PERMUTATION INFERENCE WITH A FINITE NUMBER OF HETEROGENEOUS CLUSTERS

ANDREAS HAGEMANN

Abstract. I introduce a simple permutation procedure to test conventional (non-sharp) hypotheses about the effect of a binary treatment in the presence of a finite number of large, heterogeneous clusters when the treatment effect is identified by comparisons across clusters. The procedure asymptotically controls size by applying a level-adjusted permutation test to a suitable statistic. The adjustments needed for most empirically relevant situations are tabulated in the paper. The adjusted permutation test is easy to implement in practice and performs well at conventional levels of significance with at least four treated clusters and a similar number of control clusters. It is particularly robust to situations where some clusters are much more variable than others. Examples and empirical applications are provided.

JEL classification: C01, C22, C32

Keywords: cluster-robust inference, randomization, permutation, Behrens-Fisher

1. Introduction

It has become widespread practice in economics to conduct inference that is robust to within-cluster dependence. Typical examples of clusters are states, counties, cities, schools, firms, or stretches of time. Units within the same cluster are likely to influence one another or are influenced by the same external shocks. Several analytical and computationally intensive procedures such as the bootstrap are available to account for the presence of data clusters. Most of these procedures achieve consistency by requiring the number of clusters to go to infinity. Numerical evidence by Bertrand, Duflo, and Mullainathan (2004), MacKinnon and Webb (2017), and others suggests that this type of asymptotics often translates into heavily distorted inference in empirically relevant situations when the number of clusters is small or the clusters are heterogenous. In both situations, the overall finding is that true null hypotheses are rejected far too often. In this paper, I introduce an adjusted permutation procedure that is able to asymptotically control the size of tests about the effect of a binary

Date: April 6, 2021. I would like to thank co-editor Bryan Graham, three anonymous referees, Isaiah Andrews, Michal Kolesár, Aprajit Mahajan, and several seminar audiences for useful comments and discussions. All errors are my own.

treatment in the presence of finitely many large and heterogeneous clusters. The procedure applies to difference-in-differences estimation and other situations where treatment occurs in some but not all clusters and the treatment effect of interest is identified by between-cluster comparisons.

The main theoretical insight of this paper is that classical permutation inference can be adjusted to test the null hypothesis of equality of means of two finite samples of mutually independent but arbitrarily heterogeneous normal variables. This runs counter to classical permutation testing (Hoeffding, 1952), where the data under the null are presumed to be exchangeable. The adjustment corrects the significance level of the test downwards to account for heterogeneity. I prove that this is possible for empirically relevant levels of significance if both samples consist of more than three observations. The corrections needed for all standard levels of significance are tabulated in the paper. I also show that if a random vector of interest converges weakly to multivariate normal with diagonal covariance matrix, then permutation inference remains approximately valid for that vector. To exploit this result in a cluster context, I construct asymptotically normal statistics from each cluster and then apply adjusted permutation inference to the collection of these statistics. The resulting permutation test is consistent against all fixed alternatives to the null, powerful against local alternatives, and is free of user-chosen parameters.

The strategy of using cluster-level estimates as the basis for a test goes back at least to Fama and MacBeth (1973), who without formal justification run t tests on regression coefficients obtained from year-by-year cross-sectional regressions. Their approach is generalized and formalized by Ibragimov and Müller (2010, 2016), who construct t statistics from cluster-level estimates and show that for certain combinations of numbers of clusters and significance levels these statistics can be compared to Student t critical values. The Ibragimov-Müller test and the adjusted permutation test complement one another because they both rely on finite-sample inference with heterogeneous normal variables but apply to non-nested combinations of numbers of clusters and significance levels. The empirical example in this paper features a practically relevant situation where the Ibragimov-Müller test does not apply but the adjusted permutation test does. If both tests apply, the Monte Carlo results in this paper indicate that neither test dominates the other in terms of power but the adjusted permutation test has clear advantages if the underlying data are heavy tailed.

Several other papers show that inference with a fixed number of clusters is possible under a variety of conditions: Canay, Romano, and Shaikh (2017) permute the signs of cluster-level statistics under symmetry assumptions. This approach requires the parameter of interest to be identified *within* each cluster and clusters therefore have to be paired in an ad-hoc manner for difference-in-differences estimation. This pairing has a substantial impact on the test decision and requires a large number of choices on the part of the researcher. Bester, Conley, and Hansen (2011) use standard cluster-robust covariance matrix estimators but adjust critical values under homogeneity assumptions on the clusters. Canay, Santos, and Shaikh (2020) show that certain cluster-robust versions of the wild bootstrap can be valid under strong homogeneity assumptions with a fixed number of clusters. In sharp contrast, the test developed here does not require pairing clusters or any other decisions on the part of the researcher and applies even if the clusters are arbitrarily heterogeneous.

I will use the following notation: $1\{\cdot\}$ is the indicator function, $\min\{a, b\} = a \wedge b$, and cardinality of a set A is $|A|$. The smallest integer larger than a is $\lceil a \rceil$ and the largest integer smaller than a is $\lfloor a \rfloor$. Limits are as $n \to \infty$ unless noted otherwise.

All proofs can be found in the appendix.

2. Permutation inference with heterogenous symmetric variables

In this section I show that classical permutation inference can be adjusted to test for the equality of location of two finite samples of independent symmetric variables with heterogeneous scales. The discussion focuses on heterogeneous normal variables but several of the results apply more generally.

Suppose the random vector $X = (X_1, \dots, X_q) \in \mathbb{R}^q$ has entries $X_k = \mu_1 + \sigma_k Z_k$ for $1 \leqslant k \leqslant q_1$ and $X_k = \mu_0 + \sigma_k Z_k$ for $q_1 + 1 \leqslant k \leqslant q_1 + q_0 = q$, where the $Z_1, \dots, Z_q$ are iid symmetric variables. The σ_k are not known and no estimates are assumed to be available. The number of variables q is taken as fixed throughout this paper. The goal is to construct an α-level permutation test of the hypothesis $H_0\colon \mu_1 = \mu_0$. This is a two-sample problem with "treatment" sample $X_1, \dots, X_{q_1}$ and "control" sample $X_{q_1+1}, \dots, X_q$. The test statistic T considered here is the comparison of means

$$(x_1, \dots, x_q) \mapsto T(x) = \frac{1}{q_1} \sum_{k=1}^{q_1} x_k - \frac{1}{q_0} \sum_{k=q_1+1}^{q} x_k. \tag{2.1}$$

No standardization is needed.

Let $\mathfrak{S}_q$ be the group of permutations of the set $\{1, \dots, q\}$. For $g \in \mathfrak{S}_q$, denote by $g(k)$ the value the permutation g assigns to k for $1 \leqslant k \leqslant q$. The "group action" on X in $\mathfrak{S}_q$ is the relabeling of the indices $gX = (X_{g(1)}, \dots, X_{g(q)})$. A permutation test derives its critical values from the permutation statistics $T(gX)$. Because $x \mapsto T(x)$ is invariant to the ordering of the first q_1 and last q_0 entries of x, it suffices to compute the $T(gX)$ for the set of group actions with unique combinations of $g(1), \ldots, g(q_1)$ and $g(q_1+1), \ldots, g(q)$. One way of representing this set is

$$\mathfrak{G} = \big\{g \in \mathfrak{S}_q : g(1) < \cdots < g(q_1) \text{ and } g(q_1+1) < \cdots < g(q)\big\}. \tag{2.2}$$

Denote by $T^{(1)}(X, \mathfrak{G}) \leqslant T^{(2)}(X, \mathfrak{G}) \leqslant \cdots \leqslant T^{(|\mathfrak{G}|)}(X, \mathfrak{G})$ the ordered values of $T(gX)$ as g varies over $\mathfrak{G}$ and define critical values

$$p \mapsto T^p(X, \mathfrak{G}) = T^{(\lceil (1-p)|\mathfrak{G}|\rceil)}(X, \mathfrak{G}). \tag{2.3}$$

Classical permutation inference operates under the null hypothesis that X has the same distribution as gX for all $g \in \mathfrak{S}_q$. In the present context this would be equivalent to assuming that $\mu_1 = \mu_0$ and that all σ_k are identical under the null. An argument due to Hoeffding (1952) would then show that $T^\alpha(X, \mathfrak{G})$ could be used as the critical value for an α-level test against the alternative $H_1\colon \mu_1 > \mu_0$. If the null hypothesis is weakened to $H_0\colon \mu_1 = \mu_0$ without restrictions on σ_k, a natural question to ask if there exists *any* order statistic $j \mapsto T^{(j)}(X, \mathfrak{G})$, $\lceil (1-\alpha)|\mathfrak{G}|\rceil \leqslant j < |\mathfrak{G}|$, that can be used as a critical value for an α-level test even if the classical permutation hypothesis $X \sim gX$ for all $g \in \mathfrak{S}_q$ fails. As I will discuss now, the answer to this question is affirmative for empirically relevant choices of α if q_1 and q_0 are larger than 3.

Because $T(X) \in \{T(gX) : g \in \mathfrak{G}\}$, it is always true that $T(X) \leqslant T^{(|\mathfrak{G}|)}(X, \mathfrak{G})$. The largest non-trivial critical value from $\{T(gX) : g \in \mathfrak{G}\}$ is therefore the second largest order statistic $T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})$. The following theorem shows that the probability that $T(X)$ exceeds $T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})$ is necessarily small under $H_0\colon \mu_1 = \mu_0$. In fact, this probability is so small that $T(X) > T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})$ is well below any standard choice of α for most values of q_1 and q_0. By monotonicity, the existence of a j such that $P(T(X) > T^{(j)}(X, \mathfrak{G})) \leqslant \alpha$ is then guaranteed.

Theorem 2.1 (Size for heterogeneous symmetric variables). *Let $X = (X_1, \ldots, X_q)$ with $X_k = \mu + \sigma_k Z_k$, $1 \leqslant k \leqslant q$, where $\sigma_1, \ldots, \sigma_q > 0$ and the $Z_1, \ldots, Z_q$ are iid copies of a continuous random variable Z. If Z and $-Z$ have the same distribution, then*

$$\sup_{\mu \in \mathbb{R}, \sigma_1, \ldots, \sigma_q > 0} P\big(T(X) > T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})\big) = \frac{1}{2^{q_1 \wedge q_0}}.$$

The bound in the theorem also applies if the scales $\sigma_1, \ldots, \sigma_q$ are replaced by positive random variables that do not depend on $Z_1, \ldots, Z_q$. The X_k are then called “scale mixtures” of a symmetric distribution Z. The following corollary is immediately obtained from Theorem 2.1 by conditioning on the random scales.

Corollary 2.2 (Size for symmetric scale mixtures). *Suppose* $X = (X_1, \dots, X_q)$ *with* $X_k = \mu + S_k Z_k$, $1 \leqslant k \leqslant q$, *where the* $Z_1, \dots, Z_q$ *are iid copies of a continuous random variable* Z *and* $(S_1, \dots, S_q)$ *is a possibly dependent random vector independent of* $Z_1, \dots, Z_q$ *with* $P(S_k > 0) = 1$ *for* $1 \leqslant k \leqslant q$. *If* Z *and* $-Z$ *have the same distribution, then* $\sup_{\mu \in \mathbb{R}} P(T(X) > T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})) \leqslant 1/2^{q_1 \wedge q_0}$.

The bound $1/2^{q_1 \wedge q_0}$ in Theorem 2.1 is sharp. A test that uses $T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})$ as critical value has size 0.0625, 0.0313, 0.0156, 0.0078, 0.0039 as $q_1 \wedge q_0$ increases from 4 to 8. Consequently, a 10%-level permutation test that relies only on symmetry is available with q_1 and q_0 as small as 4. One can perform a 5%-level test with $q_1 \wedge q_0 \geqslant 5$, a 5%-level two-sided test (see the discussion below (2.5) ahead) with $q_1 \wedge q_0 \geqslant 6$, a 1%-level test with $q_1 \wedge q_0 \geqslant 7$, and a 1%-level two-sided test with $q_1 \wedge q_0 \geqslant 8$.

