p-value cutoff at $\frac{\alpha}{2}$ (referred to as PALMRT$_{\frac{\alpha}{2}}$) and CPT/RPT using the median of their power ratios, despite the fact that PALMRT with an empirical p-value cutoff at $\alpha$ already achieves empirical type I error control. The median of the relative power of PALMRT$_{\frac{\alpha}{2}}$ to CPT/RPT is shown by dashed lines in the corresponding panels in Figures 3-5. Although the more stringent cutoff reduces its power, PALMRT$_{\frac{\alpha}{2}}$ still has comparable or higher power compared to CPT and RPT most of the time, especially for moderately sized dimensions of $Z$, for example, $p \in \{5, 15\}$.

7.3. *Coverage evaluation of* $\mathrm{CI}_\alpha$. In this section, we evaluate the empirical coverage and median length of confidence intervals ($\mathrm{CI}_\alpha$) constructed using Algorithm 2 ("Inversion"), Bootstrap, and normal approximation ("Normal") across Gaussian, Cauchy, and Anova designs for various $\beta$. We exclude the Paired design due to undefined Bootstrap CIs for all $p$. As CI coverage and length are consistent across different $\beta$, for the sake of space, we focus on results with a targeted F-test power of 50% and defer full results to Appendix C of the Supplementary Material [17].

Figure 6A presents the achieved coverage. The CIs from Inversion consistently meet the desired coverage. In contrast, the normal CIs exhibit slight under-coverage in the ANOVA design with Cauchy noise, and Bootstrap CIs show severe under-coverage for Cauchy noise. For the ANOVA design with $p = 15$, Bootstrap CIs are undefined and thus omitted in Figure 6A. Figure 6B shows the median CI lengths for each method. In line with previous findings, CIs from Algorithm 2 are generally wider than Normal CIs, except for $p = 1$ with Cauchy noise. Bootstrap CIs can exhibit greater variability when the designs are dominated by extreme values, such as in the Cauchy design setting.

**8. Robust identification of long-Covid biomarkers.** The MY-LC dataset contains measurements of 64 cell frequencies for 101 long-Covid (LC) participants and 84 controls (42 healthy samples, 42 convalescent samples without LC). A small percentage of measures are missing, with the number of observed samples ranging from 169 to 177 across features. Significant partial correlations between LC status and cell frequencies were identified by Klein et al. [24], after controlling for age, sex, and BMI as described in eq. (2). Among the 64 features, 26 had $p \leq 0.05$, and 5 survived multiple hypothesis correction using the BH procedure. In this work, we apply PALMRT, CPT, and RPT, which are theoretically guaranteed and validated through simulations, for robust biomarker identification.

