# INFERENCE ON QUANTILE PROCESSES WITH A FINITE NUMBER OF CLUSTERS

ANDREAS HAGEMANN

ABSTRACT. I introduce a generic method for inference on entire quantile and regression quantile processes in the presence of a finite number of large and arbitrarily heterogeneous clusters. The method asymptotically controls size by generating statistics that exhibit enough distributional symmetry such that randomization tests can be applied. The randomization test does not require ex-ante matching of clusters, is free of user-chosen parameters, and performs well at conventional significance levels with as few as five clusters. The method tests standard (non-sharp) hypotheses and can even be asymptotically similar in empirically relevant situations. The main focus of the paper is inference on quantile treatment effects but the method applies more broadly. Numerical and empirical examples are provided.

KEYWORDS: cluster-robust inference, quantiles, treatment effects, randomization inference, difference in differences

JEL CODES: C01, C21, C23

## 1. INTRODUCTION

Economic data often contain large clusters such as countries, regions, villages, or firms. Units within these clusters can be expected to influence one another or are influenced by the same political, environmental, sociological, or technical shocks. Several analytical and computer-intensive procedures such as the bootstrap are available to account for the presence of data clusters. These procedures generally achieve consistency by letting the number of clusters go to infinity. Numerical evidence

*Date*: June 15, 2023. University of Michigan Stephen M. Ross School of Business, 701 Tappan Ave, Ann Arbor, MI 48109, USA. Tel.: +1 (734) 764-2355. Fax: +1 (734) 764-2769. E-mail: hagem@umich.edu. I would like to thank two anonymous reviewers for helpful comments. All errors are my own.

by Bertrand, Duflo, and Mullainathan (2004), MacKinnon and Webb (2017), and others in the context of mean regression suggests that this type of asymptotic approximation often causes substantial size distortions when the number of clusters is small or the clusters are heterogenous. True null hypotheses are rejected far too often in both situations. Hagemann (2017) shows that this phenomenon is also present in quantile regression.

In this paper, I develop a generic method for inference on the entire quantile or regression quantile process in the presence of a finite number of large and arbitrarily heterogeneous clusters. The method, which I refer to as *cluster-randomized Kolmogorov-Smirnov* (*CRK*) test, asymptotically controls size by generating Kolmogorov-Smirnov statistics that exhibit enough distributional symmetry at the cluster level such that randomization tests (Fisher, 1935; Canay, Romano, and Shaikh, 2017) can be applied. The CRK test is not limited to the pure quantile regression setting and can be used in distributional difference-in-differences estimation (Callaway and Li, 2019) and related situations where quantile treatment effects are identified by between-cluster comparisons. The CRK test is free of user-chosen parameters, powerful against fixed and root-$n$ local alternatives, and performs well at conventional significance levels with as few as twelve clusters if parameters are identified between clusters. If parameters are identified within clusters, then even five clusters are sufficient for inference.

Quantile regression (QR), introduced by Koenker and Bassett (1978), is an important empirical tool because it can quantify the effect of a set of covariates on the entire conditional outcome distribution. An issue with QR in the presence of clustering is that estimates normalized by their asymptotic covariance kernel have standard normal marginal limit distributions but are no longer pivotal for any choice of weight matrix (Hagemann, 2017). Cluster-robust tests about the QR coefficient function therefore have asymptotic distributions that cannot be tabulated for inference about ranges of quantiles. Even if only individual quantiles are of interest, consistent covariance matrix estimation in large clusters is challenging. It requires knowledge of an explicit

ordering of the dependence structure within each cluster combined with a kernel and bandwidth choice to give distant observations less weight. Because time has a natural order, this weighting is easily done for time-dependent data but ordering data within states or villages may be difficult or impossible. The common empirical strategy of simply assuming that the clusters are small and numerous enough to satisfy a central limit theorem circumvents these issues but can lead to substantial size distortions with as few as 20 clusters (Hagemann, 2017). This remains true if a cluster-robust version of the bootstrap is used. Distortions can be especially severe if clusters differ greatly in their size and dependence structure.

I show that the CRK test is robust to each of these concerns: It performs well even when the number of clusters is small, the dependence varies from cluster to cluster, and the cluster sizes are heterogenous. The reason for this robustness is that the CRK test does not rely on clustered covariance matrices to rescale the estimates. I instead use randomization inference to generate random critical values that automatically scale to the data. There are no kernels, bandwidths, or spatio-temporal orderings of the data to choose. The test achieves consistency with a finite number of large but heterogeneous clusters under interpretable high-level conditions. Despite being based on randomization inference, the CRK test can perform standard (non-sharp) inference on entire quantile or regression quantile processes. Randomization is performed with a fixed set of estimates and does not require repeated estimation to obtain its critical values.

The randomization method underlying the CRK test was first used in the cluster context by Canay et al. (2017) as a way to perform inference on a finite-dimensional parameter with Student $t$ and Wald statistics in least squares regression. They do not consider inference on quantile functions or Kolmogorov-Smirnov statistics. Here, I considerably extend the scope of their method under explicit regularity conditions to allow for inference on the entire QR process and related objects. The proofs below

are fundamentally different from those of Canay et al. to account for the infinite-dimensional setting and do not rely on the Skorokhod almost-sure representation theorem. A practical issue with their method is that they require treated clusters to be matched ex-ante with an equal number of control clusters. Each match corresponds to a separate test and two researchers working with the same data can reach different conclusions based on which matches they choose. If there is not an equal number of treated and control clusters, then some clusters have to be combined or dropped in an ad-hoc manner. The CRK test sidesteps these issues completely and explicitly merges all potential tests into a single, uniquely determined test decision using results of Rüschendorf (1982).

Cluster-robust inference in linear regression models has a long history; recent surveys include Cameron and Miller (2015) and MacKinnon, Nielsen, and Webb (2022). Chen, Wei, and Parzen (2003), Wang and He (2007), Wang (2009), Parente and Santos Silva (2013), and Hagemann (2017) provide bootstrap and analytical methods for cluster-robust inference in QR models. Yoon and Galvao (2020) discuss the situation where clusters arise from correlation of individual units over time. All of these papers require the number of clusters to go to infinity for consistency. The CRK test differs from these papers because it is based on randomization inference and is consistent with a finite number of clusters.

Several papers show that pointwise inference with a fixed number of clusters is possible under a variety of conditions. Ibragimov and Müller (2010, 2016) use special properties of the Student $t$ statistic to perform inference on scale mixtures of normal random variables. 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. Hagemann (2019) adjusts permutation inference

for arbitrary heterogeneity at the cluster level but his bounds only apply to finite-dimensional objects. All of these methods can be used for inference at a single quantile but are not designed for simultaneous inference across ranges of quantiles. In contrast, the CRK test provides uniformly valid inference on the entire quantile process even if clusters are arbitrarily heterogeneous.

The remainder of the paper is organized as follows: Section 2 establishes new results on randomization inference on Gaussian processes. Section 3 uses these results to show consistency of the CRK test and gives specific examples where the test applies, including quantile difference-in-differences. Section 4 illustrates the finite sample behavior of the test in Monte Carlo experiments and an empirical example using Project STAR data. The appendix contains proofs.

I use the following notation and definitions: $1\{\cdot\}$ is the indicator function, cardinality of a set $A$ is $|A|$, the smallest integer greater than or equal $a$ is $\lceil a \rceil$, and the largest integer smaller than or equal $a$ is $\lfloor a \rfloor$. The minimum of $a$ and $b$ is denoted by $a \wedge b$. Limits are as $n \to \infty$ unless noted otherwise. Convergence in distribution under the parameter $\delta$ is denoted by $\overset{\delta}{\rightsquigarrow}$. A stochastic process $\{\xi(t) : t \in \mathcal{T}\}$ indexed by a set $\mathcal{T}$ is a collection of random variables $\xi(t) \colon \Omega \to \mathbb{R}$ defined on the same probability space $(\Omega, \mathcal{F}, P)$. Such a process is Gaussian if and only if $(\xi(t_1), \ldots, \xi(t_m))$ is multivariate normal for any finite collection of indices $t_1, \ldots, t_m \in \mathcal{T}$.