More generally, Theorem 2.1 implies that for many combinations of q_1, q_0, and α there exist $p \in (0,1)$ such that $\lceil (1-\alpha)|\mathfrak{G}| \rceil \leqslant \lceil (1-p)|\mathfrak{G}| \rceil < |\mathfrak{G}|$ and $P(T(X) > T^p(X, \mathfrak{G})) \leqslant \alpha$. The largest such value of p maximizes power while still controlling the size of the test. Finding this p or a close approximation of it is theoretically challenging but easily done via simulation if the distribution of Z is restricted to a single distribution. For normal distributions, the best possible p is

$$\bar{\alpha} = \sup \Bigg\{ p \in [0,1) : \sup_{\mu \in \mathbb{R}, \sigma_1, \dots, \sigma_q > 0} P\big(T(X) > T^p(X, \mathfrak{G})\big) \leqslant \alpha, \\ X \sim N\big(\mu, \operatorname{diag}(\sigma_1^2, \dots, \sigma_q^2)\big) \Bigg\}, \quad (2.4)$$

where I suppress the dependence on q_1 and q_0 to prevent notational clutter. By construction, $\bar{\alpha}$ controls the size of the permutation test not only for arbitrarily heterogeneous normal variables but also for the entire class of scale mixtures of normals. This class includes all Student t and Laplace distributions, as well as many other standard distributions (see, e.g., Gneiting, 1997). Moreover, because the critical value is from a permutation distribution, the test also controls size for all exchangeable distributions. The remainder of the paper therefore focuses on this $\bar{\alpha}$ and heterogeneous normal X but other choices of distributions are clearly possible.

Table 1 lists $\bar{\alpha}$ for common choices of α as a function of q_1 and q_0. (Appendix E outlines how these values were computed and how $\bar{\alpha}$ can be found for other choices of α, q_1, q_0.) As can be seen, the adjustment needed to make inference robust to variance heterogeneity is substantial if $q_1 \wedge q_0$ is very small but disappears quickly as $q_1 \wedge q_0$ increases. For example, for $q_1 = 4 = q_0$ a robust 10%-level test requires using the 95.62% quantile of the unadjusted test but for $q_1 = 9 = q_0$ the 91% quantile is

Table 1. Values for $\bar{\alpha}$ as defined in (2.4) as a function of q_1, q_0, and α.

		q_0								
α	q_1	4	5	6	7	8	9	10	11	12
.10	4	.0428								
	5	.0317	.0595							
	6	.0238	.0432	.0660						
	7	.0181	.0340	.0500	.0760					
	8	.0161	.0303	.0493	.0600	.0813				
	9	.0153	.0246	.0400	.0580	.0740	.0900			
	10	.0129	.0220	.0366	.0500	.0700	.0826	.0926		
	11	.0153	.0193	.0313	.0420	.0606	.0746	.0853	.0953	
	12	.0106	.0193	.0260	.0420	.0580	.0673	.0800	.0926	.0953
.05	5		.0158							
	6		.0108	.0227						
	7		.0088	.0200	.0253					
	8		.0062	.0120	.0233	.0306				
	9		.0113	.0120	.0213	.0300	.0393			
	10		.0100	.0113	.0166	.0286	.0340	.0420		
	11		.0100	.0080	.0153	.0240	.0313	.0393	.0440	
	12		.0073	.0080	.0153	.0213	.0266	.0366	.0440	.0491
.025	6			.0043						
	7			.0040	.0086					
	8			.0026	.0086	.0153				
	9			.0026	.0066	.0100	.0146			
	10			.0026	.0046	.0093	.0146	.0166		
	11			.0020	.0033	.0080	.0106	.0166	.0180	
	12			.0020	.0033	.0073	.0093	.0120	.0173	.0206
.01	7				.0026					
	8				.0013	.0026				
	9				.0013	.0020	.0033			
	10				.0013	.0020	.0033	.0040		
	11				.0013	.0020	.0033	.0040	.0066	
	12				.0013	.0013	.0026	.0033	.0053	.0066
.005	8					$\ast$				
	9					$\ast$	.0013			
	10					$\ast$	.0013	.0013		
	11					$\ast$	.0006	.0013	.0020	
	12					$\ast$	$\ast$	.0013	.0020	.0033

Note: $\ast$ means $T^{\bar{\alpha}}(X,\mathfrak{G})$ should be the second largest order statistic $T^{(|\mathfrak{G}|-1)}(X,\mathfrak{G})$. More values are available at `https://hgmn.github.io/ap`.

already sufficient for a robust 10%-level test. For larger numbers of variables the need for adjustment nearly disappears at conventional levels of significance. This is also confirmed by results in Hagemann (2019), who shows that unadjusted permutation inference in this context with the statistic $T(X)$ is consistent if the number of treated and control units grows in a balanced manner.

The test decision is now simple. For $q_1 \wedge q_0 > 3$, choose $\bar{\alpha}$ for a feasible α from Table 1 to ensure $P(T(X) > T^{\bar{\alpha}}(X, \mathfrak{G})) \leqslant \alpha$ under $H_0\colon \mu_1 = \mu_0$. The existence of such an $\bar{\alpha}$ for the comparison-of-means test statistic T is guaranteed by Theorem 2.1. For an α-level test of the null hypothesis $H_0\colon \mu_1 = \mu_0$, reject in favor of the alternative $H_1\colon \mu_1 > \mu_0$ if

$$T(X) > T^{\bar{\alpha}}(X, \mathfrak{G}). \tag{2.5}$$

For a one-sided test of level α against $\mu_1 < \mu_0$, reject if $T(-X) > T^{\bar{\alpha}}(-X, \mathfrak{G})$ or, equivalently, $T(X) < T^{(\lfloor |\mathfrak{G}| \bar{\alpha} \rfloor)}(X, \mathfrak{G})$. For a two-sided test of level 2α against $\mu_1 \neq \mu_0$, reject if $T(X) > T^{\bar{\alpha}}(X, \mathfrak{G})$ or $T(-X) > T^{\bar{\alpha}}(-X, \mathfrak{G})$. Test decisions can also be equivalently made with the p-value of the unadjusted test

$$\hat{p}(X, \mathfrak{G}) = \inf\{p \in (0,1) : T(X) > T^p(X, \mathfrak{G})\} = \frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} 1\{T(gX) \geqslant T(X)\} \tag{2.6}$$

because $T(X) > T^p(X, \mathfrak{G})$ if and only if $\hat{p}(X, \mathfrak{G}) \leqslant p$ for every $p \in (0,1)$. A p-value for a two-sided test can be defined as $2(\hat{p}(X, \mathfrak{G}) \wedge \hat{p}(-X, \mathfrak{G}))$. Reject the null hypothesis if the p-value does not exceed $\bar{\alpha}$ from Table 1 to perform an α-level test.

Appendix A contains additional results on power, stochastic approximation of $\mathfrak{G}$, and large sample approximation of X. The next section applies Theorem 2.1 to situations where X is the distributional limit of cluster-level statistics.

3. Permutation inference with heterogenous clusters

In this section, I establish large sample results for an adjusted permutation test with finitely many clusters under a single high-level condition. I then outline how these results can be applied in empirical practice.

Suppose data from q large clusters (e.g., counties, regions, schools, firms, or stretches of time) are available. Observations are independent across clusters but dependent within clusters. An intervention took place during which clusters $1 \leqslant k \leqslant q_1$ received treatment and clusters $q_1 + 1 \leqslant k \leqslant q$ did not. The quantity of interest is a treatment effect or an object related to a treatment effect that can be represented by a scalar parameter δ. Because entire clusters receive treatment, this parameter is only identified up to a location shift θ_0 within a treated cluster. Hence, only the left-hand side of

$$\theta_1 = \theta_0 + \delta$$

can be identified from such a cluster. If the clusters have similar characteristics, then θ_0 can be identified from an untreated cluster. Comparing the two clusters identifies δ.

The identification strategy outlined in the preceding paragraph is the basis for differences-in-differences estimation—arguably the most popular identification strategy in economics today—and a variety of other models. The purpose of this section is to use the results from Section 2 to develop a permutation test of the conventional (non-sharp) hypothesis

$$H_0\colon \delta = 0,$$

or, equivalently, $H_0\colon \theta_1 = \theta_0$. The idea is to obtain independent estimates $\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,q_1}$ of θ_1 and independent estimates $\hat{\theta}_{n,q_1+1}, \dots, \hat{\theta}_{n,q}$ of θ_0 so that $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,q})$ is approximately multivariate normal with diagonal covariance matrix. The following example outlines a simple situation where this is possible.

Example 3.1 (Difference in differences). Consider the regression model

$$Y_{t,k} = \theta_0 I_t + \delta I_t D_k + \beta_k' X_{t,k} + \zeta_k + U_{t,k}, \tag{3.1}$$

where k indexes individual units, t indexes time, $I_t = 1\{t > n_{0,k}\}$ indicates time periods after an intervention at a known time $n_{0,k}$, the dummy D_k indicates whether unit k eventually received treatment, and the ζ_k are individual fixed effects. Provided $U_{t,k}$ has conditional mean zero and the covariates $X_{t,k}$ vary before or after $n_{0,k}$, the data identify $\theta_1 = \theta_0 + \delta$ in a treated cluster and θ_0 in an untreated cluster. View each cluster as a separate regression and rewrite (3.1) as

$$Y_{t,k} = \begin{cases} \theta_1 I_t + \beta_k' X_{t,k} + \zeta_k + U_{t,k}, & 1 \leqslant k \leqslant q_1, \\ \theta_0 I_t + \beta_k' X_{t,k} + \zeta_k + U_{t,k}, & q_1 < k \leqslant q \end{cases} \tag{3.2}$$

and use the least squares estimates $\hat{\theta}_{n,k}$ of θ_1 and θ_0 as $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,q})$. □

The cluster-level statistics $\hat{\theta}_n$ can be combined with the results in the previous section to perform a consistent permutation test as the sample size n grows large. The test is not limited to the $\hat{\theta}_n$ constructed in the preceding example. Instead, the key high-level condition is that a centered and scaled version of some estimate $\hat{\theta}_n$ converges to a q-dimensional standard normal distribution,

$$\sqrt{n}\biggl(\frac{\hat{\theta}_{n,1} - \theta_1}{\sigma_1(\theta_1)}, \dots, \frac{\hat{\theta}_{n,q_1} - \theta_1}{\sigma_{q_1}(\theta_1)}, \frac{\hat{\theta}_{n,q_1+1} - \theta_0}{\sigma_{q_1+1}(\theta_0)}, \dots, \frac{\hat{\theta}_{n,q} - \theta_0}{\sigma_q(\theta_0)}\biggr) \rightsquigarrow N(0, I_q). \tag{3.3}$$

The $\sigma_1, \dots, \sigma_q$ may depend on θ_1 or θ_0 but are not presumed to be known or estimable by the researcher. This is an important feature of the test because consistent covariance matrix estimation would require knowledge of an explicit ordering of the dependence structure within each cluster. While ordering the data is straightforward for time-dependent data, it may be difficult or impossible to infer or credibly assume an ordering of the data within villages or schools. In contrast, (3.3) can be established under weak dependence assumptions where it is only presumed that there *exists* a possibly unknown ordering for which the dependence decays at a certain rate. El Machkouri, Volný, and Wu (2013) present easy-to-use moment bounds and limit theorems for this situation. See also Bester et al. (2011) and references therein for further results.

I now show that under the joint convergence (3.3) a permutation test based on comparison of means of $\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,q_1}$ and $\hat{\theta}_{n,q_1+1}, \dots, \hat{\theta}_{n,q}$ can be adjusted to be asymptotically of level α with a fixed number of clusters. This is possible for $q_1 \wedge q_0 > 3$ if $\bar{\alpha}$ in Table 1 is available at the desired significance level α. In that case, the test has power against fixed alternatives $\theta_1 = \theta_0 + \delta$ with $\delta > 0$ and local alternatives $\theta_1 = \theta_0 + \delta/\sqrt{n}$ converging to the null. In the latter situation, θ_0 is fixed and θ_1 implicitly depends on n. The convergence in (3.3) is then no longer pointwise in $\theta = (\theta_1, \theta_0)$ but a statement about the sequence $\theta_n = (\theta_0 + \delta/\sqrt{n}, \theta_0)$. As before, the test can be made two-sided to have power against fixed and local alternatives from either direction. Let $x \mapsto \tilde{\Phi}_{\theta_0}(x) = \prod_{1 \leqslant k \leqslant q_0} \Phi(x/\sigma_{k+q_1}(\theta_0))$.