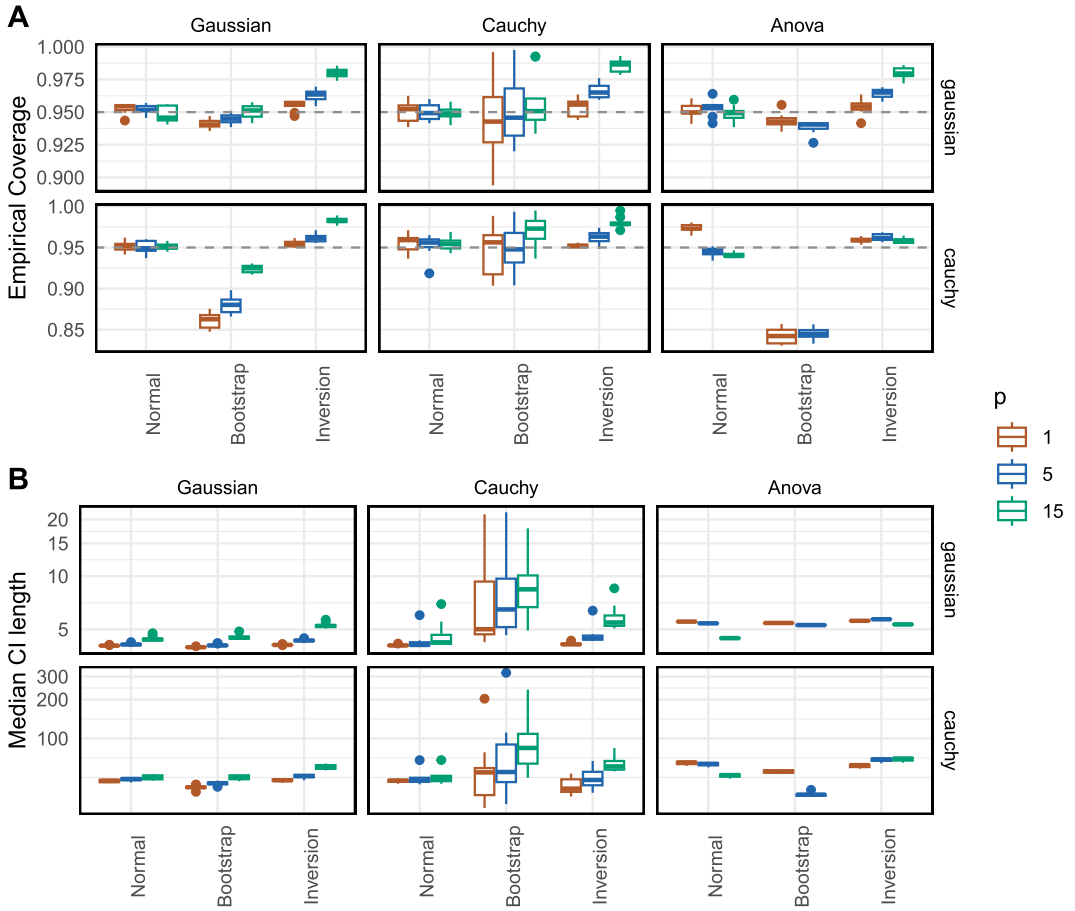


FIG. 6. *CI coverage and length comparisons, presented as boxplots, organized by designs (row names) and noise distributions (column names). Panel A displays the boxplots of coverage (y-axis) for different CI construction methods (x-axis). Panel B shows the boxplots of median CI length (y-axis) for different CI construction methods (x-axis). Both panels are colored by the feature dimension p.*

Figure 7A displays confidence intervals generated using Algorithm 2 for the 26 significant biomarkers identified by the t-test, before multiple corrections, at $\alpha = 0.05$. The center dot in each CI represents the estimated LC coefficient in eq. (2). Solid dots indicate features that were significant before correction; empty dots indicate otherwise. PALMRT confirmed 24 of these 26 biomarkers. In particular, all 5 top biomarkers, which were significant after correction using the t-test, remained significant after correction using PALMRT (solid dot with red circle). In general, the p-values from PALMRT and the t-test are highly concordant, with ratios between them for the same feature ranging from 0.6 to 1.2, except for cDC1 and the central memory CD4+ T cell with positive PD1 (CD4+ T cell (PD1+Tcm)), which achieved the smallest possible PALMRT p-values at $B = 10^5$ and also had smallest t-test p-values $\ll 10^{-5}$ (see Figure 7B).

Both CPT and RPT, in contrast, reduced discoveries before correction and yielded none afterward. Increasing the number of cyclic permutations ($m$) from $m = 19$ to $m = \lfloor \frac{n}{p} \rfloor - 1$ did not improve their power in CPT, as shown in Figure 7C. The nominal p-value cutoff at 0.05 resulted in 16 out of 26 t-test discoveries being confirmed when $m = 19$ and none being significant when $m = \lfloor \frac{n}{p} \rfloor - 1$ (see Figure 7C). RPT achieved higher power compared to CPT at the nominal p-value cutoff of 0.05, confirming 22 out of 26 t-test discoveries when $m = 19$. However, even for RPT, no discoveries were retained after correcting for multiple hypothesis