## 2. Randomization inference on Gaussian processes

In this section I study the size of randomization tests when the data come from heterogeneous Gaussian processes. I then analyze asymptotic size when a limiting experiment is characterized by such processes. The next section applies these generic results to the quantile setting.

I first introduce some notation for randomization tests that I will use throughout the paper. Let $u \mapsto X_j(u)$, $1 \leqslant j \leqslant q$, be independent mean-zero Gaussian processes indexed by $u \in \mathcal{U}$, where $\mathcal{U}$ is a compact subset of $(0, 1)$. Symmetry about zero implies

that $(X_j(u_1), \ldots, X_j(u_m))$ and $-(X_j(u_1), \ldots, X_j(u_m))$ are identically distributed. Because this is true for every finite collection of indices $u_1, \ldots, u_m \in \mathcal{U}$, $u \mapsto X_j(u)$ and $u \mapsto -X_j(u)$ have the same (finite-dimensional) distributions. Define $\mathcal{G} = \{1, -1\}^q$ as the $q$-dimensional product of $\{1, -1\}$ and, for $g = (g_1, \ldots, g_q) \in \mathcal{G}$, define $g \mapsto gx$ as the direct product $gx = (g_1 x_1, \ldots, g_q x_q)$ of $g$ and $x \in \mathbb{R}^q$. Independence and symmetry together imply that $u \mapsto X(u) = (X_1, \ldots, X_q)(u)$ and $u \mapsto gX(u)$ have the same distribution for every $g \in \mathcal{G}$ as long as $X$ has mean zero. The quantile and quantile-like processes discussed in the next section have this property under the null hypothesis. Deviations from the null cause non-zero means and therefore also asymmetry in $X$. The goal of this section is to develop a test of the null hypothesis of symmetry about zero,

$$H_0 \colon X(u) \sim gX(u), \qquad \text{all } g \in \mathcal{G}, \text{ all } u \in \mathcal{U}. \tag{2.1}$$

To test this hypothesis, I use the Kolmogorov-Smirnov-type statistic

$$T(X) = \sup_{u \in \mathcal{U}} \left( \frac{1}{q} \sum_{j=1}^{q} X_j(u) \right). \tag{2.2}$$

This statistic is large if symmetry is violated because the mean of the $X_j(u)$ is positive. I focus on one-sided tests to the right for simplicity but this is not restrictive. To test whether the mean is negative, simply use $-X$ instead of $X$ in the definition of $T$. These test statistics can be combined for two-sided tests. I explain this in detail at the end of Section 3.

Randomization inference uses distributional invariance to generate null distributions and critical values. In the present case, $X$ is distributionally invariant to all transformations $g$ contained in $\mathcal{G}$ because $X$ is symmetric. Let $T^{(1)}(X, \mathcal{G}) \leqslant T^{(2)}(X, \mathcal{G}) \leqslant \ldots \leqslant T^{(|\mathcal{G}|)}(X, \mathcal{G})$ be the $|\mathcal{G}| = 2^q$ ordered values of $T(gX)$ across $g \in \mathcal{G}$ and let

$$T^{1-\alpha}(X, \mathcal{G}) := T^{(\lceil (1-\alpha)|\mathcal{G}| \rceil)}(X, \mathcal{G}) \tag{2.3}$$

be the $1-\alpha$ quantile of these values. The randomization test function is then

$$\varphi_\alpha(X,\mathcal{G}) = 1\{T(X) > T^{1-\alpha}(X,\mathcal{G})\}. \tag{2.4}$$

If $\mathcal{U}$ is a finite set, distributional invariance under $H_0$ immediately implies $\mathrm{E}\varphi_\alpha(X,\mathcal{G}) = \mathrm{E}\varphi_\alpha(gX,\mathcal{G})$. By an argument due to Hoeffding (1952), the test function must satisfy $|\mathcal{G}|\alpha \geqslant \sum_{g\in\mathcal{G}} \varphi_\alpha(gX,\mathcal{G})$ and, after taking expectations on both sides, equality of the distributions yields $|\mathcal{G}|\alpha \geqslant \mathrm{E}\sum_{g\in\mathcal{G}} \varphi_\alpha(gX,\mathcal{G}) = \sum_{g\in\mathcal{G}} \mathrm{E}\varphi_\alpha(gX,\mathcal{G}) = |\mathcal{G}|\mathrm{E}\varphi_\alpha(X,\mathcal{G})$. This implies $\mathrm{E}\varphi_\alpha(X,\mathcal{G}) \leqslant \alpha$, which makes $T^{1-\alpha}(X,\mathcal{G})$ an $\alpha$-level critical value.

If $\mathcal{U}$ is a not finite, this argument does not immediately go through because (2.2) is a statement about possibly uncountably many $u \in \mathcal{U}$ but I have only established equivalence of the finite-dimensional distributions. However, as the following theorem shows, the conclusion that the test controls size holds nonetheless. The proof of the theorem extends Hoeffding's proof to stochastic processes with smooth sample paths by showing that (2.1) implies equality of the distributions of $(T(gX))_{g\in\mathcal{G}}$ and $(T(g\tilde{g}X))_{g\in\mathcal{G}}$ for every $\tilde{g} \in \mathcal{G}$. I prove that this is enough for Hoeffding's argument to go through as long as at least one of the processes has positive variance at every $u$.

**Theorem 2.1.** *Let $\{X_1(u)\colon u \in \mathcal{U}\}, \ldots, \{X_q(u)\colon u \in \mathcal{U}\}$ be independent mean-zero Gaussian processes with continuous sample paths indexed by the compact set $\mathcal{U} \subset (0,1)$ and let $u \mapsto X(u) := (X_1, \ldots, X_q)(u)$. If there is a $j \in \{1,\ldots,q\}$ such that $P(X_j(u) = 0) = 0$ for all $u \in \mathcal{U}$, then $\mathrm{E}\varphi_\alpha(X,\mathcal{G}) \leqslant \alpha$.*

*Remarks.* (i) If desired, the test decision can be randomized to construct an exact test. Take an independent variable $V$ with a uniform distribution on $[0,1]$ and the nonrandomized test function

$$\phi_\alpha(X,\mathcal{G}) = \begin{cases} 1 & \text{if } T(X) > T^{1-\alpha}(X,\mathcal{G}), \\ a(X) & \text{if } T(X) = T^{1-\alpha}(X,\mathcal{G}), \\ 0 & \text{if } T(X) < T^{1-\alpha}(X,\mathcal{G}), \end{cases} \tag{2.5}$$

where

$$a(X) = \frac{|\mathcal{G}|\alpha - |\{g \in \mathcal{G} : T(gX) > T^{1-\alpha}(X, \mathcal{G})|}{|\{g \in \mathcal{G} : T(gX) = T^{1-\alpha}(X, \mathcal{G})|}.$$

Using arguments of Hoeffding (1952), I show in the proof of Theorem 2.1 that the randomized test indeed satisfies $P(\phi_\alpha(X, \mathcal{G}) \geqslant V) = \alpha$. However, this type of test is uncommon in practice because rejecting the null if $\phi_\alpha(X, \mathcal{G}) \geqslant V$ bases the test decision on a single draw from the uniform distribution. A researcher could therefore draw until a desired conclusion was reached.