Theorem 3.2 (Consistency and local power). *Suppose* (3.3) *holds. If* $\theta_1 = \theta_0$*, then*

$$\lim_{n\to\infty} P_\theta\big(T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})\big) \leqslant \alpha.$$

Let $\bar{\alpha} \geqslant 1/|\mathfrak{G}|$*. If* $\theta_1 = \theta_0 + \delta$ *with* $\delta > 0$*, then* $P_\theta(T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})) \to 1$*. If* $\theta_1 = \theta_0 + \delta/\sqrt{n}$ *and the* $\sigma_1, \dots, \sigma_q$ *are continuous and positive at* θ_0*, then*

$$\lim_{n\to\infty} P_{\theta_n}\big(T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})\big) \geqslant \int_0^1 \prod_{1\leqslant j \leqslant q_1} \Phi\left(\frac{\delta - \tilde{\Phi}_{\theta_0}^{-1}(t)}{\sigma_j(\theta_0)}\right) dt. \tag{3.4}$$

Remarks. (i) Because $T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})$ if and only if $T(a(\hat{\theta}_n - \theta_0 1_q)) > T^{\bar{\alpha}}(a(\hat{\theta}_n - \theta_0 1_q), \mathfrak{G})$, where $a > 0$ and 1_q is a q-vector of ones, the root-n rate in (3.3) and in the theorem can be replaced by any other rate as long as the asymptotic normal distribution in (3.3) is still attained. Several semiparametric or nonstandard estimators are therefore covered by the theorem.

(ii) It is sometimes of interest in applications to test the null hypothesis $H_0\colon \theta_1 = \theta_0 + \lambda$ for a given λ. In that case, define $\Lambda = (\lambda 1\{k \leqslant q_1\})_{1 \leqslant k \leqslant q}$ and reject if $T(\hat{\theta}_n - \Lambda) > T^{\bar{\alpha}}(\hat{\theta}_n - \Lambda, \mathfrak{G})$. Replace θ_0 by $\theta_0 + \lambda$ in Theorem 3.2 and use part (i) of this remark to see that this leads to a consistent test.

(iii) If evaluating all elements of $\mathfrak{G}$ is too costly, the computational burden can be reduced by working with a random sample $\mathfrak{G}_m$ of m random draws from $\mathfrak{G}$. As long as $m \to \infty$ and then $n \to \infty$, the theorem and parts (i)-(ii) of this remark also hold for $\mathfrak{G}_m$ with the exception of the local power bound if $\bar{\alpha}|\mathfrak{G}|$ happens to be an integer. In that case, the inequality (3.4) holds after subtracting $P(\hat{p}(Y, \mathfrak{G}) = \bar{\alpha})/2$ from its right-hand side, where $Y = (\sigma_1(\theta_0)Z_1, \dots, \sigma_q(\theta_0)Z_q)$, the $Z_1, \dots, Z_q$ are independent standard normal, and $\hat{p}$ is defined in (2.6). This corrects for the inherent discreteness of the test. (See also the discussion in Appendix A.) □

Example 3.3 (Difference in differences, cont.). Suppose there are $n_{0,k}$ pre-intervention and $n_{1,k}$ post-intervention periods for unit k. The data from the $n_k = n_{0,k} + n_{1,k}$ time periods available for unit k are the k-th cluster. Let $n = \sum_{k=1}^q n_k$. In the absence of covariates (i.e., $\beta_k \equiv 0$), each least squares estimate in (3.2) satisfies

$$\sqrt{n}(\hat{\theta}_{n,k} - \theta_0) = \left(\frac{n}{n_{1,k}}\right)^{1/2} n_{1,k}^{-1/2} \sum_{t=n_{0,k}+1}^{n_k} U_{t,k} - \left(\frac{n}{n_{0,k}}\right)^{1/2} n_{0,k}^{-1/2} \sum_{t=1}^{n_{0,k}} U_{t,k}$$

under H_0. If the pre-intervention and post-intervention periods are long in the sense that $n/n_{0,k} \to c_{0,k} \in (0, \infty)$ and $n/n_{1,k} \to c_{1,k} \in (0, \infty)$ for $1 \leqslant k \leqslant q$, then condition (3.3) already holds if $n^{-1/2}(\sum_{t=1}^{n_{0,k}} U_{t,k}, \sum_{t=n_{0,k}+1}^{n_k} U_{t,k})$ is independent across $1 \leqslant k \leqslant q$ and has a non-degenerate normal limiting distribution for each k. A large number of central limit theorems for time dependent data exist; see, e.g., White (2001). Alternatively, if relatively few post-intervention periods are available so that $n_1 = \sum_{k=1}^q n_{1,k}$ satisfies $n_1/n_{0,k} \to 0$ and $n_1/n_{1,k} \to c_{1,k} \in (0, \infty)$ for $1 \leqslant k \leqslant q$, the scale invariance of the test allows replacement of the $\sqrt{n}$ in (3.3) by $\sqrt{n_1}$. Then (3.3) holds if $n_{0,k}^{-1/2} \sum_{t=1}^{n_{0,k}} U_{t,k} = O_P(1)$ and $n_{1,k}^{-1/2} \sum_{t=n_{0,k}+1}^{n_k} U_{t,k}$ obeys a central limit theorem for $1 \leqslant k \leqslant q$. This argument also applies if relatively few pre-intervention periods are available with the roles of $n_{0,k}$ and $n_{1,k}$ reversed. If the pre-intervention and post-intervention periods are short, Theorem 2.1 implies that the permutation test can still be applied if $(U_{t,k})_{1 \leqslant t \leqslant n_k}$ is multivariate normal for $1 \leqslant k \leqslant q$.

The calculations in the preceding paragraph can be adjusted to include covariates. Similar calculations also apply if each cluster is a collection of individual-level data over time, although in that case more general limit theory is needed. See, e.g, Jenish and Prucha (2009) and El Machkouri et al. (2013) for appropriate results. □

Appendix B provides more practical guidance for the implementation of the adjusted permutation test and applies the test to several standard econometric models. The next section compares adjusted permutation inference to other methods.

4. Numerical results

This section studies the behavior of the adjusted permutation test and related methods in a Monte Carlo experiment and in data from a randomized trial. The discussion focuses on one-sided tests to the right but the results apply more generally.

Example 4.1 (Difference in differences, cont.). This example explores the behavior of the adjusted permutation (AP hereafter) test, the Ibragimov and Müller (2016, IM) test (see Appendix C for a description and more results), the Bester et al. (2011, BCH) test, and a clustered wild bootstrap (Cameron, Gelbach, and Miller, 2008, WCB) in a version of a Monte Carlo experiment in Conley and Taber (2011). The BCH test estimates parameters by least squares in the pooled sample and standardizes this estimate with the usual cluster-robust covariance matrix with a degrees-of-freedom adjustment. The resulting statistic is compared to the $1-\alpha$ quantile of t distribution with $q-1$ degrees of freedom. BCH show that this test is valid for certain ranges of q and α under regularity conditions if the distribution of the covariates is very similar across clusters. The WCB takes the same statistic but compares it to the bootstrap distribution of the statistic obtained from the cluster-robust version of the wild bootstrap using the Rademacher distribution and with the null hypothesis imposed on the data. This procedure is outlined in detail in Cameron et al. (2008). It is valid with $q \to \infty$ (Djogbenou, MacKinnon, and Nielsen, 2019) under mild homogeneity conditions and valid for fixed q under strong homogeneity conditions (Canay et al., 2020). The bootstrap shown here uses 199 repetitions.

The data generating process is the model in (3.1) specialized to

$$Y_{t,k} = \theta_0 I_t + \delta I_t D_k + \beta_1 X_{1,t,k} + \beta_2 X_{2,t,k} + \beta_3 X_{3,t,k} + \zeta_k + U_{t,k},$$

$$U_{t,k} = \rho U_{t-1,k} + V_{t,k}, \qquad X_{1,t,k} = \gamma I_t D_k + W_{t,k},$$

with $\theta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, $\zeta_k \equiv 1$, $\rho = 0.5$, and $\gamma = 0.8$. As before, $I_t = 1\{t > n_{0,k}\}$ is a post-intervention indicator and D_k is a treatment indicator. There are $n_{0,k} \equiv 10$ pre-intervention and $n_{1,k} \equiv 10$ post-intervention periods, six clusters received treatment, and six did not. I consider $(X_{2,t,k}, X_{3,t,k}, V_{t,k}, W_{t,k}) \sim N(0, \sigma_k^2 \mathrm{I})$ for every $1 \leqslant k \leqslant q$ and t. The experiment varies $\delta \in \{0,1,2,3\}$ and the heterogeneity of the clusters as follows: for $h \in \{1,3,5,7\}$, the last h clusters had $\sigma_{q-h+1} = \cdots = \sigma_q = 20$ and the remaining $q-h$ clusters had $\sigma_1 = \cdots = \sigma_{q-h} = 1$. The number h can therefore be viewed as a measure of heterogeneity of the clusters.

Table 2 shows the rejection frequencies of the four tests outlined above under the null and the alternative. Each entry was computed from 10,000 Monte Carlo

Table 2. Rejection frequencies of the adjusted permutation test (AP) test, Ibragimov-Müller (IM) test, Bester-Conley-Hansen (BCH) test, wild cluster bootstrap (WCB), and an oracle version of the Canay-Romano-Shaikh (CRS) test for increasing degrees of heterogeneity h in Example 4.1.

	AP	IM	BCH	WCB	oracle CRS	AP	IM	BCH	WCB	oracle CRS
h	$\delta = 0$ (size)					$\delta = 1$ (power)				
1	.0244	.0086	.0265	.0392	.0474	.2826	.1176	.2930	.3981	.4570
3	.0316	.0287	.0641	.0538	.0513	.1214	.0706	.1433	.1493	.1627
5	.0377	.0507	.0787	.0635	.0451	.0549	.0662	.1086	.0887	.0792
7	.0358	.0475	.0735	.0634	.0442	.0438	.0560	.0924	.0791	.0659
	$\delta = 2$ (power)					$\delta = 3$ (power)				
1	.5541	.3142	.5631	.6326	.6036	.6227	.4797	.7001	.7054	.6775
3	.1896	.1263	.2375	.2435	.2410	.2445	.1900	.3448	.3420	.3056
5	.0728	.0889	.1566	.1325	.1192	.0982	.1188	.2214	.1897	.1565
7	.0533	.0707	.1306	.1110	.0908	.0715	.0915	.1715	.1488	.1168

simulations and all methods were faced with the same data. As can be seen, all tests were conservative when there was little heterogeneity ($h = 1$). However, the BCH test and the WCB were no longer able to control size as the heterogeneity increased. The over-rejection in both methods led to higher rejection frequencies under the alternative, which therefore should not be viewed as evidence of their power. The AP test rejected far more false null hypotheses than the IM test when there was little heterogeneity. As the heterogeneity increased, the IM test had a slight advantage. The BCH test and the WCB performed well at $h = 1$. However, even then there was little cost to using the AP test. It rejected nearly as many false nulls as the BCH test and at most 11.55 percentage points fewer false nulls than the WCB but was able to control size.

Several other methods for inference specifically designed for difference in differences such as Donald and Lang (2007) and Conley and Taber (2011) are available. Here I focus only on methods that apply more broadly and that are valid with a fixed number of clusters. The test of Canay et al. (2017, CRS) technically applies here but requires matching each treated cluster with a control cluster. In the present example, there are $6! = 720$ potential matches and equally many potential tests. A substantial multiple testing correction would therefore be required. However, if a pilot study or pre-analysis plan prescribed the cluster pairs, the (randomized) CRS test would be asymptotically similar and therefore provides a useful benchmark for the AP test. To this end, Table 2 shows results of an oracle version of the CRS test that presumes that a pre-analysis plan is in place. As can be seen, the AP test compares well to the CRS test while completely avoiding the multiple testing issue. □

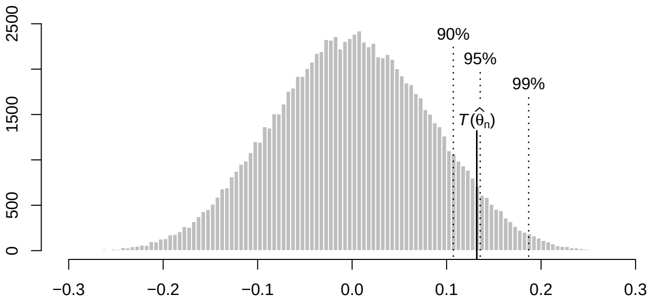

FIGURE 1. Histogram of the permutation distribution of $T(\hat{\theta}_n) \approx 0.132$ (solid black line) from Example 4.2 with 90%, 95%, and 99% critical values (dotted lines).