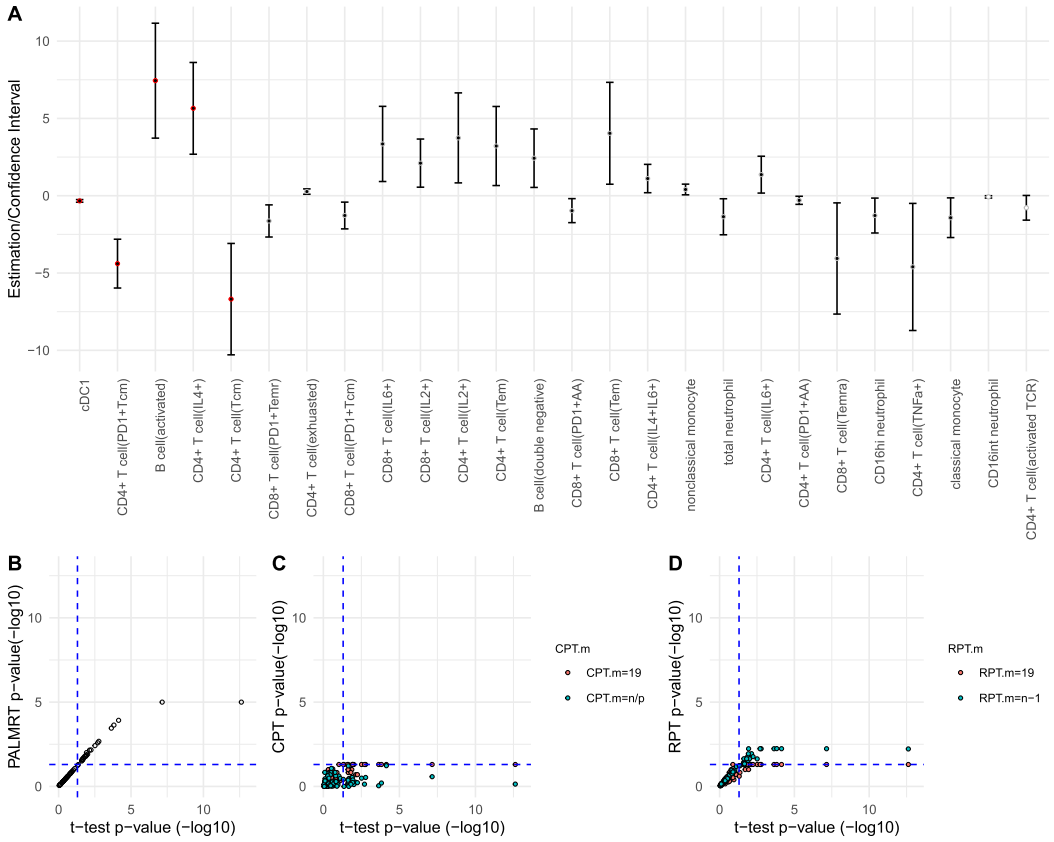


FIG. 7. *Biomarker discovery using PALMRT and CPT. Panel A displays the estimated coefficient* (*center dot*) *and constructed CI* (*segmented line*)*for PALMRT for the* 26 *significant t-test findings before correction, with the x-axis label showing the feature name* (*ordered based on significance*). *The center dot is solid if p-value* $\leq \alpha$ *and empty otherwise using PALMRT. If a feature is significant after correction using PALMRT, the center dot is further highlighted with red circle. Panel B shows the negative log*10 *of the PALMRT p-value against that from the t-test for* 64 *features. Panel C shows the negative log*10 *of the CPT p-value against that from the t-test for* 64 *features, where a dot is colored red if CPT used* $m = 19$ (*CPT.m* $= 19$) *and blue if CPT used* $m = \lfloor \frac{n}{p} \rfloor - 1$ (*CPT.m* $= n/p$). *Panel D shows the negative log*10 *of the RPT p-value against that from the t-test for* 64 *features, where a dot is colored red if RPT used* $m = 19$ (*RPT.m* $= 19$) *and blue if RPT used* $m = n - 1$ (*RPT.m* $= n - 1$). *The vertical/horizontal dashed lines represent the p-value level of* 0.05 *in Panels B-D.*

tests, whether using $m = 19$ or $m = n - 1$, the largest $m$ allowed by the RPT algorithm (see Figure 7D).