(ii) Similar arguments arise in the context of conformal prediction (Vovk, Gammerman, and Shafer, 2005) with exchangeable data. Such arguments do not apply here because $(T(gX))_{g\in\mathcal{G}}$ is generally not exchangeable. □

If $X$ is only an approximation in the sense that $X_n \rightsquigarrow X$ in $\ell^\infty(\mathcal{U})^q$, the space of bounded maps from $\mathcal{U}$ to $\mathbb{R}^q$, then the conclusions of the theorem still hold as long as the non-degeneracy conditions are strengthened. Here and in the following I tacitly assume that a process is indexed by a compact $\mathcal{U} \subset (0, 1)$ and that $\ell^\infty(\mathcal{U})^q$ is equipped with the Borel $\sigma$-field induced by the uniform norm topology.

**Theorem 2.2.** *If $X_n \rightsquigarrow X = \{(X_1, \ldots, X_q)(u) \colon u \in \mathcal{U}\}$, where the $\{X_j(u) \colon u \in \mathcal{U}\}$ are independent mean-zero Gaussian processes with continuous sample paths that satisfy $P(X_j(u) = -X_j(u')) = 0$ for all $u, u' \in \mathcal{U}$ and $1 \leqslant j \leqslant q$, then $\mathrm{E}\varphi_\alpha(X_n, \mathcal{G}) \to \mathrm{E}\varphi_\alpha(X, \mathcal{G})$.*

*Remarks.* (i) For the non-degeneracy assumption $P(X_j(u) = -X_j(u')) = 0$ to fail, a Gaussian process with uniformly continuous sample paths has to traverse, with certainty, from $X_j(u)$ to $X_j(u') = -X_j(u)$ while maintaining a positive variance along the entire path. The process would have to have identical variances at time $u$ and $u'$ but be perfectly negatively correlated at those times, which is impossible for Brownian bridges and related processes that typically arise in a quantile context. Still, such Gaussian processes exist and have to be ruled out.

(ii) The main difficulty of the proof of Theorem 2.2 is that the critical value $T^{1-\alpha}(X_n, \mathcal{G})$ does not settle down in the limit and is highly dependent on $T(X)$. The assumptions of Theorem 2.2 rule out degeneracies in the limit process that could lead to ties in the order statistics of $\{T(gX) : g \in \mathcal{G}\}$. This would put probability mass on the boundary of the set $\{T(X) > T^{1-\alpha}(X, \mathcal{G})\}$ and prevent application of the portmanteau lemma. Canay et al. (2017) use a delicate construction based on Skorokhod's representation theorem to account for the randomness in the limit. While these results could be extended from vectors to processes, I instead give a direct proof that I can also use to analyze the behavior of the test under both local and global alternatives when I discuss quantile processes in the next section.

(iii) Similar but less involved arguments show that if the supremum in the test statistic (2.4) is replaced by an integral over $\mathcal{U}$, then Theorems 2.1 and 2.2 continue to hold. However, this implicitly changes (2.1) to an hypothesis about the symmetry of $\int_{\mathcal{U}} X(u)du$. Other forms of the test statistic can also lead to valid tests, although the smoothness conditions described in parts (i) and (ii) of this remark may change. □

## 3. Inference on quantile processes with a finite number of clusters

This section gives high level conditions under which asymptotically valid inference on quantile processes and related objects can be performed even if the underlying data come from a fixed number of heterogeneous clusters.

### 3.1. **Inference when parameters are identified within clusters.**

Suppose data from $q$ large clusters (e.g., counties, regions, schools, firms, or stretches of time) are available. Throughout the paper, the number of clusters $q$ remains fixed and does not grow with the number of observations $n$. Observations are independent across clusters but dependent within clusters. Data from each cluster $1 \leqslant j \leqslant q$ separately identify a quantile or quantile-like scalar function $\delta : \mathcal{U} \to \mathbb{R}$. The $\delta$ can be estimated by $\hat{\delta}_j$ using data from only cluster $j$ such that a total of $q$ separate estimates $(\hat{\delta}_1, \ldots, \hat{\delta}_q) =: \hat{\delta}$ of $u \mapsto \delta(u)$ are available. The goal is to use randomization inference on a centered and

scaled version of $\hat{\delta}$ to develop tests of the null hypothesis

$$H_0\colon \delta(u) = \delta_0(u), \quad \text{all } u \in \mathcal{U}, \tag{3.1}$$

for some known function $\delta_0 : \mathcal{U} \to \mathbb{R}$. The following two examples describe simple but empirically relevant situations that fit this framework.

**Example 3.1 (Regression quantiles).** Suppose an outcome $Y_{i,j}$ of individual $i$ in cluster $j$ can be represented as $Y_{i,j} = X_{i,j}\delta(U_{i,j}) + Z_{i,j}'\beta_j(U_{i,j})$, where $u \mapsto X_{i,j}\delta(u) + Z_{i,j}'\beta_j(u)$ is strictly increasing in $u$ and $U_{i,j}$ is standard uniform conditional on covariates $(X_{i,j}, Z_{i,j})$. Here $X_{i,j}$ is the scalar covariate of interest and the $Z_{i,j}$ are additional controls. Monotonicity implies that the $u$-th conditional quantile of $Y_{i,j}$ is $X_{i,j}\delta(u) + Z_{i,j}'\beta_j(u)$ and linear QR as in Koenker and Bassett (1978) can provide estimates $(\hat{\delta}_j, \hat{\beta}_j)$ of $(\delta, \beta_j)$ for each cluster. Testing (3.1) with $\delta_0 \equiv 0$ tests whether $Y_{i,j}$ and $X_{i,j}$ are associated at any quantile after controlling for $Z_{i,j}$.

Several related models fit the framework of this example: (i) The $\beta_j$ can be constant across clusters. This does not impact the null hypothesis or the computation of the $\hat{\delta}_j$. (ii) The $\delta$ can vary by cluster in the QR model $Y_{i,j} = X_{i,j}\delta_j(U_{i,j}) + Z_{i,j}'\beta(U_{i,j})$ under the alternative. This has no impact on the computation of the $\delta_j$ and the null hypothesis simply becomes $H_0\colon \delta_1 = \cdots = \delta_q = \delta_0$. Identical $\delta_j$ are required only under the null hypothesis. (iii) If $\beta_j \equiv 0$ and $X_{i,j} \equiv 1$, then $u \mapsto \hat{\delta}(u)$ reduces to the $u$-th unconditional empirical quantile of $Y_{i,j}$. The null (3.1) can then be used to test whether $\delta$ has a specific functional form, e.g., a standard normal quantile function. □

**Example 3.2 (Quantile treatment effects).** Consider predetermined pairs $\{(j, j+q) : 1 \leqslant j \leqslant q\}$ of $2q$ groups. Suppose the first $q$ groups received treatment, indicated by $D_j = 1\{j \leqslant q\}$, and the remaining groups did not. Groups here could be manufacturing plants or villages. Treatment could be management consulting or introduction of a new technology. Denote treatment and control potential outcomes by $Y_j(1) \sim F_{Y(1)}$ and $Y_j(0) \sim F_{Y(0)}$, respectively. The observed outcome is $Y_j = D_jY_j(1) + (1 - D_j)Y_j(0)$.