Example 4.2 (Achievement awards; Angrist and Lavy 2009). In this example, I reanalyze data from a randomized trial of Angrist and Lavy (2009) in Israel. Their intervention provided cash rewards to low-achieving high school students if they performed well on the Bagrut, a sequence of certification exams that is the formal prerequisite for university admission in Israel. I follow the analysis in Table 5 of Angrist and Lavy (2009) and focus on 32 schools in the sample for which Bagrut rates from 2000 to 2002 are available. Of these schools, 15 received treatment and 17 did not. Because 5 schools did not comply with treatment, the estimates below should be interpreted as intent-to-treat effects. Following Angrist and Lavy, I investigate the performance of girls in the June 2001 exams who were close to achieving Bagrut certification in the sense that they were ranked above the median of the credit-weighted January 2001 scores of girls. The sample also includes all girls who were above the median in 2000 and 2002. The 2948 girls who met these criteria had an above 50% chance of achieving Bagrut certification. I view each school over time as a cluster, which yields an average cluster size of approximately 92 students.

Angrist and Lavy (2009) report a large number of specifications. I consider a version of their fixed-effects model and estimate $Y_{i,t,k} = \theta_0 I_t + \delta D_k I_t + \eta J_t + \beta top_i + \zeta_k + U_{i,t,k}$, where i indexes students, t indexes time, k indexes schools, $Y_{i,k}$ indicates Bagrut status, D_k is the treatment indicator, I_t equals 1 in 2001 and is 0 otherwise, J_t equals 1 in 2002 and is 0 otherwise, top_i indicates whether a student is in the top quartile of the pre-Bagrut grade distribution of girls in the cohort, and ζ_k is a school fixed effect. Angrist and Lavy estimate several related specifications by logit in their Table 5. They report heteroskedasticity-robust standard errors for that table and argue that clustering is accounted for by their fixed effects. For simplicity and ease of interpretation, I estimate the model by least squares. The model predicts an average increase in the probability of receiving Bagrut status by 0.114 relative to a mean of

0.539 with a robust standard error of 0.037. A null of no effect against the alternative that δ is positive is rejected at any conventional significance level if standard normal critical values are used. This is in line with Table 5, col. (3) of Angrist and Lavy (2009), who report significant effects ranging from 0.093 to 0.168 with standard errors ranging from 0.039 to 0.045 for this sample and several subsamples.

To apply the adjusted permutation test, I view each cluster as an individual regression and separately estimate each of the $q = 32$ equations in

$$Y_{i,t,k} = \begin{cases} \theta_1 I_t + \eta J_t + \beta top_i + \zeta_k + U_{i,t,k}, & 1 \leqslant k \leqslant 15, \\ \theta_0 I_t + \eta J_t + \beta top_i + \zeta_k + U_{i,t,k}, & 15 < k \leqslant 32. \end{cases}$$

Note that ζ_k is now simply the constant term in each regression. The resulting test statistic $T(\hat{\theta}_n) \approx 0.132$ can be viewed as an alternative point estimate of δ and is comparable in magnitude to the estimates reported in Angrist and Lavy (2009). However, as can be seen in Figure 1, which plots the permutation distribution from 100,000 draws together with the corresponding critical values, the adjusted permutation test only rejects the null of no effect in favor of a positive effect at the 10% level and barely does not reject at the 5% level. If the fixed effects in the regression do not fully account for the within-cluster dependence in the data, the positive effect for girls may therefore be far less significant than previously reported. This result in also line with Angrist and Lavy, who find substantial but statistically marginal positive effects for girls across a wide variety of plausible specifications when they use cluster-robust standard errors. Also note that the 5% and 10% level one-sided tests performed here are outside the feasible range of the Ibragimov and Müller (2016) test. For the Canay et al. (2017) test, there are $17!/2 \approx 1.78 \times 10^{14}$ ways of testing if 15 treated clusters are paired with 15 control clusters and two control clusters are dropped. In 1,000 randomly chosen unique pairings, the Canay et al. (2017) test rejected the null of no effect against $\delta > 0$ for 425 pairings at the 5% level and in 48 pairings at the 1% level. Any desired conclusion could be reached by choosing a specific pairing. □

Appendix C contains additional numerical examples and empirical applications.

References

Angrist, J. and V. Lavy (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review 99*, 301–331.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*, 249–275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*, 137–151.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–427.

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

Canay, I. A., A. Santos, and A. M. Shaikh (2020). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics*, forthcoming.

Conley, T. G. and C. R. Taber (2011). Inference with "difference in differences" with a small number of policy changes. *Review of Economics and Statistics 93*, 113–125.

Djogbenou, A., J. G. MacKinnon, and M. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*, 393–412.

Donald, S. G. and K. Lang (2007). Inference with difference-in-differences and other panel data. *Review of Economics and Statistics 89*, 221–233.

El Machkouri, M., D. Volný, and W. B. Wu (2013). A central limit theorem for stationary random fields. *Stochastic Processes and their Applications 123*, 1–14.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*, 607–636.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation 59*, 375–384.

Hagemann, A. (2019). Placebo inference on treatment effects when the number of clusters is small. *Journal of Econometrics 213*, 190–209.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics 23*, 169–192.

Ibragimov, R. and U. Müller (2010). *t*-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics 28*, 453–468.

Ibragimov, R. and U. Müller (2016). Inference with few heterogenous clusters. *Review of Economics and Statistics 98*, 83–06.

Jenish, N. and I. R. Prucha (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics 150*, 86–98.

MacKinnon, J. G. and M. D. Webb (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics 32*, 233–254.

White, H. (2001). *Asymptotic Theory for Econometricians* (revised ed.). Academic Press, San Diego.

SUPPLEMENTAL APPENDIX TO "PERMUTATION INFERENCE WITH A FINITE NUMBER OF HETEROGENEOUS CLUSTERS"[*]

This supplemental appendix is organized as follows: Appendix A presents additional theoretical results, some of which are of potentially independent interest. Appendix B provides a step-by-step procedure for implementing the adjusted permutation test and applies that procedure in several examples. Appendix C contains additional numerical results and comparisons with the test of Ibragimov and Müller (2016). Appendix D contains proofs. Appendix E presents a simple algorithm for simulating critical values beyond those found in Table 1 in the main text.

Appendix A. Additional theoretical results

I start with a discussion of the behavior of the test under the alternative $H_1\colon \mu_1 > \mu_0$. (Tests in the other direction follow by considering $-X$ instead of X.) Let $\delta = \mu_1 - \mu_0$ and denote by Φ the standard normal distribution function. The distribution function of $\max_{1\leqslant k\leqslant q_0} X_{q_1+k}$ is equal to $x \mapsto \prod_{1\leqslant k\leqslant q_0} \Phi(x/\sigma_{k+q_1}) =: \tilde{\Phi}(x)$ and therefore has a continuous and strictly increasing inverse. The following result gives a simple lower bound on the power of a permutation test as a function of δ, Φ, $\tilde{\Phi}$, and the standard deviations in the treatment group. Here I assume that the α under consideration is feasible, i.e., the corresponding $\bar{\alpha}$ satisfies $\lceil (1-\bar{\alpha})|\mathfrak{G}|\rceil < |\mathfrak{G}|$ or, equivalently, $\bar{\alpha} \geqslant 1/|\mathfrak{G}|$. Otherwise the test becomes trivial because the null is never rejected.

Theorem A.1 (Power). *Suppose $X = (X_1, \dots, X_q)$ with independent $X_k \sim N(\mu + \delta 1\{k \leqslant q_1\}, \sigma_k^2)$, $1 \leqslant k \leqslant q$. Let $\bar{\alpha} \geqslant 1/|\mathfrak{G}|$. Then, for every $\sigma_1, \dots, \sigma_q > 0$,*

$$\inf_{\mu\in\mathbb{R}} P\big(T(X) > T^{\bar{\alpha}}(X, \mathfrak{G})\big) \geqslant \int_0^1 \prod_{1\leqslant j\leqslant q_1} \Phi\left(\frac{\delta - \tilde{\Phi}^{-1}(t)}{\sigma_j}\right) dt.$$

As can be expected, the power of the test is driven by the strength of the signal δ relative to the noise represented by the standard deviations $\sigma_1, \dots, \sigma_q$. For example, a small treatment effect δ can be drowned out by large variation in the control group because $t \mapsto \tilde{\Phi}^{-1}(t)$ will then be positive and large for most values of t. However, the power of the test is not inherently limited. The integrand on the right is bounded by 1 and converges to 1 as $\delta \to \infty$ pointwise for every t. The integral and consequently the power of the permutation test therefore approach 1 by dominated convergence as

[*] Andreas Hagemann, University of Michigan.

$\delta \to \infty$. Both the bound and this result can be generalized to the symmetric scale mixtures from Corollary 2.2; see Lemma D.1 for details.

Next, I discuss several aspects of the practical implementation of the permutation test (2.5). First, one can still perform an asymptotic α-level test if the observed data or statistic X_n converges in distribution to the X considered in Theorem 2.1 or Corollary 2.2. The reason is that the g that order $T(gX_n)$ and $T(gX)$ as g varies over $\mathfrak{G}$ eventually coincide if sufficiently many entries of X are smooth. The proof is a consequence of arguments in Canay et al. (2017).

Proposition A.2 (Large sample approximation). *Let $X_n \rightsquigarrow X \in \mathbb{R}^q$ and let T be as in* (2.1)*. If X has independent entries of which more than $q_1 \wedge q_0$ are continuously distributed, then*

$$\lim_{n\to\infty} P\big(T(X_n) > T^{(j)}(X_n, \mathfrak{G})\big) = P\big(T(X) > T^{(j)}(X, \mathfrak{G})\big), \qquad \textit{every } 1 \leqslant j \leqslant |\mathfrak{G}|.$$

Second, if evaluating $T(gX)$ over all elements of $\mathfrak{G}$ is too costly because $|\mathfrak{G}| = \binom{q}{q_1}$ is large, the computational burden can be reduced by working with a random sample $\mathfrak{G}_m$ of m draws from the uniform distribution on $\mathfrak{G}$. This is often referred to as "stochastic approximation." The following result shows that the critical values $T^p(X, \mathfrak{G}_m)$ and $T^p(X, \mathfrak{G})$ lead to identical test decisions for any p and large m as long as $p|\mathfrak{G}|$ is not an integer. If $p|\mathfrak{G}|$ *is* in fact an integer, the stochastic approximation can be marginally more conservative. The reason is that $p \mapsto T^p(X, \mathfrak{G})$ can vary discontinuously at integer values of $p|\mathfrak{G}|$. The stochastic approximation then hits the order statistic just above $T^p(X, \mathfrak{G})$ with nonzero probability. The same arguments apply if the identity transformation is always included in $\mathfrak{G}_m$, which is common practice for randomization tests.

Proposition A.3 (Stochastic approximation). *Let $X_n \in \mathbb{R}^q$ be an arbitrary random vector possibly depending on n. Suppose $\mathfrak{G}_m$ is a collection of m random draws from $\mathfrak{G}$ independent of X_n. Then*

$$\lim_{m\to\infty} P\big(T(X_n) > T^p(X_n, \mathfrak{G}_m)\big) \leqslant P\big(T(X_n) > T^p(X_n, \mathfrak{G})\big), \qquad \textit{every } p \in (0,1),$$

with equality unless $p|\mathfrak{G}| \in \mathbb{N}$. The result remains true if one of the members of $\mathfrak{G}_m$ is replaced by the identity with probability one.