**9. Discussions.** We introduce a novel conformal test, PALMRT, designed for hypothesis testing in linear regression. This test and its corresponding confidence intervals are efficiently computed by evaluating statistic pairs, which are formed by augmenting the original regression problem with row-permuted versions of $(x, Z)$. PALMRT achieves little power loss compared to conventional tests like the F/t-test and FL-test, and differs from CPT and RPT which also enjoy a worst-case coverage guarantee. Unlike CPT, PALMRT eliminates the need for complex optimization to construct the test and consistently outperforms CPT in both simulated and real-data scenarios. In comparison to RPT, another recently proposed method, developed independently to address the challenges in CPT, PALMRT does not require specially designed permutations that form a group, and its performance remains stable and does not deteriorate as the number of permutations increases.

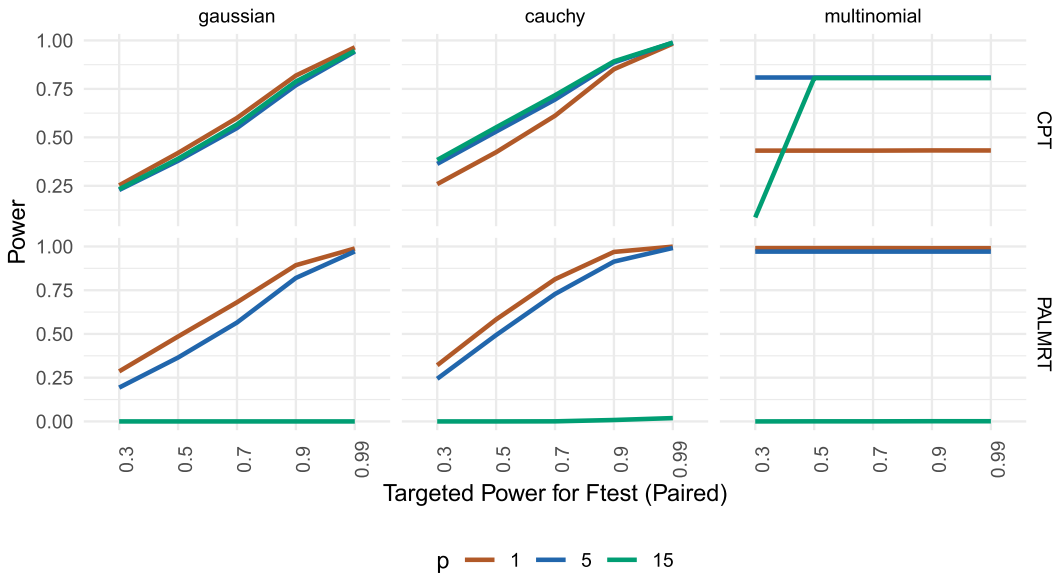


FIG. 8. *Power analysis in Anova design, presented as line plots, organized by methods (row names) and noise distributions (column names). Each line plot shows the median power of a given method and dimension $p$, plotted against the targeted F-test power for various signal sizes. Different colors indicate varying feature dimensions $p$.*

Of note, the improvement over CPT is not universal. For example, in special designs like the paired design where there are many duplicate rows in $Z$, CPT can be more powerful. Figure 8 illustrates the comparative power of CPT and PALMRT in signal detection under the paired design. While PALMRT generally matches or exceeds the power of CPT when $p = 1, 5$, it fails to detect signals at $\alpha = 0.05$ when $p = 15$. CPT, conversely, successfully identifies an effective pre-ordering and $\eta$ to construct a nontrivial test. While such settings are rare in practice, this observation raises an intriguing theoretical inquiry: Can the power of PALMRT be enhanced by filtering the random permutations $\{x_{\pi_b}, Z_{\pi_b}\}$ to avoid near co-linearity among $x$, $Z$, $x_{\pi_b}$, $Z_{\pi_b}$? We earmark this question for future exploration.