For each group $j$, the experimenter observes identically distributed but potentially highly dependent copies $Y_{i,j}$ of $Y_j$ representing workers $i$ within group $j$. View each pair $(j, j+q)$ for $1 \leqslant j \leqslant q$ as a cluster and define the quantile treatment effect (QTE) as

$$u \mapsto \delta(u) = F^{-1}_{Y(1)}(u) - F^{-1}_{Y(0)}(u).$$

This QTE can be estimated as difference of the empirical quantiles

$$u \mapsto \hat{\delta}_j(u) = \hat{F}^{-1}_{Y_j}(u) - \hat{F}^{-1}_{Y_{j+q}}(u)$$

or, alternatively, as the coefficient on $D_j$ in a QR of $Y_{i,j}$ on a constant and $D_j$ using data only from cluster $j$. The situation where $\delta$ varies with $j$ is again included in the analysis as long as the null hypotheses is $\delta_1 = \cdots = \delta_q = \delta_0$. Estimation remains unchanged. I discuss the more complex scenario where the counterfactual $F_{Y(0)}$ has to be identified through difference-in-differences methods in Example 3.6 ahead. □

The $\hat{\delta}$ is neither limited to the estimators discussed in the preceding two examples nor does it need to have a special functional form. However, I assume that it can be approximated by a Gaussian process as in Theorem 2.2. Let $1_q$ be a $q$-vector of ones.

**Assumption 3.3.** *The stochastic process $\{\hat{\delta}(u) \colon u \in \mathcal{U}\}$ with $\hat{\delta}(u) \in \mathbb{R}^q$ satisfies*

$$X_n := \{\sqrt{n}(\hat{\delta} - \delta 1_q)(u) : u \in \mathcal{U}\} \overset{\delta}{\rightsquigarrow} X = \{(X_1, \ldots, X_q)(u) \colon u \in \mathcal{U}\}, \qquad (3.2)$$

*where the components of $X$ are independent mean-zero Gaussian processes with continuous sample paths, $P(X_j(u) = -X_j(u')) = 0$ for all $u, u' \in \mathcal{U}$ and $1 \leqslant j \leqslant q$.*

Examples of $X_n$ that can satisfy this assumption include unconditional quantile functions, coefficient functions in quantile regressions, quantile treatment effects, and other quantile-like objects. El Machkouri, Volný, and Wu (2013) present invariance principles and moment bounds that can be used to establish the convergence condition (3.2) under explicit weak dependence conditions.

I now connect the results from Section 2 about heterogeneous Gaussian processes to tests about $\delta$ under Assumption 3.3. The key property is that if $H_0$ in (3.1) does not hold, then $\sqrt{n}(\hat{\delta}-\delta_0 1_q) = X_n+\sqrt{n}(\delta-\delta_0)1_q$. The $X_n$ converges to a symmetric process but $\sqrt{n}(\delta-\delta_0)(u)$ grows without bound for some $u$, which makes the distribution of $\sqrt{n}(\hat{\delta}-\delta_0 1_q)$ highly asymmetric. Testing for symmetry using randomization inference is therefore informative about the hypothesis that $\delta = \delta_0$. I refer to a test that uses $\hat{\delta}-\delta_0 1_q$ in place of $X$ in test function (2.4) as the *cluster-randomized Kolmorogov-Smirnov* (*CRK*) test. From a practical perspective, the function $\delta_0$ is almost always $\delta_0 \equiv 0$. This tests the null of no effect at any quantile but more general hypotheses can be considered.

The test function $x \mapsto \varphi_\alpha(x, \mathcal{G})$ is invariant to scaling of $x$ by positive constants. If $H_0\colon \delta = \delta_0$ is true, then the CRK test satisfies

$$T(\hat{\delta}-\delta_0 1_q) > T^{1-\alpha}(\hat{\delta}-\delta_0 1_q, \mathcal{G})$$

if and only if $T(X_n) > T^{1-\alpha}(X_n, \mathcal{G})$. That the CRK test is an asymptotic $\alpha$-level test is then an immediate consequence of Theorems 2.1 and 2.2.

**Theorem 3.4 (Size).** *Suppose Assumption 3.3 holds. If $H_0\colon \delta = \delta_0$ is true, then* $\lim_{n\to\infty} \mathrm{E}\varphi_\alpha(\hat{\delta}-\delta_0 1_q, \mathcal{G}) \leqslant \alpha$.

*Remarks.* (i) The canonical limit of quantile and regression quantile processes such as those in Examples 3.1 and 3.2 is a scaled version of a $q$-dimensional Brownian bridge. That process easily satisfies the non-standard condition $P(X_j(u) = -X_j(u')) = 0$ imposed by Assumption 3.3.

(ii) The inequality in the theorem becomes an equality if $(1-\alpha)2^q$ is an integer. In that case, the test in the limit experiment is "similar," i.e., it has rejection probability exactly equal to $\alpha$ for all Gaussian processes that satisfy Assumption 3.3. The CRK test can therefore be asymptotically similar in some situations. If desired, the test decision can be randomized to make the CRK test similar in the limit for all $\alpha$. □

To analyze the power of the CRK test, I consider fixed alternatives $\delta(u) = \delta_0(u)+\lambda(u)$ with a positive function $u \mapsto \lambda(u)$, and local alternatives $\delta(u) = \delta_0(u) + \lambda(u)/\sqrt{n}$ converging to the maintained null hypothesis $H_0\colon \delta = \delta_0$. In the local case, $\delta_0$ is fixed but $\delta$ now depends on $n$ and the convergence (3.2) is under the sequence of functions $\delta = \delta_0 + \lambda/\sqrt{n}$. As the following results show, the CRK test has power against both types of alternatives.

**Theorem 3.5 (Global and local power).** *Suppose Assumption 3.3 holds and $\alpha \geqslant 1/2^q$. If $H_1\colon \delta = \delta_0 + \lambda$ is true with $\lambda\colon \mathcal{U} \to [0,\infty)$ continuous and $\sup_{u\in\mathcal{U}} \lambda(u) > 0$, then $\lim_{n\to\infty} \mathrm{E}\varphi_\alpha(\hat{\delta} - \delta_0 1_q, \mathcal{G}) = 1$. If $H_1\colon \delta = \delta_0 + \lambda/\sqrt{n}$ is true with $\sup_{u\in\mathcal{U}} \lambda(u) > \mathrm{E}\sup_{u\in\mathcal{U}} X_j(u)$, $1 \leqslant j \leqslant q$, then*

$$\lim_{n\to\infty} \mathrm{E}\varphi_\alpha(\hat{\delta} - \delta_0 1_q, \mathcal{G}) \geqslant \prod_{j=1}^{q}\left(1 - e^{-[\sup \lambda(u) - \mathrm{E}\sup X_j(u)]^2/2\sup \mathrm{E}X_j^2(u)}\right) > 0,$$

*where the suprema in the exponent are over $u \in \mathcal{U}$.*

*Remarks.* (i) The lower bound used for the local power result comes from the Borell-Tsirelson-Ibragimov-Sudakov (Borell-TIS) inequality (see, e.g., Adler and Taylor, 2007, p. 50). For large $q$, the bound is relatively crude but for small $q$, the only crude part is the assumption that $\delta$ is moderately large when compared to $X$. This is reflected in the condition that $\sup_{u\in\mathcal{U}} \lambda(u) > \mathrm{E}\sup_{u\in\mathcal{U}} X_j(u)$ instead of $\sup_{u\in\mathcal{U}} \lambda(u) > 0$. The bound can be made arbitrarily close to 1 by choosing $\sup_{u\in\mathcal{U}} \lambda(u)$ large enough.

(ii) If $(1-\alpha)|\mathcal{G}| > |\mathcal{G}| - 1$, the power of the test is identically zero. In that case $T^{1-\alpha}(X,\mathcal{G}) = \max_{g\in\mathcal{G}} T(gX)$ and $T(X) > T^{1-\alpha}(X,\mathcal{G})$ becomes impossible because $T(X)$ is contained in $\{T(gX) : g \in \mathcal{G}\}$. I therefore I focus on the case $(1-\alpha)|\mathcal{G}| \leqslant |\mathcal{G}| - 1$, which is equivalent to $\alpha \geqslant 1/2^q$.