As a referee points out, the choice of m is important in practice. In particular, it seems if $|\mathfrak{G}|$ is large, then m must be large as well to provide an accurate stochastic approximation of the test decision. However, this is only true if the p-value $\hat{p}(X, \mathfrak{G}_m)$, as defined in (2.6), is very close to $\bar{\alpha}$. If $\hat{p}(X, \mathfrak{G}_m)$ is much larger than $\bar{\alpha}$ for a given m, there is often enough information to conclude that $\hat{p}(X, \mathfrak{G})$ is highly unlikely to be smaller than $\bar{\alpha}$. The same is true if the direction of the inequalities is reversed. The reason is that $\mathrm{E}(\hat{p}(X, \mathfrak{G}_m) \mid X) = \hat{p}(X, \mathfrak{G})$ and, for almost every realization of X, the central limit theorem implies that $\sqrt{m}(\hat{p}(X, \mathfrak{G}_m) - \hat{p}(X, \mathfrak{G}))$ converges to mean-zero normal with variance $\hat{p}(X, \mathfrak{G})(1 - \hat{p}(X, \mathfrak{G}))$. It is therefore easy to test hypotheses of the form $\hat{p}(X, \mathfrak{G}) \geqslant \bar{\alpha}$ or $\hat{p}(X, \mathfrak{G}) \leqslant \bar{\alpha}$ with a very small error tolerance β. For example, if $\hat{p}(X, \mathfrak{G}_m) > \bar{\alpha}$ for a given m, one can check whether $\hat{p}(X, \mathfrak{G}) \leqslant \bar{\alpha}$ can be rejected at this m. If not, one can add draws from $\mathfrak{G}$ until the decision becomes possible. This idea is, in fact, the basis for the widely-used algorithm of Davidson and MacKinnon (2000) for determining a sufficient number of bootstrap repetitions in models where the bootstrap is expensive to compute. Their algorithm can be adapted to the present problem with only notational changes.

Algorithm A.4 (Choosing m if $|\mathfrak{G}|$ is very large). Choose a starting value m (e.g., 10,000), a step size m' (e.g., 1,000), a maximal number of permutations $m_{\max}$ (e.g., 100,000), and an error tolerance β (e.g., .001).

(1) If $\hat{p}(X, \mathfrak{G}_m) < \bar{\alpha}$, test the null hypothesis $\hat{p}(X, \mathfrak{G}) \geqslant \bar{\alpha}$ by rejecting in favor of $\hat{p}(X, \mathfrak{G}) < \bar{\alpha}$ if $\sqrt{m}(\hat{p}(X, \mathfrak{G}_m) - \bar{\alpha})/\sqrt{\bar{\alpha}(1 - \bar{\alpha})} < \Phi^{-1}(\beta)$. Stop if the null is rejected and use $\hat{p}(X, \mathfrak{G}_m)$ as if it were $\hat{p}(X, \mathfrak{G})$.

(2) If $\hat{p}(X, \mathfrak{G}_m) > \bar{\alpha}$, test the null hypothesis $\hat{p}(X, \mathfrak{G}) \leqslant \bar{\alpha}$ by rejecting in favor of $\hat{p}(X, \mathfrak{G}) > \bar{\alpha}$ if $\sqrt{m}(\hat{p}(X, \mathfrak{G}_m) - \bar{\alpha})/\sqrt{\bar{\alpha}(1 - \bar{\alpha})} > \Phi^{-1}(1 - \beta)$. Stop if the null is rejected and use $\hat{p}(X, \mathfrak{G}_m)$ as if it were $\hat{p}(X, \mathfrak{G})$.

(3) Stop if $m + m' > m_{\max}$ and use $\hat{p}(X, \mathfrak{G}_m)$ as if it were $\hat{p}(X, \mathfrak{G})$. Otherwise draw m' additional permutations from $\mathfrak{G}$, set $m = m + m'$, and restart from step (1).

Finally, the two approximation results in Propositions A.2 and A.3 can be combined with Theorem 2.1 to obtain

$$\lim_{n\to\infty} \lim_{m\to\infty} P\big(T(X_n) > T^{\bar{\alpha}}(X_n, \mathfrak{G}_m)\big) \leqslant \alpha,$$

i.e., adjusted permutation inference with an asymptotically normally distributed vector with heterogeneous variances remains approximately valid even if the set of permutations is drawn at random. It should also be noted that Proposition A.3 is generic and can be restated for other statistics T and finite groups with appropriate notational changes. Proposition A.2 can be extended to other statistics and groups under smoothness conditions.

Appendix B. Additional examples

I first present a brief summary of how the permutation test can be implemented in practice. By Theorem 3.2, the following procedure provides an asymptotically α-level test in the presence of a finite number of large clusters that are arbitrarily heterogeneous. The test is free of nuisance parameters, does not require matching clusters or any other decisions on part of the researcher, can be two-sided or one-sided in either direction, and is able to detect all fixed and $1/\sqrt{n}$-local alternatives.

Algorithm B.1 (Permutation test adjusted for cluster heterogeneity).

(1) Order the data such that clusters $1 \leqslant k \leqslant q_1$ received treatment and clusters $q_1+1 \leqslant k \leqslant q_1+q_0 = q$ did not. Compute for each $k = 1, \dots, q$ and using only data from cluster k an estimate $\hat{\theta}_{n,k}$ of either θ_1 or θ_0 depending on whether k received treatment or not so that the difference $\theta_1 - \theta_0$ is the treatment effect of interest. (Examples are provided below and in the main text.) Define $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,q})$ and compute $T(\hat{\theta}_n) = q_1^{-1}\sum_{k=1}^{q_1} \hat{\theta}_{n,k} - q_0^{-1}\sum_{k=q_1+1}^{q} \hat{\theta}_{n,k}$.

(2) For the desired α, choose $\bar{\alpha}$ from Table 1.

(3) Compute the set of permutations $\mathfrak{G}$ defined in (2.2). Alternatively, draw a large random sample of permutations $\mathfrak{G}_m$ and replace $\mathfrak{G}$ by $\mathfrak{G}_m$ in step (4).

(4) Reject the null hypothesis of no effect of treatment $H_0\colon \theta_1 = \theta_0$ against
   (a) $\theta_1 > \theta_0$ if $T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})$ for a test with asymptotic level α,
   (b) $\theta_1 < \theta_0$ if $T(-\hat{\theta}_n) > T^{\bar{\alpha}}(-\hat{\theta}_n, \mathfrak{G})$ for a test with asymptotic level α,
   (c) $\theta_1 \neq \theta_0$ if $T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})$ or $T(-\hat{\theta}_n) > T^{\bar{\alpha}}(-\hat{\theta}_n, \mathfrak{G})$ for a test with asymptotic level 2α,

   where $T^{\bar{\alpha}}(\cdot, \mathfrak{G})$, defined in (2.3), is the $\lceil (1-\bar{\alpha})|\mathfrak{G}| \rceil$-th largest value of the permutation distribution of $T(\cdot)$.

I now discuss two additional examples of how the cluster-level statistics $\hat{\theta}_n$ can be constructed such that the condition (3.3) required for Theorem 3.2 holds. For simplicity, the discussion focuses on (3.3) under the null hypothesis $H_0 : \theta_1 = \theta_0$ but the arguments apply more broadly.

Example B.2 (Regression with cluster-level treatment). Consider a linear regression model

$$Y_{i,k} = \theta_0 + \delta D_k + \beta_k' X_{i,k} + U_{i,k},$$

where i indexes individuals within clusters $1 \leqslant k \leqslant q$. The parameter of interest is the coefficient δ on the treatment dummy D_k indicating whether cluster k received treatment or not. The regression also includes covariates $X_{i,k}$ that vary within each cluster and have coefficients β_k that may vary across clusters. The condition $\mathrm{E}(U_{i,k} \mid D_k, X_{i,k}) = 0$ identifies $\theta_1 = \theta_0 + \delta$ within a treated cluster and θ_0 within an untreated cluster. The preceding display can then be written as

$$Y_{i,k} = \begin{cases} \theta_1 + \beta_k' X_{i,k} + U_{i,k}, & 1 \leqslant k \leqslant q_1, \\ \theta_0 + \beta_k' X_{i,k} + U_{i,k}, & q_1 < k \leqslant q. \end{cases}$$

View these as q separate regressions and use the least squares estimates of the constants θ_1 and θ_0 as $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,q})$. Also note that permuting $\hat{\theta}_n$ is identical to permuting the vector of the observed treatment indicators that labels each of these q regressions as coming from either a treated or an untreated cluster. The same types of arguments as in Example 3.1 can be used to establish a central limit theorem for $\hat{\theta}_n$.

Under suitable conditions, the δ in this example can be interpreted as an average treatment effect in a potential outcomes framework. See, e.g., Słoczyński (2018) and references therein for a precise discussion. The goal here is to make permutation inference about δ. This should not be confused with testing the "sharp" null hypothesis that the treatment and control potential outcomes under the intervention are identical. Testing sharp nulls is often associated with permutation testing and is a much stronger restriction than that the *average* effect δ on the outcomes be zero. Rosenbaum (1984) explains how to use permutation inference to test sharp nulls in the presence of covariates under assumptions on the propensity score. □

Example B.3 (Binary choice with cluster-level treatment). Consider a version of the model in Example B.2 as the latent model $Y_{i,k} = \theta_0 + \delta D_k + \beta_0' X_{i,k} + U_{i,k}$ in a binary choice setting. Here $U_{i,k}$ has a known, smooth, and symmetric distribution function F and is independent of $(D_k, X_{i,k})$. Only $1\{Y_{i,k} > 0\}$, $X_{i,k}$, and D_k are observed. Each cluster has n_k observations and can be viewed as a separate binary choice model

$$P(Y_{i,k} > 0 \mid X_{i,k}) = \begin{cases} F(\theta_1 + \beta_0' X_{i,k}), & 1 \leqslant k \leqslant q_1, \\ F(\theta_0 + \beta_0' X_{i,k}), & q_1 < k \leqslant q. \end{cases}$$

If the treatment effect of interest is $F(\theta_1 + \beta_0' x) - F(\theta_0 + \beta_0' x)$ for some x, then $H_0\colon \theta_1 = \theta_0$ corresponds to the null hypothesis of no treatment effect. Let $\psi_{\theta,\beta}(y, x) = (1, x')'(1\{y > 0\} - F(\theta + \beta' x))$ and suppose the moment condition $\mathrm{E}\psi_{\theta_0,\beta_0}(Y_{i,k}, X_{i,k}) = 0$ holds for every i and k. The corresponding Z-estimates $(\hat{\theta}_{n,k}, \hat{\beta}_{n,k}')'$ for the k-th cluster are zeros of $\Psi_{n,k}(\theta, \beta) = n_k^{-1} \sum_{i=1}^{n_k} \psi_{\theta,\beta}(Y_{i,k}, X_{i,k})$. Denote the derivative of $\Psi_{n,k}$ with respect to (θ, β') by $\dot{\Psi}_{n,k}$.

Using the same limit theory as outlined in Example 3.3, it is possible to argue under regularity conditions that $\dot{\Psi}_{n,k}$ converges pointwise in probability to a limit $\dot{\Psi}_k$ and $(\hat{\theta}_{n,k}, \hat{\beta}_{n,k}) \xrightarrow{\mathrm{P}} (\theta_0, \beta_0)$. If $\dot{\Psi}_k(\theta_0, \beta_0)$ is non-singular and $\sqrt{n}\Psi_{n,k}(\theta_0, \beta_0) = O_P(1)$, then

$$\sqrt{n}(\hat{\theta}_{n,k} - \theta_0) = e_1' \dot{\Psi}_k(\theta_0, \beta_0)^{-1} \sqrt{n}\Psi_{n,k}(\theta_0, \beta_0) + o_P(1),$$

where e_1 is a conformable vector with a 1 in the first position and 0 otherwise. Condition (3.3) is satisfied if a central limit theorem applies to $\sqrt{n}\Psi_{n,k}(\theta_0, \beta_0)$. Because this is a scaled average of mean-zero random vectors, the same references as in Example 3.3 can be used to establish a central limit theorem. □

Appendix C. Additional numerical results

This section presents a detailed comparison of the Ibragimov and Müller (2016) and adjusted permutation tests in Monte Carlo experiments and empirical examples.