Another potential direction for future exploration is to extend the idea underlying PALMRT to settings beyond exchangeable noises, such as symmetric noises, by designing suitable bivariate functions compatible with the noise assumptions. We believe the PALMRT framework can be adapted to design and analyze test procedures in more general settings, as key conditions like Condition 4 and Proposition 6.1 are not specific to permutations.

As a brief detour from our main discussion, the augmentation step in PALMRT is reminiscent of the seminal work by Barber and Candès [2], which introduced the concept of knockoffs. This approach generates a knockoff copy $\tilde{X}$ of the original feature matrix $X$, for example, $X \leftarrow (x, Z)$ in our notation with dimension $n \times (p + 1)$. The key requirement of the knockoff copy is that swapping any pair $(\tilde{X}_j, X_j)$ leaves the covariance structure unchanged. Consequently, important quantities in regression analysis—such as OLS or Lasso coefficients $(\hat{\beta}_1, \ldots, \hat{\beta}_{p+1}, \tilde{\beta}_1, \ldots, \tilde{\beta}_{p+1})$—are invariant under these swaps, provided the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Gaussian. Knockoffs control the FDR by retaining features $X_j$ for which $|\hat{\beta}_j| - |\tilde{\beta}_j|$ is sufficiently large. Although both PALMRT and knockoffs operate under a fixed design and have similar sample size requirements, their objectives differ. Knockoffs aim to control FDR across $p + 1$ features under the Gaussian noise assumption, often in more complex model-fitting contexts, whereas PALMRT focuses on computing individual p-values for partial correlations under a more relaxed exchangeability condition for the residual errors. Recent advancements in derandomized knockoffs [29] allow for the computation of modified e-values [32, 33] through repeated runs with different $\tilde{X}$ copies. However,

achieving small p-values or high e-values necessitates a large $n >$ e-value or $n > (1/\text{p-value})$ at least, a constraint not shared by PALMRT. This makes PALMRT particularly advantageous for exploratory analyses aimed at uncovering partial correlations among a potentially large set of response and primary feature pairs, after adjusting for a limited number of covariates.

**Acknowledgments.** The author would like to thank Dr.Ĩwasaki and her team for providing access to the MY-LC data. We would also like to thank the reviewers and editors for their invaluable suggestions in improving this manuscript.

**Funding.** L.G. was supported in part by NSF Grant DMS-2310836.

## SUPPLEMENTARY MATERIAL

**Supplement to "A conformal test of linear models via permutation-augmented regressions"** (DOI: 10.1214/24-AOS2421SUPP; .pdf). In the Supplementary Material, we provide (1) omitted proofs for theoretical results in the main paper. (2) Additional results from numerical experiments.