(iii) The test also has power against alternatives where $\lambda$ varies with the cluster index $j$ and at least some of the $\lambda_j$ are large. However, a precise statement without additional conditions on the relative sizes of the $\lambda_j$ is involved. I do not pursue this here to prevent notational clutter. □

3.2. **Inference when parameters are identified across clusters.** In applications, the treatment effect is often not identified from within a cluster but by comparisons across two clusters. This is the case, for example, if treatment is assigned at random at the cluster level or if identification comes from comparing changes in one cluster to changes in another cluster in a quasi-experimental context. In this situation, each individual pairing of a treated cluster $j$ with a control cluster $k$ is generally informative about the treatment effect of interest $\delta$ and each $(j,k)$ pair gives rise to an estimate $\hat{\delta}_{j,k}$ of $\delta$ that could be used in a CRK-type test. The following example illustrates this for difference-in-differences estimation of quantile treatment effects.

**Example 3.6 (Quantile difference in differences).** Let $\Delta Y_t(0) = Y_t(0) - Y_{t-1}(0)$ be time differences of untreated outcomes. Periods $t \in \{0, -1\}$ are pre-intervention periods and $t = 1$ is the post-intervention period; $Y_1(1)$ is a treated potential outcome and $Y_t$ are observed outcomes. Denote by $F_{Y|D=d}$ the distribution of a variable $Y$ conditional on the treatment indicator taking on the value $d \in \{0, 1\}$. Callaway and Li (2019) show that the distribution $F_{Y_1(0)|D=1}(y)$ of the untreated potential outcome of a treated observation at time $t = 1$ can be identified as

$$P\Big(F^{-1}_{\Delta Y_1|D=0}\Big(F_{\Delta Y_0|D=1}(\Delta Y_0)\Big) + F^{-1}_{Y_0|D=0}\Big(F_{Y_{-1}|D=1}(Y_{-1})\Big) \leqslant y \mid D = 1\Big) \tag{3.3}$$

as long as a distributional version of the standard parallel trends assumption and some additional stability and smoothness conditions hold. This identifies the quantile treatment on the treated (QTT) effect

$$u \mapsto \delta(u) = F^{-1}_{Y_1(1)|D=1}(u) - F^{-1}_{Y_1(0)|D=1}(u),$$

where $F^{-1}_{Y_1(1)|D=1}(u)$ can be estimated by the sample quantile $\hat{F}^{-1}_{Y_1|D=1}(u)$. To estimate the counterfactual quantile, Callaway and Li replace $P$ and every $F$ in (3.3) with sample equivalents. This yields the estimated QTT

$$u \mapsto \hat{F}^{-1}_{Y_1|D=1}(u) - \hat{F}^{-1}_{Y_1(0)|D=1}(u). \tag{3.4}$$

Callaway and Li show that $\sqrt{n}(\hat{F}^{-1}_{Y_1|D=1} - \hat{F}^{-1}_{Y_1(0)|D=1} - \delta)$ converges to a well-behaved Gaussian process under mild regularity conditions.

Suppose that data come from $q_1$ states that received treatment and $q_0$ states that did not. View a single state over time as a cluster. Then two clusters are enough to compute (3.4): $\hat{F}^{-1}_{Y_1|D=1}$ can be computed from a treated cluster $j$ and $\hat{F}^{-1}_{Y_1(0)|D=1}$ can be computed from $j$ and an untreated cluster $k$. Denote by $\hat{\delta}_{j,k}$ the QTT estimated in this fashion using only data from clusters $j$ and $k$. Each $(j,k)$ pair provides a valid estimate of $\delta$ and each $\hat{\delta}_{j,k}$ could potentially be used in a CRK-type test of the null hypothesis $H_0\colon \delta = \delta_0$. □

I again assume that centered and scaled $\hat{\delta}_{j,k}$ converge in distribution to non-degenerate Gaussian processes with smooth sample paths as in Assumption 3.3. I only adjust this condition for the fact that estimates are constructed from pairwise combination of clusters. Let $q_1$ be the number of treated clusters and let $q_0$ be the number of control clusters.

**Assumption 3.7.** *The process $\{\sqrt{n}(\hat{\delta}_{j,k} - \delta)(u) : u \in \mathcal{U}\}$ converges, jointly in $j$ and $k$, in distribution to mean-zero Gaussian processes $X_{j,k}$ with continuous sample paths that satisfy $P(X_{j,k}(u) = -X_{j,k}(u')) = 0$ for all $u, u' \in \mathcal{U}$, $1 \leqslant j \leqslant q_1$, and $1 \leqslant k \leqslant q_0$. If both $j \neq j'$ and $k \neq k'$, then $X_{j,k}$ and $X_{j',k'}$ are independent.*

A naïve test of $H_0\colon \delta \equiv \delta_0$ would now take $X_{n,j,k} := \sqrt{n}(\hat{\delta}_{j,k} - \delta_0)$ and generate randomization distributions from $\{X_{n,j,k} : 1 \leqslant j \leqslant q_1, 1 \leqslant k \leqslant q_0\}$ via sign changes. However, $X_{n,j,k}$ and $X_{n,j,k'}$ are dependent for any choice of $j, k, k'$ because $j$ is used twice. This remains true even in large samples and if the data from all $q_1 + q_0$ groups are independent. Dependence causes problems because $(X_{n,j,k}, X_{n,j,k'})$ and $(X_{n,j,k}, -X_{n,j,k'})$ generally do not have the same joint distribution even when $n \to \infty$. Invariance under transformations with $g$ therefore fails. This issue can be avoided if one works with a subset of $\{X_{n,j,k} : 1 \leqslant j \leqslant q_1, 1 \leqslant k \leqslant q_0\}$ that uses each $j$ and $k$ only once. While this solves the dependence issue, it introduces another problem: each of

the $q_1$ treatment groups now has to be paired with exactly one of the $q_0$ control groups. Unless these pairings are determined before the data are analyzed, two researchers working with the same data and methodology could arrive at different conclusions because they chose different pairings. To address this problem, I now develop a method that maintains invariance under sign changes but avoids any decisions on the part of the researcher.

I first introduce some notation. If $q_1 \leqslant q_0$, there are $q_0 \times (q_0 - 1) \times \cdots \times (q_0 - q_1 + 1)$ ways of choosing $q_1$ ordered elements out of $(1, \ldots, q_0)$. Identify each such choice with an $h$ and denote the collection of all $h$ by $\mathcal{H}$. The ordering within $\mathcal{H}$ will not affect the test decision. For each $h \in \mathcal{H}$, denote by

$$\hat{\delta}_{[h]} = (\hat{\delta}_{1,h(1)}, \hat{\delta}_{2,h(2)}, \ldots, \hat{\delta}_{q_1,h(q_1)}), \qquad q_1 \leqslant q_0, \tag{3.5}$$

the vector that matches the subset of control groups associated with the label $h = (h(1), \ldots, h(q_1))$ to the (unpermuted) treated groups. If there are more treated than control groups such that $q_1 > q_0$, permute treated groups instead and take $h$ as enumerating ways of choosing $q_0$ elements out of $(1, \ldots, q_1)$ to define

$$\hat{\delta}_{[h]} = (\hat{\delta}_{h(1),1}, \hat{\delta}_{h(2),2}, \ldots, \hat{\delta}_{h(q_0),q_0}), \qquad q_1 > q_0. \tag{3.6}$$

By construction, the entries of $\hat{\delta}_{[h]}$ are independent of one another but $\hat{\delta}_{[h]}$ and $\hat{\delta}_{[h']}$ for $h, h' \in \mathcal{H}$ are potentially highly dependent.