Example C.1 (Equality of means). The adjusted permutation test developed here and the Ibragimov and Müller (2016) test both rely on results about the behavior of heterogeneous normal variables applied to certain test statistics. For the adjusted permutation test, this statistic is the comparison on means T. For the Ibragimov-Müller test, it is the studentized two-sample statistic

$$\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{\frac{1}{q_1(q_1-1)}\sum_{k=1}^{q_1}(X_k - \bar{X}_1)^2 + \frac{1}{q_0(q_0-1)}\sum_{k=q_1+1}^{q}(X_k - \bar{X}_0)^2}},$$

where $\bar{X}_1 = q_1^{-1}\sum_{k=1}^{q_1} X_k$ and $\bar{X}_0 = q_0^{-1}\sum_{k=q_1+1}^{q} X_k$. This statistic is compared to the quantiles of the Student t distribution with $(q_1 \wedge q_0) - 1$ degrees of freedom. This example investigates the relative performance of the two tests.

As in Section 2, suppose $X = (X_1, \dots, X_q) \in \mathbb{R}^q$ has independent entries $X_k = \mu_0 + (\mu_1 - \mu_0)1\{k \leqslant q_1\} + \sigma_k Z_k$ with Z_k distributed as $N(0, 1)$. The results reported here use $\mu_0 = 0$. To investigate the impact of heterogeneity on the two tests, I considered the following six configurations of $\sigma_1, \dots, \sigma_q$:

(a) $\sigma_1, \dots, \sigma_q = 1$,

(b) $\sigma_1, \dots, \sigma_{q-1} = 1$, $\sigma_q = 100$

(c) $\sigma_1, \dots, \sigma_{q_1-1} = 1$, $\sigma_{q_1} = 100$, $\sigma_{q_1+1}, \dots, \sigma_{q-1} = 1$, $\sigma_q = 100$,

(d) $\sigma_1, \dots, \sigma_{q_1} = 1$, $\sigma_{q_1+1}, \dots, \sigma_q = 3$

(e) $\sigma_1, \dots, \sigma_{q_1/2} = 3$, $\sigma_{q_1/2+1}, \dots, \sigma_{q_1+q_0/2} = 1$, $\sigma_{q_1+q_0/2+1}, \dots, \sigma_q = 3$,

(f) $\sigma_1, \dots, \sigma_{q_1/2} = 1$, $\sigma_{q_1/2+1}, \dots, \sigma_{q_1+q_0/2} = 3$, $\sigma_{q_1+q_0/2+1}, \dots, \sigma_q = 9$.

Configurations (a), (d), (e), and (f) are taken from Ibragimov and Müller (2016).

Rows (a)-(f) of Figure 2 correspond to the six configurations (a)-(f) and show the rejection frequencies of the adjusted permutation test (black lines) and the Ibragimov-Müller test (grey) at the 5% level (dashed line) as μ_1 increases. The null hypothesis is correct at $\mu_1 = 0$. The columns correspond, from left to right, to the sample sizes $(q_1 = 8, q_0 = 8)$, $(q_1 = 8, q_0 = 16)$, and $(q_1 = 16, q_0 = 16)$. Each horizontal coordiate was computed from 10,000 Monte Carlo replications. As can be seen, the variation in σ_k led to marked differences in power at different levels of heterogeneity. The adjusted permutation test was able to reject far more false nulls than the Ibragimov-Müller test for small μ_1 when there were few large variances as in (b) and (c). For instance, in (b) with $(q_1 = 8, q_0 = 8)$ at $\mu_1 = 1$ the adjusted permutation test rejected in 47.62% of all cases whereas the Ibragimov-Müller test rejected in only 6.36% of all cases. This difference eventually disappeared for large μ_1. However, neither test is more powerful. With slightly different variances within or across groups as in (d) and (f), the Ibragimov-Müller test had an advantage when the sample sizes differed substantially. The differences between the two tests were much smaller for the other configurations. Other samples sizes (not shown) led to qualitatively similar results.

As a referee points out, it would be interesting to compare the performance of the adjusted permutation test and the Ibragimov-Müller test in fat-tailed settings. Just like the adjusted permutation test, the Ibragimov-Müller test can be used with mixtures of normals, which includes models with infinite variances. I therefore repeated the above experiments with standard Cauchy distributed Z_k instead of standard normal distributions, holding all else equal. The results are plotted in Figure 3. As can be seen, within the scope of the configurations for (a)-(f), the adjusted permutation test was more powerful than the Ibragimov-Müller test for *every* configuration at all sample sizes and for all values of μ_1. In sharp contrast to the situation with standard normal Z_k, this was true even when the samples sizes differed. □

A reviewer also recommends comparing the conclusions of adjusted permutation inference and the Ibragimov-Müller test in empirical examples discussed in Ibragimov and Müller (2016), which include tests of hypotheses on January effects and a randomized trial of Bloom, Eifert, Mahajan, McKenzie, and Roberts (2013).

Example C.2 (January effects; Keim 1983). Keim (1983) investigates January effects in stock returns. He considers excess returns in portfolios constructed from firms in the top and bottoms decile of size, as measured by market value of equity on the New York Stock Exchange (NYSE) and American Stock Exchange (now called NYSE American) over the period 1963-1979. To test whether the January effect is

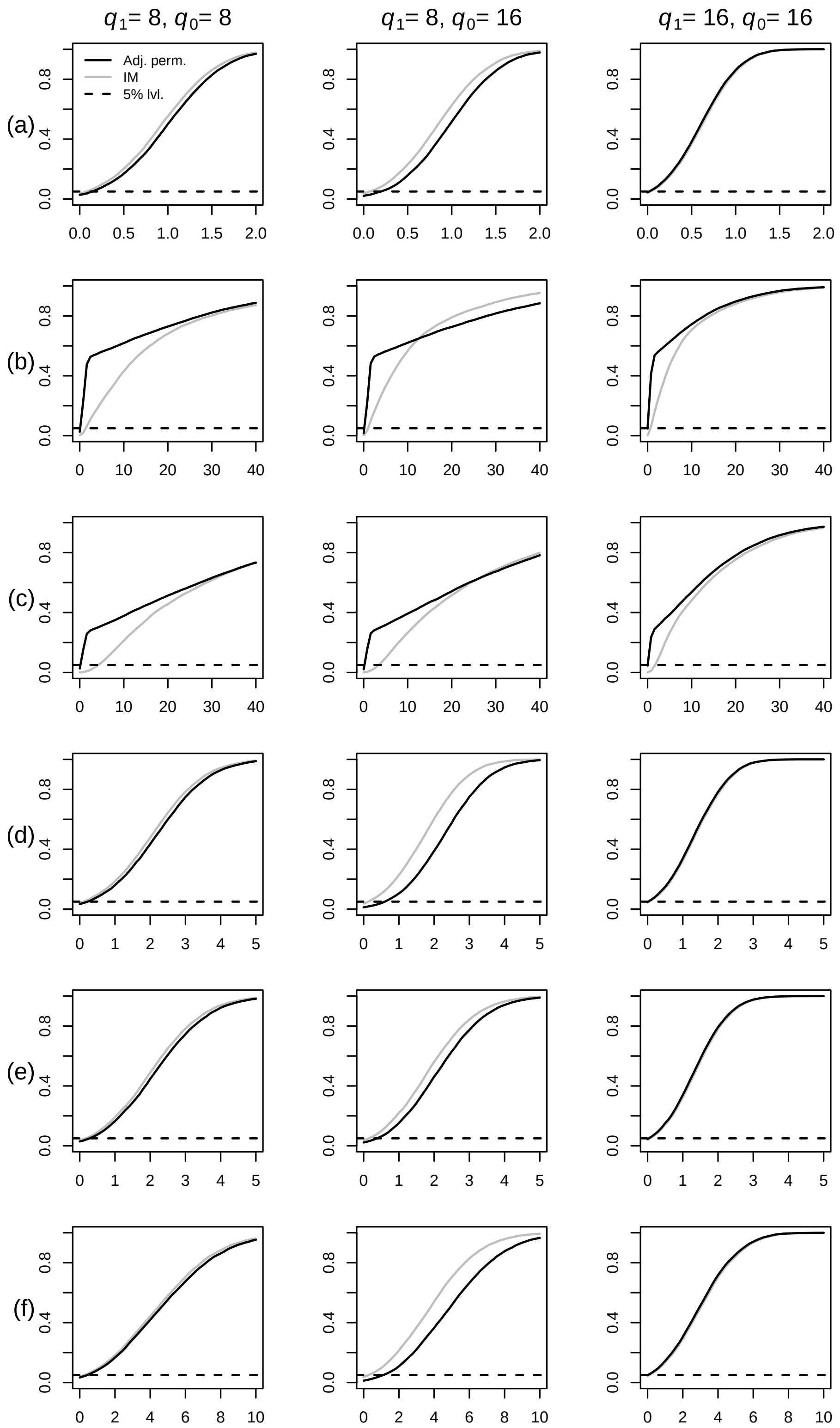

FIGURE 2. Rejection frequencies of the adjusted permutation test (black lines) and the Ibragimov-Müller test (IM, grey) for models (a)-(f) (rows) in Example C.1 for $q_1 = q_0 = 8$ (left), $q_1 = 8, q_0 = 16$ (middle), and $q_1 = q_0 = 16$ (right).

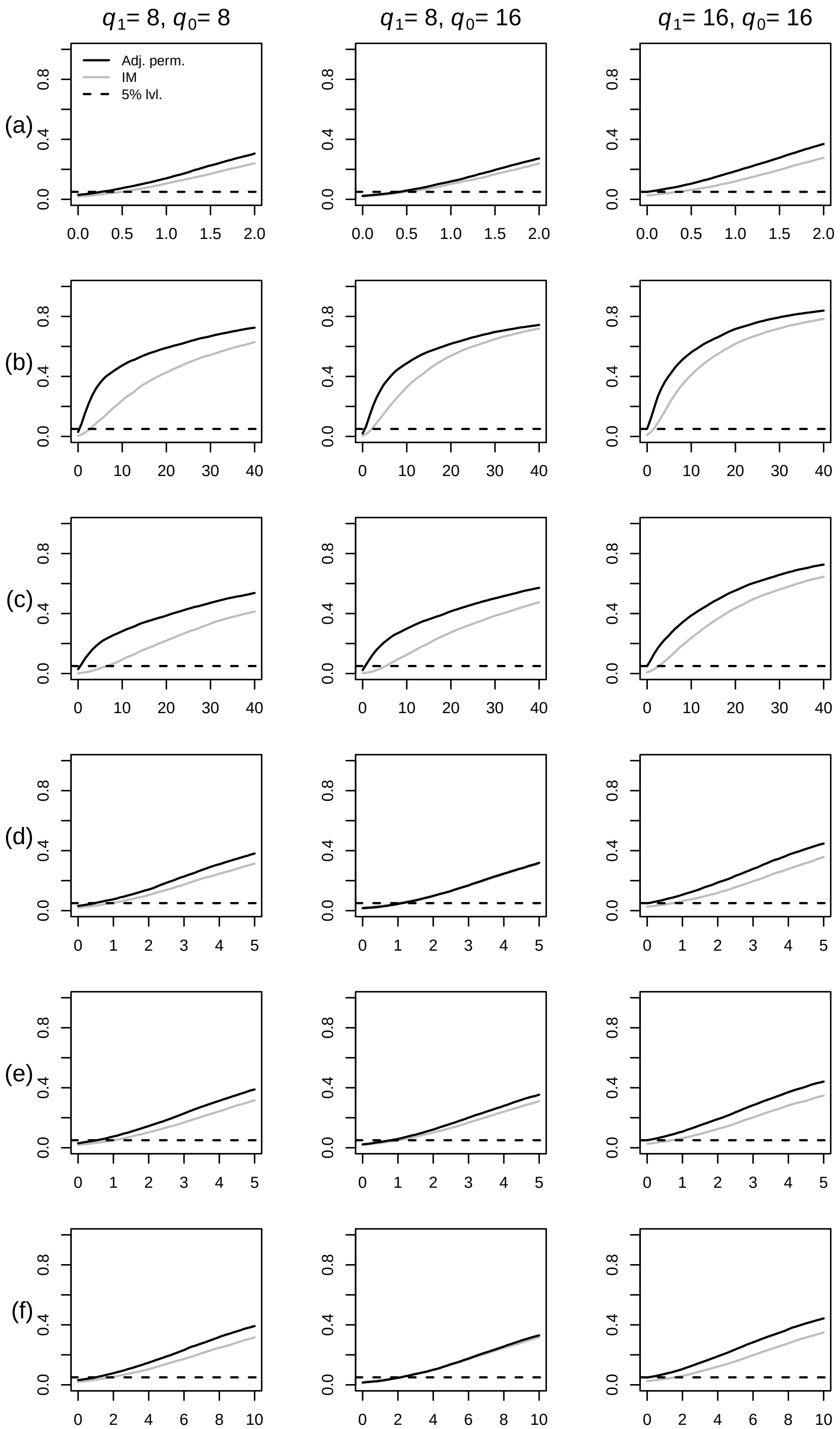

FIGURE 3. Rejection frequencies of the adjusted permutation test (black lines) and the Ibragimov-Müller test (IM, grey) as in Figure 2 but with Cauchy distributions.

time invariant, Ibragimov and Müller assume that the data are suitably approximated by a scale mixture of normals and implement their test by comparing the January excess returns for 1963-1969 to the January excess returns for 1970-1979. They do not reject the null hypothesis of time invariance at the 5% level but reject at the 10% level. The adjusted permutation test does not reject at either significance level. □

Example C.3 (Modern management practices; Bloom et al. 2013). In this example, I reanalyze data form a randomized trial of Bloom et al. (2013). Their intervention provided five months of extensive management consulting from a large international consulting firm to eleven randomly selected Indian textile plants. A control group of six randomly selected plants received only one month of diagnostic consulting. The experiment ran from 2008 to 2011 and several key performance measures were collected before, during, and after the intervention. These measures include data on quality defects, inventory, output, and total factor productivity. Here I focus on output because it is the only measure that has data for all 17 firms available. For the effect on output in their main results in their Table II, Bloom et al. (2013) run a regression of the log of picks (one pick is a single rotation of a weaving shuttle) on a treatment dummy, time fixed effects, and firm fixed effects. They find a 9% increase in output as a result of the intervention.

