# Residual permutation test for high-dimensional regression coefficient testing

Tengyao Wang

London School of Economics

Workshop on Robustness Meets Causality

Jul 2024

# Collaborators

Kaiyue Wen
Tsinghua University

Yuhao Wang
Tsinghua University

# Regression coefficient testing

- We revisit one of the oldest problem in statistics: coefficient testing in linear model:

$$Y = X\beta + Zb + \epsilon$$

with $X \in \mathbb{R}^{n\times p}$ and $Z \in \mathbb{R}^n$ having fixed design and $\epsilon$ random noise in $\mathbb{R}^n$.

- We want to test

$$H_0 : b = 0 \quad \text{versus} \quad H_1 : b \neq 0.$$

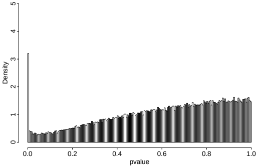
- **Goal**:
  - develop test with non-asymptotic valid size
  - understand difficulty of the problem in terms of the tail property of $\epsilon$.

# Classically ...

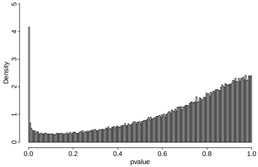
- Assuming $\epsilon$ has i.i.d. Gaussian entries, Fisher (1921) proposed the ANOVA procedure
$$\frac{\mathrm{RSS}_X - \mathrm{RSS}_{X,Z}}{\mathrm{RSS}_{X,Z}} \sim F_{1,n-p-1}, \quad \text{under } H_0.$$
- The Gaussian error assumption can be relaxed to rotationally invariant or symmetric around zero noise (Hartigan, 1970; Meinshausen, 2015).
- Asymptotically, when $n \to \infty$ and $p$ is fixed, the above test statistic is asymptotically $\chi^2_1$.

# Size validity of ANOVA

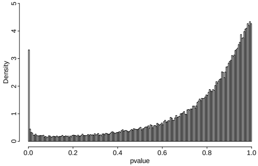
- ANOVA can have poor finite-sample size control (nominal size = 0.01)




$n = 300, p = 100$
Gaussian design
$t_1$ noise
empirical size 0.0181




$n = 300, p = 100$
$t_1$ design
$t_1$ noise
empirical size 0.0243




$n = 600, p = 200$
$t_1$ design
$t_1$ noise
empirical size 0.0141

- $p$-value distribution is far from uniform
- Large spike around 0, causing poor size control, especially for small nominal size.
- Important to develop a distribution-free and finite-sample valid test!

# Other tests in the literature

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- Permutation-based test can often achieve distribution-free size validity.
  - Freedman and Lane (1983) introduced a test based on permuting the regression residuals.
  - DiCiccio and Romano (2017) considered a permutation test using studentised partial correlations of $Y$ and $Z$ given $X$.
  - Toulis (2019) studied a test based on permuting residuals of $Y$ against $(Z, X)$.
- However, these tests only have asymptotic size controls.
- Cyclic permutation test of Lei and Bickel (2021) achieves finite-sample validity, assuming $n/p \geq 1/\alpha - 1$.

# A warm-up 1-bit test



- We assume only that $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ has **exchangeable** components.
- Given permutation matrices $P_1, \ldots, P_K \in \mathbb{R}^{n \times n}$
  - Let $\Pi_k \in \mathbb{R}^{n \times n}$ be the projection onto the orthogonal complement of the column span of $(X, P_k X)$ and write

$$\langle u, v \rangle_{\Pi_k} := u^\top \Pi_k v.$$

  - Under $H_0$, for a fixed $k \in \{1, \ldots, K\}$,

$$\langle Z, Y \rangle_{\Pi_k} = \langle Z, \epsilon \rangle_{\Pi_k} \stackrel{\mathrm{d}}{=} \langle Z, P_k \epsilon \rangle_{\Pi_k} = \langle Z, P_k Y \rangle_{\Pi_k}$$

  - Residuals of regression $Y$ and $P_k Y$ against $(X, P_k X)$ should have be equally likely to correlate with $Z$ under the null.
  - Each $P_k$ gives a 1-bit test of $H_0$.