To address the issue that there are multiple ways of combining clusters, I use an adjustment based on the randomization $p$-value

$$p(X, \mathcal{G}) = \inf\{p \in (0,1) : T(X) > T^p(X, \mathcal{G})\} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} 1\{T(gX) \geqslant T(X)\}. \tag{3.7}$$

Testing with this $p$-value is equivalent to a test with a critical value because $T(X) > T^{1-\alpha}(X, \mathcal{G})$ if and only if $p(X, \mathcal{G}) \leqslant \alpha$. The multiple comparisons adjustment is based on an inequality of Rüschendorf (1982). It states that arbitrary, possibly dependent variables $U_h$ indexed by $h \in \mathcal{H}$ with the property that $P(U_h \leqslant u) \leqslant u$ for every

$u \in [0,1]$ satisfy

$$P\left(\frac{2}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} U_h \leqslant u\right) \leqslant u, \qquad \text{every } u \in [0,1]. \tag{3.8}$$

This specific form of the inequality is given in Vovk and Wang (2020). Here the indexing set $\mathcal{H}$ is arbitrary and does not need to be related to permutations. The only condition is that $H = |\mathcal{H}| \geqslant 2$. The randomization $p$-value $p(\hat{\delta}_{[h]} - \delta_0 1_{q_1 \wedge q_0}, \mathcal{G})$ for testing whether the treatment effect of interest equals $\delta_0$ can be expected to behave like the $U_h$ in (3.8) in a large enough sample. Combining $p$-values of the CRK test to reject the null if

$$\frac{2}{H}\sum_{h\in\mathcal{H}} p(\hat{\delta}_{[h]} - \delta_0 1_{q_1 \wedge q_0}, \mathcal{G}) \tag{3.9}$$

does not exceed $\alpha$ should then asymptotically control size. The following theorem confirms that this is indeed true.

**Theorem 3.8 (Size with combined $p$-values).** *Suppose Assumption 3.7 holds. If $\delta = \delta_0$, then*

$$\limsup_{n\to\infty} P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(\hat{\delta}_{[h]} - \delta_0 1_{q_1 \wedge q_0}, \mathcal{G}) \leqslant \alpha\right) \leqslant \alpha.$$

*Remarks.* (i) The theorem can be improved slightly if $\alpha|\mathcal{G}|H/2$ is not an integer. In that case, the limit superior in the theorem is a proper limit that equals $P((2/H)\sum_{h\in\mathcal{H}} p(X_{[h]}, \mathcal{G}) \leqslant \alpha)$, where $X_{[h]}$ is the weak limit of $\sqrt{n}(\hat{\delta}_{[h]} - \delta_0 1_{q_1 \wedge q_0})$. This is because the sum in the preceding display can vary discontinuously at certain values. The limit inferior is $P((2/H)\sum_{h\in\mathcal{H}} p(X_{[h]}, \mathcal{G}) < \alpha)$.

(ii) Results of Vovk and Wang (2020) suggest that other ways of combining $p$-values such as exp(1) times the geometric mean of the $p$-value instead of a twice the average $p$-value are likely to be applicable here as well. However, the proof of the theorem given here relies crucially on the properties of the Rüschendorf inequality. In the Monte Carlo experiments in the next section, I do not find evidence that other ways of combining $p$-values lead to better results. □

The price paid for not matching treated and control clusters before the analysis is lower relative power. When $p$-values are averaged, Rüschendorf's inequality essentially decreases $\alpha$ to $\alpha/2$ to control size. Meng (1993) shows that the constant 2 cannot be improved. Still, as I establish below, the test has power against global and local alternatives if $\alpha > 1/2^{q_1 \wedge q_0 - 1}$, which is slightly stronger than what is needed in Theorem 3.5. Compared to Theorem 3.5, I also do not state an explicit bound for the local power analysis because applying the Borel-TIS inequality to the averaged $p$-values directly yields only relatively crude results. I instead show that if the alternatives $\lambda/\sqrt{n}$ converging to the null hypothesis are scaled up by a constant $c$, the test can detect these alternatives in the limit experiment with arbitrary accuracy if $c$ is large enough, that is, if first $n \to \infty$ and then $c \to \infty$.

**Theorem 3.9 (Global and local power with combined $p$-values).** *Suppose Assumption 3.7 holds and $\alpha > 1/2^{q_1 \wedge q_0 - 1}$. If $H_1 \colon \delta = \delta_0 + \lambda$ with $\lambda \colon \mathcal{U} \to [0, \infty)$ continuous and $\sup_{u \in \mathcal{U}}(u) > 0$, then $\lim_{n \to \infty} P((2/H) \sum_{h \in \mathcal{H}} p(\hat{\delta}_{[h]} - \delta_0 1_{q_1 \wedge q_0}, \mathcal{G}) \leqslant \alpha) = 1$. If $H_1 \colon \delta = \delta_0 + c\lambda/\sqrt{n}$, then*

$$\lim_{c \to \infty} \liminf_{n \to \infty} P\left( \frac{2}{H} \sum_{h \in \mathcal{H}} p(\hat{\delta}_{[h]} - \delta_0 1_{q_1 \wedge q_0}, \mathcal{G}) \leqslant \alpha \right) = 1.$$

3.3. **Implementation.** I now turn to some practical aspects of the CRK test. I discuss (i) what to do if $\mathcal{G}$ is large, (ii) what to do if $\mathcal{H}$ is large, and (iii) how to implement the test with a step-by-step guide. First, $\mathcal{G}$ can be prohibitively large if the number of clusters is large. If computing the entire randomization distribution is too costly, then $\mathcal{G}$ can be approximated by a random sample $\mathcal{G}_m$ consisting of $m$ draws from $\mathcal{G}$ with replacement. This is often referred to as "stochastic approximation." The theorems presented in Sections 3.1 and 3.2 continue to hold if $\mathcal{G}_m$ is used in place of $\mathcal{G}$ as long as a limit superior or inferior as $m \to \infty$ is applied before $n \to \infty$. The order of limits is not restrictive because, in a given sample of size $n$, the number of draws can $m$ always be made as large as computationally feasible. Under stochastic approximation,

the statement in Theorem 3.4 becomes $\lim_{n\to\infty} \limsup_{m\to\infty} \mathrm{E}\varphi_\alpha(\hat{\delta} - \delta_0 1_q, \mathcal{G}_m) \leqslant \alpha$, whereas statements about power use a limit inferior. Limit superior and inferior are needed here because of potential discontinuities but can be replaced by regular limits for most values of $\alpha$. Theorems 3.4, 3.8, and 3.9 hold without additional conditions but the conditions of Theorem 3.5 have to be strengthened marginally to avoid a discontinuity at $\alpha = 1/2^q$.

**Proposition 3.10.** *Suppose $\mathcal{G}_m$ consists of $m$ iid draws from $\mathcal{G}$. If every instance of $\mathcal{G}$ is replaced by $\mathcal{G}_m$, then*

(i) *Theorem 3.4 holds if* $\lim_{n\to\infty}$ *is replaced by* $\lim_{n\to\infty} \limsup_{m\to\infty}$,

(ii) *Theorem 3.5 holds if every* $\lim_{n\to\infty}$ *is replaced by* $\lim_{n\to\infty} \liminf_{m\to\infty}$ *and* $\alpha > 1/2^q$,

(iii) *Theorem 3.8 holds if* $\limsup_{n\to\infty}$ *is replaced by* $\limsup_{n\to\infty} \limsup_{m\to\infty}$,

(iv) *Theorem 3.9 holds if* $\lim_{n\to\infty}$ *is replaced by* $\lim_{n\to\infty} \liminf_{m\to\infty}$ *and* $\liminf_{n\to\infty}$ *is replaced by* $\liminf_{n\to\infty} \liminf_{m\to\infty}$.