Bloom et al. (2013) use, among other methods, the Ibragimov and Müller (2016) test to conduct inference. The adjusted permutation test also applies and can be computed as outlined in Examples 3.1 and 3.3. Both the Ibragimov-Müller and the adjusted permutation test find a significant positive effect on log output at the 5% level but not at the 1% level, which confirms that the results of Bloom et al. (2013) remain valid even if methods designed for a small number of arbitrarily heterogeneous clusters are used. □

Appendix D. Proofs

Proof of Theorem 2.1 and Corollary 2.2. Denote the distribution function of an arbitrary random variable Y by F_Y. We have $T(X) > T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})$ if and only if $T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G})$. Because the test statistic is location invariant, assume without loss of generality that $\mu = 0$. Denote by $X_{(1)}, X_{(2)}, \dots, X_{(q)}$ the order statistics of X. Then $T^{(|\mathfrak{G}|)}(X, \mathfrak{G}) = q_1^{-1} \sum_{k=1}^{q_1} X_{(k+q_0)} - q_0^{-1} \sum_{k=1}^{q_0} X_{(k)}$. Because $T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G})$ and $\min\{X_1, \dots, X_{q_1}\} < \max\{X_{q_1+1}, \dots, X_q\}$ cannot be true at the same time and $\min\{X_1, \dots, X_{q_1}\} > \max\{X_{q_1+1}, \dots, X_q\}$ implies $T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G})$, it follows that

$P(T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G}))$ equals

$$P\big(\min\{X_1, \dots, X_{q_1}\} > \max\{X_{q_1+1}, \dots, X_q\}\big) \\ + P\big(T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G}), \min\{X_1, \dots, X_{q_1}\} = \max\{X_{q_1+1}, \dots, X_q\}\big).$$

Suppose $X_k = S_k Z_k$, $1 \leqslant k \leqslant q$, where the S_k is nonzero with probability one and the Z_k has a continuous distribution. The second line of the preceding display must then be zero conditional on $S = (S_1 \dots, S_q)$ and the same must therefore hold unconditionally. The first line conditional on $S_1 = \sigma_1, \dots, S_q = \sigma_q$ for fixed scales $\sigma_1, \dots, \sigma_q$ is, by independence, equivalent to the statement $P(\min\{X_1, \dots, X_{q_1}\} > \max\{X_{q_1+1}, \dots, X_q\})$ with $X_k = \sigma_k Z_k$ for $1 \leqslant k \leqslant q$. In the following, I will therefore work with $X_k = \sigma_k Z_k$ first and return to the unconditional case later.

Let $V = \max\{X_1, \dots, X_{q_1}\}$ and $W = \max\{X_{q_1+1}, \dots, X_q\}$. Symmetry of $X_1, \dots, X_{q_1}$ and independence of V and W imply

$$P\big(\min\{X_1, \dots, X_{q_1}\} > W\big) = P\big(\min\{-X_1, \dots, -X_{q_1}\} > W\big) = P(V + W < 0).$$

Suppose $q_1 < q_0$. The two maxima V and W must satisfy

$$P(V + W < 0) = P\left(\bigcap_{k=1}^{q_1} \bigcap_{l=1}^{q_0} \{X_k + X_{l+q_1} < 0\}\right) \leqslant P\left(\bigcap_{k=1}^{q_1} \{X_k + X_{k+q_1} < 0\}\right).$$

Define $Y_k = X_k + X_{k+q_1}$. Note that the Y_k are independent across $1 \leqslant k \leqslant q_1$ and symmetric because $P(X_k + X_{k+q_1} \leqslant y) = P(-X_k - X_{k+q_1} \leqslant y) = P(-Y_k \leqslant y)$. The right-hand side of the preceding display then equals $P(\max\{Y_1, \dots, Y_{q_1}\} < 0) = F_Y(0)^{q_1}$. Conclude from symmetry that $P(V + W < 0) \leqslant 0.5^{q_1}$. Repeat the argument with $q_1 > q_0$ to obtain

$$P(V + W < 0) \leqslant \max\{0.5^{q_1}, 0.5^{q_0}\} = 2^{\min\{q_1, q_0\}},$$

as desired. To see that this bound is tight, assume first that $q_1 \geqslant q_0$. Choose $\sigma_1 = \cdots = \sigma_{q_1} = 1$, $\sigma = \sigma_{q_1+1} = \cdots = \sigma_q$, and let $U = \max\{Z_{1+q_1}, \dots, Z_q\}$. Then $P(V + W < 0) = \mathrm{E}F_V(-\sigma U)$. If $U > 0$, then $F_V(-\sigma U) \to 0$ almost surely as $\sigma \to \infty$. If $U < 0$, then $F_V(-\sigma U) \to 1$ almost surely as $\sigma \to \infty$. Conclude from dominated convergence that $\mathrm{E}F_V(-\sigma U) \to P(U < 0) = 0.5^{q_0}$. If $q_1 < q_0$, switch V and W. This proves the theorem.

For the corollary, return to $X_k = S_k Z_k$ and redefine V, W accordingly. It is still true that $P(V + W \mid S) \leqslant 1/2^{\min\{q_1, q_0\}}$ almost surely and therefore $P(T(X) > T^{(|\mathfrak{G}|-1)}(X, \mathfrak{G})) \leqslant 1/2^{\min\{q_1, q_0\}}$, as required for the corollary. □

Define $V = \max\{S_1 Z_1, \dots, S_{q_1} Z_{q_1}\}$ and $W = \max\{S_{q_1+1} Z_{q_1+1}, \dots, S_q Z_q\}$. Let $w \mapsto F_W(w \mid S)$ be the distribution function of W conditional on S.

Lemma D.1. *Suppose $X = (X_1, \dots, X_q)$ with $X_k = \mu + \delta 1\{k \leqslant q_1\} + S_k Z_k$, $1 \leqslant k \leqslant q$, where the $Z_1, \dots, Z_q$ are iid copies of a random variable Z with continuous distribution function and $S = (S_1, \dots, S_q)$ is a random vector independent of $Z_1, \dots, Z_q$ with $P(S_k > 0) = 1$ for $1 \leqslant k \leqslant q$. If Z and $-Z$ have the same distribution, then*

$$\inf_{\mu \in \mathbb{R}} P\big(T(X) > T^{\bar{\alpha}}(X, \mathfrak{G})\big) \geqslant \mathrm{E} \int_0^1 \prod_{1 \leqslant j \leqslant q_1} F_Z\left(\frac{\delta - F_W^{-1}(t \mid S)}{S_j}\right) dt.$$

The right-hand side converges to 1 as $\delta \to \infty$.

Proof of Theorem D.1. This proof is similar to the proof of Theorem 2.2. As before, consider $T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G})$ and assume without loss of generality the case $\mu = 0$ so that $\min\{X_1, \dots, X_{q_0}\}$ has the same distribution as $\delta - V$. Because $T^{(|\mathfrak{G}|)}(X, \mathfrak{G}) = q_1^{-1} \sum_{k=1}^{q_1} X_{(k+q_0)} - q_0^{-1} \sum_{k=1}^{q_0} X_{(k)}$, continuity implies

$$P\big(T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G})\big) = P(V + W < \delta). \tag{D.1}$$

Independence of V and W conditional on S and continuity imply that there is an independent standard uniform U such that the preceding display equals

$$\mathrm{E} P\big(V < \delta - F_W^{-1}(U \mid S) \mid S\big) = \mathrm{E} \int_0^1 P\big(V < \delta - F_W^{-1}(t \mid S) \mid S\big) dt,$$

where the equality follows from Tonelli's theorem. By independence, distribution function of V conditional on S is $v \mapsto \prod_{1 \leqslant j \leqslant q_1} F_Z(v/S_j)$. The first result now follows because $P(T(X) > T^{\bar{\alpha}}(X, \mathfrak{G})) \geqslant P(T(X) = T^{(|\mathfrak{G}|)}(X, \mathfrak{G}))$. The second result follows from (D.1) as $\delta \to \infty$. □

Proof of Theorem A.1. This follows immediately from Lemma D.1 by letting $S = (\sigma_1, \dots, \sigma_q)$ with probability one and $F_Z = \Phi$. □

Proof of Proposition A.2. Following Canay et al. (2017), I only have to show that for any two distinct $g, g' \in \mathfrak{G}$, either $T(gx) = T(g'x)$ for all $x \in \mathbb{R}^q$ or $P(T(gX) \neq T(g'X)) = 1$. Let $w_{g(k)} = q_1^{-1} 1\{g(k) \leqslant q_1\} - q_0^{-1} 1\{g(k) > q_1\}$ and notice that $g \neq g'$ implies that $w_{g(k)} \neq w_{g'(k)}$ for at least two $k', k'' \in \{1, \dots, q\}$. By the pigeonhole principle, $X_{k'}$ or $X_{k''}$ must be continuously distributed. Then $T(gX) - T(g'X) = \sum_{k=1}^{q} (w_{g(k)} - w_{g'(k)}) X_k$ is continuously distributed by independence and therefore $P(T(gX) - T(g'X) = 0) = 0$. □

Proof of Proposition A.3. All limits are as $m \to \infty$. Let $\mathfrak{G}_m = \{G_1, \ldots, G_m\}$ be a collection of m draws from the uniform distribution on $\mathfrak{G}$, in which case $\mathrm{E}(\hat{p}(X, \mathfrak{G}_m) \mid X) = \hat{p}(X, \mathfrak{G})$. For almost every realization of X, the central limit theorem implies that $\sqrt{m}(\hat{p}(X, \mathfrak{G}_m) - \hat{p}(X, \mathfrak{G}))$ converges to mean-zero normal with variance $\hat{p}(X, \mathfrak{G})(1 - \hat{p}(X, \mathfrak{G}))$. Because $\hat{p}(X, \mathfrak{G}) \geqslant 1/|\mathfrak{G}|$, this variance can only be zero if $\hat{p}(X, \mathfrak{G}) = 1$. This occurs if and only if $T(gX) \geqslant T(X)$ for all $g \in \mathfrak{G}$, which also implies $\hat{p}(X, \mathfrak{G}_m) = 1$ for such X.