## REFERENCES

[1] ANDERSON, M. J. and ROBINSON, J. (2001). Permutation tests for linear models. *Aust. N. Z. J. Stat.* **43** 75–88.
[2] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**.
[3] BARBER, R. F., CANDES, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021). Predictive inference with the jackknife+. *Ann. Statist.* **49**.
[4] DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application* **1**. Cambridge University Press, Cambridge.
[5] DICICCIO, T. J. and EFRON, B. (1996). Bootstrap confidence intervals. *Statist. Sci.* **11** 189–228.
[6] DICICCIO, T. J. and ROMANO, J. P. (1988). A review of bootstrap confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **50** 338–354.
[7] DRAPER, N. R. and STONEMAN, D. M. (1966). Testing for the inclusion of variables in einear regression by a randomisation technique. *Technometrics* **8** 695–699.
[8] EDGINGTON, E. and ONGHENA, P. (2007). *Randomization Tests*. CRC press, Boca Raton.
[9] EFRON, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82** 171–185.
[10] EFRON, B. and NARASIMHAN, B. (2020). The automatic construction of bootstrap confidence intervals. *J. Comput. Graph. Statist.* **29** 608–619. https://doi.org/10.1080/10618600.2020.1714633
[11] EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. CRC press, Boca Raton.
[12] FISHER, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *J. R. Stat. Soc.* **85** 597–612.
[13] FISHER, R. A. (1970). Statistical methods for research workers. In *Breakthroughs in Statistics*: *Methodology and Distribution* 66–70. Springer, Berlin.
[14] FISHER, R. A. et al. (1924). 036: On a distribution yielding the error functions of several well known statistics.
[15] FREEDMAN, D. and LANE, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econom. Statist.* **1** 292–298.
[16] GARTHWAITE, P. H. (1996). Confidence intervals from randomization tests. *Biometrics* 1387–1393.
[17] GUAN, L. (2024). Supplement to "A conformal test of linear models via permutation-augmented regressions." https://doi.org/10.1214/24-AOS2421SUPP
[18] GUPTA, C., KUCHIBHOTLA, A. K. and RAMDAS, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognit.* **127** 108496.
[19] HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.* 927–953.
[20] HALL, P. and WILSON, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* 757–762.
[21] HAN, Y., XU, M. and GUAN, L. (2023). Conformalized semi-supervised random forest for classification and abnormality detection. ArXiv preprint arXiv:2302.02237.
[22] KENNEDY, F. E. (1995). Randomization tests in econometrics. *J. Bus. Econom. Statist.* **13** 85–94.

[23] KIM, B., XU, C. and BARBER, R. (2020). Predictive inference is free with the jackknife+-after-bootstrap. *Adv. Neural Inf. Process. Syst.* **33** 4138–4149.
[24] KLEIN, J., WOOD, J., JAYCOX, J., DHODAPKAR, R. M., LU, P., GEHLHAUSEN, J. R., TABACHNIKOVA, A., GREENE, K., TABACOF, L. et al. (2023). Distinguishing features of long COVID identified through immune profiling. *Nature* 1–3.
[25] LEI, L. and BICKEL, P. J. (2021). An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika* **108** 397–412.
[26] MANLY, B. F. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology* **70**. CRC press, Boca Raton.
[27] 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.
[28] PITMAN, E. J. G. (1937). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Suppl. J. R. Stat. Soc.* **4** 225–232.
[29] REN, Z. and BARBER, R. F. (2022). Derandomized knockoffs: Leveraging e-values for false discovery rate control. ArXiv preprint arXiv:2205.15461.
[30] TER BRAAK, C. J. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and Related Techniques*: *Proceedings of an International Conference, Held in Trier, FRG, June* 4–8, 1990 79–85. Springer, Berlin.
[31] VOVK, V., NOURETDINOV, I., MANOKHIN, V. and GAMMERMAN, A. (2018). Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications* 37–51. PMLR.
[32] VOVK, V. and WANG, R. (2021). E-values: Calibration, combination and applications. *Ann. Statist.* **49** 1736–1754.
[33] WANG, R. and RAMDAS, A. (2022). False discovery rate control with e-values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 822–852.
[34] WEN, K., WANG, T. and WANG, Y. (2022). Residual permutation test for high-dimensional regression coefficient testing. ArXiv preprint arXiv:2211.16182.
[35] WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-Based Multiple Testing*: *Examples and Methods for P-Value Adjustment* **279**. Wiley, New York.
[36] WINKLER, A. M., RIDGWAY, G. R., WEBSTER, M. A., SMITH, S. M. and NICHOLS, T. E. (2014). Permutation inference for the general linear model. *NeuroImage* **92** 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060