*If $\alpha \notin \{j/|G| : 1 \leqslant j \leqslant |G|\}$, then* $\liminf_{m\to\infty}$ *and* $\limsup_{m\to\infty}$ *can be replaced by* $\lim_{m\to\infty}$ *in* (i)-(iv).

Second, the number of elements of $\mathcal{H}$ can similarly be large if the number of clusters is large or if there is a large discrepancy between the number of treated and the number of control clusters. In that case one can again work with a random subset $\mathcal{I}$ of $\mathcal{H}$. The crucial difference to the preceding result is that both Theorems 3.8 and 3.9 continue to hold even if $\mathcal{I}$ consists of only a finite number of random draws. In fact, the result goes through for any $\mathcal{I}$ as long as $\mathcal{I}$ is independent of the data.

**Proposition 3.11.** *Let $\mathcal{I}$ with $|\mathcal{I}| \geqslant 2$ be a fixed or random subset of $\mathcal{H}$ independent of the data. Then Theorems 3.8 and 3.9 continue to hold if $\mathcal{H}$ is replaced by $\mathcal{I}$.*

Finally, the following two algorithms outline and summarize how to apply the CRK test in practice. By Theorems 3.4 and 3.8, the procedures provide an asymptotically

$\alpha$-level test in the presence of a finite number of large clusters that are arbitrarily heterogeneous. They are free of nuisance parameters and do not require any decisions on the part of the researcher. By Theorems 3.5 and 3.9, the tests are able to detect fixed and $1/\sqrt{n}$-local alternatives. The first algorithm describes the CRK test when the parameters are identified within clusters. The second algorithm describes the between-cluster case, which is needed for distributional difference in differences. The tests can be two-sided or one-sided in either direction.

**Algorithm 3.12 (CRK test for parameters identified within clusters).**

(1) Compute for each $j = 1, \dots, q$ and using only data from cluster $j$ an estimate $\hat{\delta}_j$ of a parameter of interest $\delta$. (See Examples 3.1 and 3.2.) Define $\hat{\delta} = (\hat{\delta}_1, \dots, \hat{\delta}_q)$.

(2) Compute $\mathcal{G}$, the set of all vectors of length $q$ with entries 1 or $-1$, or replace $\mathcal{G}$ with a large random sample $\mathcal{G}_m$ from $\mathcal{G}$ in the following.

(3) Reject the null hypothesis $H_0 \colon \delta(u) = \delta_0(u)$ for all $u$ (e.g., $\delta_0 \equiv 0$ tests for no effect of treatment) against

   (a) $\delta(u) > \delta_0(u)$ for some $u$ if $T(\hat{\delta} - \delta_0 1_q) > T^{1-\alpha}(\hat{\delta} - \delta_0 1_q, \mathcal{G})$ for a test with asymptotic level $\alpha$,

   (b) $\delta(u) < \delta_0(u)$ for some $u$ if $T(\hat{\delta} - \delta_0 1_q) < T^{\alpha}(\hat{\delta} - \delta_0 1_q, \mathcal{G})$ for a test with asymptotic level $\alpha$,

   (c) $\delta(u) \neq \delta_0(u)$ for some $u$ if (a) or (b) are true for a test with asymptotic level $2\alpha$,

   where $T$ is defined in (2.2) and $T^{1-\alpha}(\cdot, \mathcal{G})$ is the $\lceil (1-\alpha)|\mathcal{G}| \rceil$-th largest value of the randomization distribution of $T$, defined in (2.3).

**Algorithm 3.13 (CRK test for parameters identified between clusters).**

(1) Compute $\mathcal{H}$, as defined above (3.5), or replace $\mathcal{H}$ with a large subset $\mathcal{I}$ in the following.

(2) Compute $\mathcal{G}$, the set of all vectors of length $q$ with entries 1 or $-1$, or replace $\mathcal{G}$ with a large random sample $\mathcal{G}_m$ from $\mathcal{G}$ in the following.

(3) For each $h$, compute $\hat{\delta}_{[h]}$ from (3.5) if $q_1 \leqslant q_0$ or from (3.6) if $q_1 > q_0$. (See Example 3.6.) Use (3.7) and (3.9) to compute

$$\frac{2}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} p(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1, q_0\}}, \mathcal{G}) \leqslant \alpha \tag{3.10}$$

(4) Reject the null hypothesis $H_0 \colon \delta(u) = \delta_0(u)$ for all $u$ (e.g., $\delta_0 \equiv 0$ tests for no effect of treatment) against
   (a) $\delta(u) > \delta_0(u)$ for some $u$ if (3.10) is true for a test with asymptotic level $\alpha$,
   (b) $\delta(u) < \delta_0(u)$ for some $u$ if (3.10) is true when $\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1, q_0\}}$ is replaced by $-(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1, q_0\}})$ for a test with asymptotic level $\alpha$,
   (c) $\delta(u) \neq \delta_0(u)$ for some $u$ if (a) or (b) are true for a test with asymptotic level $2\alpha$.

In some contexts, Algorithm 3.12 can be used even if the parameter of interest is identified by comparisons between treated and untreated clusters. For this to work, the researcher has to merge each treated cluster with an untreated cluster into a single cluster to recover within-cluster identification. If the number of treated clusters and control clusters is equal, then every treated cluster can be matched with a control cluster according to some rule. If the number of clusters is not equal, then two or more clusters can be merged to force an equal number of treated and control clusters. The merged clusters can then be reinterpreted as clusters and Algorithm 3.12 can be applied to these new clusters. While this comes with a large number of decisions, it is a valid method for inference if these decisions are made before the data are analyzed. For example, when estimating quantile treatment effects, a pre-analysis plan can be put in place that prescribes how clusters that received treatment will be merged with clusters that did not receive treatment. This reduces the problem to the one described in Example 3.2.

The next section investigates the finite sample performance of Algorithms 3.12 and 3.13 in several sitations.

## 4. Numerical results

This section presents several Monte Carlo experiments to investigate the small-sample properties of the CRK test in comparison to other methods of inference. I discuss significance tests on quantile regression coefficient functions (Example 4.1), inference in experiments when parameters are identified between clusters (Example 4.2), and estimation of QTEs in Project STAR (Example 4.3). I test one-sided hypotheses to the right but the results apply more broadly.

**Example 4.1 (Regression quantiles, cont.).** In this example, I adapt an experiment of Hagemann (2017) and use the data generating process (DGP)

$$Y_{i,j,k} = U_{i,j,k} + U_{i,j,k} Z_{i,j,k},$$

where $U_{i,j,k} = \sqrt{\varrho} V_{j,k} + \sqrt{1-\varrho} W_{i,j,k}$ with $\varrho \in [0,1)$; $V_{j,k}$ and $W_{i,j,k}$ are standard normal, independent of one another, and independent across indices. This ensures that the $U_{i,j,k}$ are standard normal and, for a given $j,k$, any pair $U_{i,j,k}$ and $U_{i',j,k}$ has correlation $\varrho$. The $Z_{i,j,k}$ satisfy $Z_{i,j,k} = X_{i,j,k}^2/3$ with $X_{i,j,k}$ standard normal independent of $U_{i,j,k}$ to ensure that the $U_{i,j,k}Z_{i,j,k}$ have mean zero and variance one. Both $X_{i,j,k}$ and $U_{i,j,k}$ are independent across $j$ and $k$, and $X_{i,j,k}$ is also independent across $i$. I discard information on $k$ after data generation and drop the $k$ subscripts in the following because they are not assumed to be known. This induces a dependence structure where each cluster $j = 1, \dots, q$ consists of several (unknown) neighborhoods $k = 1, \dots, K$ where observations are dependent if they come from the same $k$ but are independent otherwise. If $K \to \infty$ and the size of the neighborhoods is fixed or grows slowly with $K$, then this dependence structure is compatible with Assumptions 3.3 and 3.7 because it generates the weak dependence needed for central limit theory. In the experiments ahead, I set $K$ to either 10 or 20 and draw the size of each neighborhood from the uniform distribution on $\{5, 6, \dots, 15\}$. The DGP in the preceding display