# Combining the 1-bit tests into a level-$\alpha$ test

- Recall that under $H_0$, the 1-bit test compares the magnitude of

$$a_k := \langle Z, Y \rangle_{\Pi_k} = \langle Z, \epsilon \rangle_{\Pi_k} \quad \text{and} \quad b_k := \langle Z, P_k Y \rangle_{\Pi_k} = \langle Z, P_k \epsilon \rangle_{\Pi_k}.$$

- To combine 1-bit tests from projections $P_0 = I_n, P_1, \ldots, P_K$, define

$$a^* := \min_{\ell \in \{1,\ldots,K\}} a_\ell \quad \text{and} \quad b_k^* := \min_{\ell \in \{1,\ldots,K\}} \langle Z, P_k \epsilon \rangle_{\Pi_\ell}.$$

# Combining the 1-bit tests into a level-$\alpha$ test

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- Recall that under $H_0$, the 1-bit test compares the magnitude of

$$a_k := \langle Z, Y \rangle_{\Pi_k} = \langle Z, \epsilon \rangle_{\Pi_k} \quad \text{and} \quad b_k := \langle Z, P_k Y \rangle_{\Pi_k} = \langle Z, P_k \epsilon \rangle_{\Pi_k}.$$

- To combine 1-bit tests from projections $P_0 = I_n, P_1, \ldots, P_K$, define

$$a^* := \min_{\ell \in \{1,\ldots,K\}} a_\ell \quad \text{and} \quad b_k^* := \min_{\ell \in \{1,\ldots,K\}} \langle Z, P_k \epsilon \rangle_{\Pi_\ell}.$$

- If $P_0, \ldots, P_K$ form a group, then

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

is a valid (and almost exact) $p$-value at any size-$\alpha$.

# Combining the 1-bit tests into a level-$\alpha$ test

- Recall that under $H_0$, the 1-bit test compares the magnitude of

$$a_k := \langle Z, Y\rangle_{\Pi_k} = \langle Z, \epsilon\rangle_{\Pi_k} \quad \text{and} \quad b_k := \langle Z, P_k Y\rangle_{\Pi_k} = \langle Z, P_k \epsilon\rangle_{\Pi_k}.$$

- To combine 1-bit tests from projections $P_0 = I_n, P_1, \ldots, P_K$, define

$$a^* := \min_{\ell \in \{1,\ldots,K\}} a_\ell \quad \text{and} \quad b_k^* := \min_{\ell \in \{1,\ldots,K\}} \langle Z, P_k \epsilon\rangle_{\Pi_\ell}.$$

- If $P_0, \ldots, P_K$ form a group, then

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

  is a valid (and almost exact) $p$-value at any size-$\alpha$.
- Unfortunately, $\phi^*$ is not computable from data: $\langle Z, P_k \epsilon\rangle_{\Pi_\ell} \neq \langle Z, P_k Y\rangle_{\Pi_\ell}$.

# An admissible test with valid size control



- Instead of $\phi^*$, we use

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

which stochastically dominates $\phi^*$.

# An admissible test with valid size control

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- Instead of $\phi^*$, we use

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

which stochastically dominates $\phi^*$.

- Computational complexity: same as running $K$ OLS regressions, so $O(Kp^2n)$.

- In addition to using Euclidean inner products, we can also construct test using any function $T(\Pi_k Z, \Pi_k Y)$.

# Size control under the null

- $\phi$ has finite-sample size validity under weak assumptions.

**Theorem.** Assume $Y = X\beta + Zb + \epsilon$ with $\epsilon$ having exchangeable components and $p < n/2$. If $\{P_0, P_1, \ldots, P_K\}$ forms a group, then $\phi$ defined above satisfies

$$\mathbb{P}(\phi \leq \alpha) \leq \frac{\lfloor \alpha(K+1) \rfloor}{K+1} \leq \alpha,$$

for all $\alpha \in [0, 1]$.

# Sketch of proof for size control

- Since $\epsilon$ is exchangeable, it is invariant under group action of $\mathcal{P} = \{P_1, \ldots, P_K\}$.
- The set $\{\Pi_1, \ldots, \Pi_K\}$ is also invariant under action of the group $\mathcal{P}$.
- Hence the test statistics $a^*$, $b_1^*$,...,$b_K^*$ are invariant under group action of $\mathcal{P}$, in particular, rank of $a^*$ is uniformly distributed in $\{1, \ldots, K\}$.

# Power analysis: additional assumptions



- To analyse the power of the test, we need more assumptions on the design.

# Power analysis: additional assumptions

- To analyse the power of the test, we need more assumptions on the design.
- (Assumption A3) We assume that $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ have i.i.d. components distributed from a centred distribution $\mathbb{P}_\epsilon$, and that

$$Z = X\gamma + e$$

  with $e = (e_1, \ldots, e_n)^\top$ independent from $\epsilon$ with i.i.d. components distributed from a centred distribution $\mathbb{P}_e$.
- (Assumption A4) Assume additionally that the permutation matrices $P_1, \ldots, P_K$ satisfies $\mathrm{tr}(P_k) = 0$ and $|\mathrm{tr}(\Pi_0 P_k)| \leq \sqrt{2p}K$.