By the equivalence of p-values and critical values, $T(X) > T^p(X, \mathfrak{G}_m)$ if and only if $\hat{p}(X, \mathfrak{G}_m) \leqslant p$ and therefore

$$P\big(T(X) > T^p(X, \mathfrak{G}_m)|X\big) = P\Big(\sqrt{m}\big(\hat{p}(X, \mathfrak{G}_m) - \hat{p}(X, \mathfrak{G})\big) \leqslant \sqrt{m}\big(p - \hat{p}(X, \mathfrak{G})\big)|X\Big).$$

Since $P(\sqrt{m}(p(X, \mathfrak{G}_m) - p(X, \mathfrak{G})) \leqslant t \mid X)$ converges almost surely to a (possibly degenerate) normal distribution function, for every $\varepsilon > 0$ and almost every realization of X there is an M (possibly depending on ε and X) such that the limit of $P(\sqrt{m}(p(X, \mathfrak{G}_m) - p(X, \mathfrak{G})) \leqslant -M \mid X)$ is at most ε and $P(\sqrt{m}(p(X, \mathfrak{G}_m) - p(X, \mathfrak{G})) \leqslant M \mid X)$ is at least $1 - \varepsilon$. If $p > p(X, \mathfrak{G})$, then $\sqrt{m}(p - p(X, \mathfrak{G}))$ is eventually larger than every such M. If $p < p(X, \mathfrak{G})$, then $\sqrt{m}(p - p(X, \mathfrak{G}))$ is eventually smaller than $-M$. If $p = p(X, \mathfrak{G})$, which cannot occur if $p(X, \mathfrak{G}) = 1$, the preceding display converges almost surely to 0.5. Conclude that the preceding display converges almost surely to $1\{p(X, \mathfrak{G}) < p\} + 1\{p(X, \mathfrak{G}) = p\}/2$. The dominated convergence theorem then implies

$$P\big(T(X) > T^p(X, \mathfrak{G}_m)\big) \to P\big(\hat{p}(X, \mathfrak{G}) < p\big) + 0.5P\big(\hat{p}(X, \mathfrak{G}) = p\big).$$

The right hand side is equal to $P(\hat{p}(X, \mathfrak{G}) \leqslant p)$ if $P(\hat{p}(X, \mathfrak{G}) = p) = 0$, which is the case if $p|\mathfrak{G}|$ is not an integer because infinitesimal changes in p cannot change $P(\hat{p}(X, \mathfrak{G}) \leqslant p)$. If $P(\hat{p}(X, \mathfrak{G}) = p)$ is nonzero, then the preceding display is smaller than $P(\hat{p}(X, \mathfrak{G}) \leqslant p)$.

If $\mathfrak{G}'_m = \{id, G_2, \ldots, G_m\}$ then, both unconditionally and conditional on X,

$$\sqrt{m}\big(\hat{p}(X, \mathfrak{G}'_m) - \hat{p}(X, \mathfrak{G}_m)\big) = \frac{1 - 1\{T(G_1X) \geqslant T(X)\}}{\sqrt{m}} \xrightarrow{\mathrm{P}} 0.$$

The proof now follows from the arguments for $\mathfrak{G}_m$. □

Proof of Theorem 3.2. Suppose $\theta_1 = \theta_0$. Let 1_q denote a q-vector of ones and $X = (X_1, \ldots, X_q) \sim N(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)(\theta_0))$. Notice that $T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})$ if and only if

$$T\big(\sqrt{n}(\hat{\theta}_n - \theta_0 1_q)\big) > T^{\bar{\alpha}}\big(\sqrt{n}(\hat{\theta}_n - \theta_0 1_q), \mathfrak{G}\big).$$

Hence, it suffices to prove the result with $X_n = \sqrt{n}(\hat{\theta}_n - \theta_0 1_q)$ in place of $\hat{\theta}_n$. Because $X_n \rightsquigarrow X$, the desired result for $\theta_1 = \theta_0$ follows from Proposition A.2 and Theorem 2.1.

Suppose $\theta_1 = \theta_0 + \delta/\sqrt{n}$. Let $X_n = \sqrt{n}(\hat{\theta}_{n,k} - \theta_{1\{k\leqslant q_1\}})_{1\leqslant k\leqslant q}$ and $\Delta = (\delta 1\{k \leqslant q_1\})_{1\leqslant k\leqslant q}$. Then $X_n + \Delta \rightsquigarrow X + \Delta$ by the assumed continuity and the Slutsky lemma. By construction, $T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})$ is equivalent to $T(X_n + \Delta) > T^{\bar{\alpha}}(X_n + \Delta, \mathfrak{G})$. Proposition A.2 then implies

$$P\big(T(X_n + \Delta) > T^{\bar{\alpha}}(X_n + \Delta, \mathfrak{G})\big) \to P\big(T(X + \Delta) > T^{\bar{\alpha}}(X + \Delta, \mathfrak{G})\big).$$

Now apply the lower bound developed in Theorem A.1 to the right-hand side.

Suppose $\theta_1 = \theta_0 + \delta$. Let $\Delta_n = \sqrt{n}(\delta 1\{k \leqslant q_1\})_{1\leqslant k\leqslant q}$ so that $T(\hat{\theta}_n) \leqslant T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})$ is equivalent to $T(X_n) \leqslant T^{\bar{\alpha}}(X_n + \Delta_n, \mathfrak{G}) - T(\Delta_n)$. For a large $M > 0$, the probability that the latter event occurs is bounded above by

$$P\big(T(X_n) \leqslant -M\big) + P\big(T^{\bar{\alpha}}(X_n + \Delta_n, \mathfrak{G}) - T(\Delta_n) > -M\big). \tag{D.2}$$

The first term is bounded above by $\sup_n P(|T(X_n)| \geqslant M)$. This can be made as small as desired by choosing M large enough because the continuous mapping theorem implies that $T(X_n)$ is uniformly tight. By the properties of quantile functions, the second term in the preceding display is equal to

$$P\left(\sum_{g\in\mathfrak{G}} 1\big\{T(gX_n) + T(g\Delta_n) - T(\Delta_n) > -M\big\} > |\mathfrak{G}|\bar{\alpha}\right).$$

Because $T(g\Delta_n) - T(\Delta_n) = 0$ for $g \in \bar{\mathfrak{G}} = \{g \in \mathfrak{G} : \sum_{k=1}^{q_1} 1\{g(k) \leqslant q_1\} = q_1\}$ and $T(g\Delta_n) - T(\Delta_n) \leqslant -2\sqrt{n}\delta \to -\infty$ for $g \in \mathfrak{G} \setminus \bar{\mathfrak{G}}$, uniform tightness of $T(gX_n)$ for every $g \in \mathfrak{G}$ implies $P(1\big\{T(gX_n) + T(g\Delta_n) - T(\Delta_n) > -M\big\} = 1) = P(T(gX_n) + T(g\Delta_n) - T(\Delta_n) > -M)$ converges to 0 for every given M if $g \in \mathfrak{G} \setminus \bar{\mathfrak{G}}$. In addition, $T(gX_n) = T(X_n)$ for $g \in \bar{\mathfrak{G}}$ and hence the preceding display is within $o(1)$ of

$$P\big(|\bar{\mathfrak{G}}| 1\big\{T(X_n) > -M\big\} > |\mathfrak{G}|\bar{\alpha}\big),$$

which equals zero if $|\bar{\mathfrak{G}}| \leqslant |\mathfrak{G}|\bar{\alpha}$. Let $n \to \infty$ and then $M \to \infty$ in (D.2) to conclude $P(T(\hat{\theta}_n) > T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})) \to 1$ if $|\bar{\mathfrak{G}}| \leqslant |\mathfrak{G}|\bar{\alpha}$. Because $|\bar{\mathfrak{G}}| = q_1!(q - q_1)!$ and $|\mathfrak{G}| = q!$, this proves the result for $\bar{\alpha} \geqslant 1/\binom{q}{q_1}$. If $|\bar{\mathfrak{G}}| > |\mathfrak{G}|\bar{\alpha}$ or, equivalently, $\lceil \binom{q}{q_1}(1 - \bar{\alpha}) \rceil = \binom{q}{q_1}$, then $T^{\bar{\alpha}}(\hat{\theta}_n, \mathfrak{G})$ is the maximal order statistic and the power of the test is zero for any sample size. □

Appendix E. Numerical computation of $\bar{\alpha}$

This section provides two algorithms for the numerical computation of $\bar{\alpha}$ as in Table 1. For the algorithms, notice that it is of no loss of generality to assume that the standard deviations $\sigma_1, \dots, \sigma_q$ are restricted to the interval $(0, 1]$ because both sides of $T(X) > T^{(j)}(X, \mathfrak{G})$ can be divided by the largest standard deviation without altering the test decision.

Algorithm E.1 (q_1 and q_0 small). (1) Choose j, starting with $j = |\mathfrak{G}| - 2$.

(2) Draw a large number R of iid copies $V^1, \dots, V^R$ of a q-vector V with independent $\mathrm{Beta}(a, b)$ entries, e.g., $\mathrm{Beta}(0.1, 0.1)$.

(3) For each $1 \leqslant r \leqslant R$, draw a large number S of iid copies $X^1, \dots, X^S$ of $X \sim N(0, \operatorname{diag} V^r)$ and approximate $P(T(X) > T^{(j)}(X, \mathfrak{G}))$ by

$$\frac{1}{S} \sum_{s=1}^{S} 1\big\{T(X^s) > T^{(j)}(X^s, \mathfrak{G})\big\}.$$

(4) If there is an r in $1, \dots, R$ for which the number from step (3) is larger than α (or, alternatively, $\alpha + \eta$ for a small tolerance $\eta > 0$), let $j^* = j + 1$. If not, decrease j by 1 and restart at step (1).

(5) Define $\bar{\alpha} = 1 - j^* / \binom{q}{q_1}$.

Algorithm E.2 (q_1 or q_0 large). (1) Choose a large number m. Choose j, starting with $j = m - 2$.

(2) Draw a large number R of iid copies $V^1, \dots, V^R$ of a q-vector V with independent $\mathrm{Beta}(a, b)$ entries, e.g., $\mathrm{Beta}(0.1, 0.1)$.

(3) For each $1 \leqslant r \leqslant R$, draw a large number S of iid copies $X^1, \dots, X^S$ of $X \sim N(0, \operatorname{diag} V^r)$ and approximate $P(T(X) > T^{(j)}(X, \mathfrak{G}))$ by

$$\frac{1}{S} \sum_{s=1}^{S} 1\big\{T(X^s) > T^{(j)}(X^s, \mathfrak{G}_m)\big\}.$$

(4) If there is an r in $1, \dots, R$ for which the number from step (3) is larger than α (or, alternatively, $\alpha + \eta$ for a small tolerance $\eta > 0$), let $j^* = j + 1$. If not, decrease j by 1 and restart at step (1).

If $\binom{q}{q_1} < 1,500$, Table 1 uses three passes of Algorithm E.1 with $a = b = 0.1$ and $R = 3,000$. The first pass computes steps (1)-(3) with $S = 1,000$. The second pass takes, for each j, the top 1% values of $1 \leqslant r \leqslant R$ that led to the highest rejections and computes steps (3)-(5) with $S = 10,000$. If $\binom{q}{q_1} \geqslant 1,500$, Table 1 uses three passes of Algorithm E.2 with $a = b = 0.1$, $R = 3,000$, and $m = 1,500$. The first pass computes steps (1)-(3) with $S = 1,000$. The second pass takes, for each j, the top 1% values of $1 \leqslant r \leqslant R$ that led to the highest rejections and computes steps (3)-(5) with $S = 10,000$. Independently of the size of $\binom{q}{q_1}$, I compute a third pass with $S = 10,000$ where each entry of V takes on either 0.001 or 1 and add these as values of V^r to step (3) of each algorithm. This third pass and the Beta$(0.1, 0.1)$ distribution are used here because the highest rejection rates seem to occur near the boundaries of the parameter space where the Beta$(0.1, 0.1)$ distribution has most of its mass.

Additional references

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2013). Does management matter? evidence from india. *The Quarterly Journal of Economics 128*, 1–51.

Davidson, R. and J. G. MacKinnon (2000). Bootstrap tests: how many bootstraps? *Econometric Reviews 19*, 55–68.

Keim, D. B. (1983). Size related anomalies and stock return seasonality: further empirical evidence. *Journal of Financial Economics 12*, 13–32.

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

Słoczyński, T. (2018). A general weighted average representation of the ordinary and two-stage least squares estimands. Working paper, Department of Economics, Brandeis University.

Department of Economics, University of Michigan, 611 Tappan Ave, Ann Arbor, MI 48109, USA. Tel.: +1 (734) 764-2355. Fax: +1 (734) 764-2769

Email address: hagem@umich.edu

URL: umich.edu/~hagem