# Power analysis: discussion of assumptions

- Assumption (A4) is relatively mild. It can be shown that a group $P_0, P_1, \ldots, P_K$ satisfying (A4) always exists and we can find a random algorithm that has a success probability of $1 - 1/K$ in finding such a permutation group in each iteration.

- Assumption (A3) appears more stringent. The nodewise regression structural assumption of $Z$ is similar to the assumption in debiased Lasso.
- We can relax the linear structural assumption on $Z$ to allow for nonlinearity.

# Power analysis: signal strength upper bound

- We are interested in how the minimal testable signal strength $b$ is related to the tail heaviness of $e$ and $\epsilon$.

**Theorem.** Suppose $Y = X\beta + Zb + \epsilon$ where $\epsilon$ and $Z$ satisfies Assumption (A3) and

$$0 < \mathbb{E}|e_1|^2 < \infty \quad \text{and} \quad 0 < \mathbb{E}|\epsilon_1|^{1+t} < \infty$$

for some $t \in [0, 1]$. Assume $P_0, P_1, \ldots, P_K$ satisfies Assumption (A4). In the asymptotic regime where $b$ and $p$ vary with $n$ in such a way that $n > (3 + m)p$ for some constant $m > 0$ and

$$|b| \gtrsim n^{-t/(1+t)} \text{ if } t < 1 \quad \text{or} \quad |b| \gg n^{-1/2} \text{ if } t = 1,$$

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

# Sketch of proof for power analysis

- We need to show that

$$a_\ell = \langle Z, Y \rangle_{\Pi_\ell} = \langle e, be + \epsilon \rangle_{\Pi_\ell}$$

dominates

$$b_k = \langle Z, P_k Y \rangle_{\Pi_k} = \langle e, bP_k e + P_k \epsilon \rangle_{\Pi_k}$$

for all $\ell, k \in [K]$.

# Sketch of proof for power analysis

- We need to show that

$$a_\ell = \langle Z, Y \rangle_{\Pi_\ell} = \langle e, be + \epsilon \rangle_{\Pi_\ell}$$

  dominates

$$b_k = \langle Z, P_k Y \rangle_{\Pi_k} = \langle e, bP_k e + P_k \epsilon \rangle_{\Pi_k}$$

  for all $\ell, k \in [K]$.
- It suffices to show that for al $k \in [K]$

$$\begin{aligned} \langle e, \epsilon \rangle_{\Pi_k} &= o_p(bn) \\ \langle e, P_k \epsilon \rangle_{\Pi_k} &= o_p(bn) \\ \langle e, e \rangle_{\Pi_k} &= n - 2p + o_p(n) \\ \langle e, P_k e \rangle_{\Pi_k} &\leq p + \sqrt{2p}K + o_p(n) \end{aligned}$$

- The key step is to analyse the correlation of $e$ and $\epsilon$ on the projection space of $\Pi_k$.

# Power analysis: signal strength lower bound

- Let $\mathcal{D}_t$ be the class of distributions with $t$-th order moment bounded between $[1, 2]$.
- If $\mathbb{P}_e \in \mathcal{D}_2$ and $\mathbb{P}_\epsilon \in \mathcal{D}_{1+t}$, then a signal strength of $|b| \gtrsim n^{-t/(1+t)}$ is sufficient for RPT to be asymptotically powerful.
- The following result shows that this signal strength requirement is essentially optimal.

**Theorem.** Fix $t \in (0, 1]$. Suppose $Y = X\beta + Zb + \epsilon$ where $\epsilon$ and $Z$ satisfies Assumption (A3). For any $\eta \in (0, 1)$, there exists $c_\eta > 0$ depending only on $\eta$ such that for any fixed design $X$,

$$\inf_{\text{test } \varphi} \Bigg\{ \sup_{\substack{\mathbb{P}_\epsilon \in \mathcal{D}_{1+t} \\ \mathbb{P}_e \in \mathcal{D}_1 \\ \beta, \gamma \in \mathbb{R}^p}} \mathbb{P}_0(\varphi = 1) + \sup_{\substack{\mathbb{P}_\epsilon \in \mathcal{D}_{1+t} \\ \mathbb{P}_e \in \mathcal{D}_1 \\ \beta, \gamma \in \mathbb{R}^p}} \sup_{b \ge c_\eta n^{-t/(1+t)}} \mathbb{P}_b(\varphi = 0) \Bigg\} \ge 1 - \eta.$$

# Numerical simulations: size control

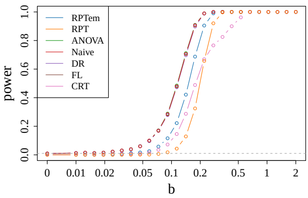
THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- Empirical size under the null for various design and noise distributions.
- We compare against DiCiccio and Romano (2017), Freedman and Lane (1983) and CRT of Candès et al. (2018).

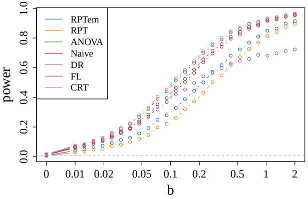
| | | | | $\text{RPT}_{\text{EM}}$ | | RPT | | DR | | FL | | CRT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $p$ | $\boldsymbol{X}$ | noise | 1% | 0.5% | 1% | 0.5% | 1% | 0.5% | 1% | 0.5% | 1% | 0.5% |
| 300 | 100 | $\mathcal{G}$ | $\mathcal{G}$ | 0 | 0 | 0 | 0 | 0.98 | 0.5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

# Numerical simulations: power curves

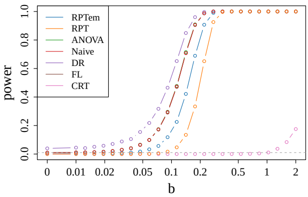
- Empirical power curves against signal size $b$ for various design and noise distributions.



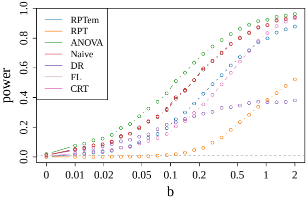

(a) Gaussian design, Gaussian noise




(b) Gaussian design, $t_1$ noise




(c) $t_1$ design, Gaussian noise




(d) $t_1$ design, $t_1$ noise

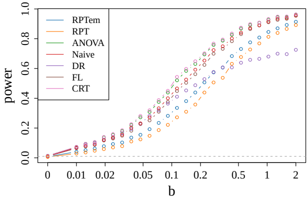
# Numerical simulations: misspecification



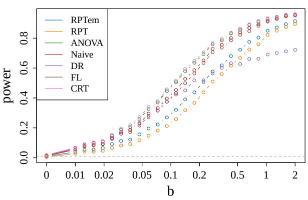
- Instead of a linear model of $Z$ on $X$ and $\epsilon \perp\!\!\!\perp e$, we allow
  - $Z$ to depend nonlinearly on $X$: $Z_i = f(X_i\gamma) + e_i$, where $f : t \mapsto 1/(1 + e^{-t})$ is the sigmoid function.
  - $e$ and $\epsilon$ to be dependent: $e$ has independent $t_1$ entries, and $\epsilon$ has either $t_1$ and $2t_1$ entries dependent on the sign of entries of $e$.

# Numerical simulations: misspecification

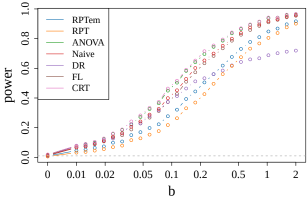




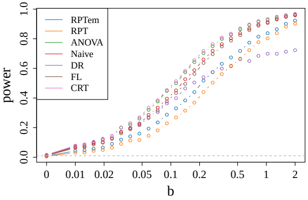

(a) independent noise, linear relation




(b) dependent noise, linear relation




(c) independent noise, nonlinear relation




(d) dependent noise, nonlinear relation

# Summary



- We propose a finite-sample valid permutation-based test for a single regression coefficient in a high-dimensional setting
- Key idea: compute projected correlation on the subspace orthogonal to both the original and permuted design matrix.
- Optimal power result showing minimal detectable signal $b$ in terms of tail-heaviness of the noise under suitable modelling assumption of design.
- R Package available on `github.com/wangtengyao/ResPerm`.

Main reference:


- Wen, K., Wang, T. and Wang, Y. (2022) Residual permutation test for high-dimensional regression coefficient testing. *Preprint*, arxiv:2211.16182.

**Thank you!**

# References




- Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. Roy. Statist. Soc., Ser. B*, **80**, 551—577.
- DiCiccio, C. J. and Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *J. Amer. Statist. Assoc.*, **112**, 1211–1220.
- Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.
- Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Statist.*, **1**, 292-–298.
- Hartigan, J. (1970). Exact confidence intervals in regression problems with independent symmetric errors. *Ann. Math. Statist.*, **41**, 1992–1998.
- Lei, L. and Bickel, P. J. (2021). An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, **108**, 397—412.
- Meinshausen, N. (2015). Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. Roy. Statist. Soc., Ser. B*, **77**, 923–945.
- Toulis, P. (2019). Invariant Inference via Residual Randomization. *arXiv preprint*, arXiv:1908.04218.