corresponds to the QR model

$$Q(u \mid X_{i,j}, Z_{i,j}) = \beta_0(u) + \beta_1(u)X_{i,j} + \beta_2(u)Z_{i,j} \tag{4.1}$$

with $\beta_1(u) \equiv 0$ and $\beta_0(u) = \Phi^{-1}(u) = \beta_2(u)$, where $\Phi$ is the standard normal distribution function. For the CRK test, I estimated (4.1) separately for each cluster, obtained $q$ estimates of $\beta_1$ and applied Algorithm 3.12 with 1,000 new draws from $\mathcal{G}$ for each Monte Carlo replication.

To the best of my knowledge, there are no other methods of inference designed specifically for quantile functions or Kolmogorov-Smirnov statistics in data with few large clusters. I therefore compare the CRK test to inference with the wild gradient bootstrap (Hagemann, 2017), a cluster-robust version of the bootstrap that requires the number of clusters $q \to \infty$ for consistency. The wild gradient bootstrap is the default option for cluster-robust inference in the `quantreg` package in `R`. I use the package default settings with Mammen bootstrap weights and 200 bootstrap simulations. Alternative analytical methods for cluster-robust inference in quantile regressions exist but can only perform pointwise inference because the QR process as $q \to \infty$ generally has an analytically intractable distribution. Hagemann (2017) shows that the wild gradient bootstrap can conduct uniform inference on quantile regression functions and that it outperforms other methods for pointwise inference in this context. However, Hagemann (2017) notes that size distortions can occur when fewer than 20 clusters are present. I therefore focus on this situation in the following.

Figure 1 shows the rejection frequencies of a true null hypothesis $H_0\colon \beta_1(u) = 0$ for all $u$ as a function of the number of clusters $q \in \{5, 6, \dots, 20\}$ for the wild gradient bootstrap (left) and the CRK test (right) at the 5% level (short-dashed line). The figure shows rejection frequencies in 5,000 Monte Carlo replications for each horizontal coordinate with (i) $K = 10$ neighborhoods per cluster with intra-neighborhood correlation $\varrho = .5$ (solid lines), (ii) $K = 20$ with $\varrho = .5$ (long-dashed), and (iii) $K = 10$ with $\varrho = .1$ (dotted). Both methods were faced with the same data

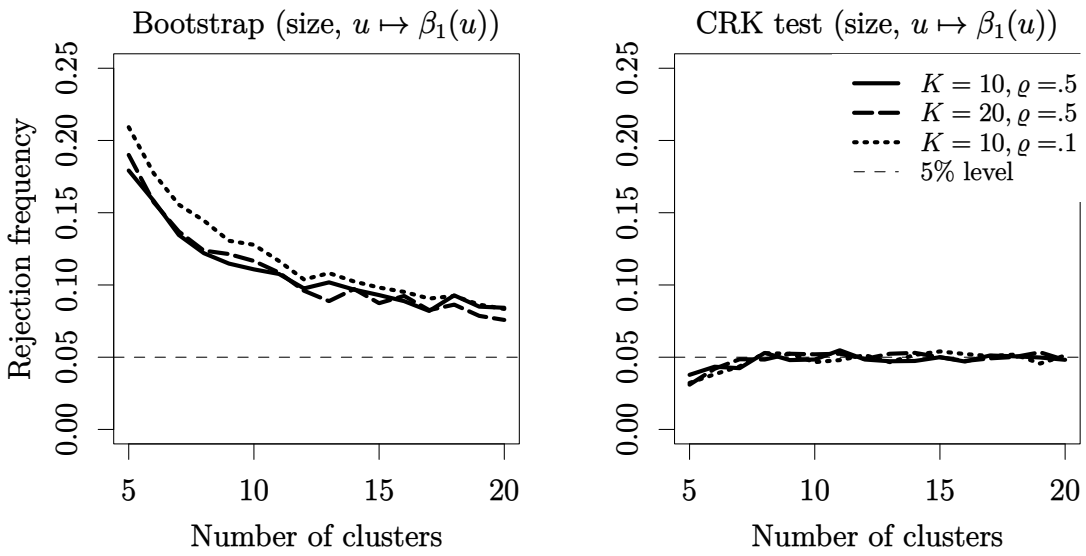


FIGURE 1. Rejection frequencies in Example 4.1 of a true null $H_0\colon \beta_1(u) = 0$ for all $u$ as a function of the number of clusters for the bootstrap (left) and the CRK test (right) with (i) $K = 10$ neighborhoods per cluster with intra-neighborhood correlation $\varrho = .5$ (solid lines), (ii) $K = 20$ with $\varrho = .5$ (long-dashed), and (iii) $K = 10$ with $\varrho = .1$ (dotted). Short-dashed line equals nominal level .05.

and I estimated $\beta_1$ at $u = .1, .2, \ldots, .9$ for both methods. As can be seen, the wild gradient bootstrap over-rejected mildly with 20 clusters but over-rejected substantially for smaller numbers of clusters. It exceeded a 10% rejection rate if only 12 clusters were available. With 5 clusters, the wild gradient bootstrap falsely discovered an effect in up to 20.9% of all cases ($K = 10, \varrho = .1$). In contrast, the CRK test rejected at or slightly below nominal level for all $q$ and all configurations of $K$ and $\varrho$.

I also experimented with a large number of alternative DGPs under the null. I considered (not shown) larger neighborhoods, different values of $\varrho$, different spatial dependence structures such as (spatial) autoregressive models, and different distributions for $X_{i,j,k}$. However, I found that these changes had little qualitative impact on the results described in the preceding paragraph or in Hagemann (2017). The wild gradient bootstrap generally performed very well but experienced size distortions with fewer than 20 clusters. The CRK test rejected at or slightly below nominal level in all situations I investigated.

I now turn to the behavior of the test under the alternative. I repeated the experiment but now tested the incorrect null hypothesis $H_0\colon \beta_2(u) = 0$ for all $u \in \mathcal{U}$. Figure 2 shows the rejection frequencies of this null against the alternative $H_1\colon \beta_2(u) > 0$ for some $u \in \mathcal{U}$, where $\mathcal{U}$ was either $(0, 1)$ (black) or $(.5, 1)$ (grey). The null hypothesis is false in both situations but the case where $\mathcal{U} = (0, 1)$ is more challenging because $\beta_2(u) < 0$ for all $u < .5$ so that estimates below the median provide evidence in the direction away from the alternative. I again considered (i) $K = 10$ neighborhoods per cluster with intra-neighborhood correlation $\varrho = .5$ (solid lines), (ii) $K = 20$ with $\varrho = .5$ (long-dashed), and (iii) $K = 10$ with $\varrho = .1$ (dotted). As could be expected, the bootstrap rejected a large fraction of null hypotheses mostly because it was unable to control the size of the test. However, it had high power when the number of clusters was above 20 and the size distortions disappeared (not shown). The CRK test had high power while maintaining size control even when the number of clusters was below 20. For example, at $q = 12$ it detected a deviation from the null between 22.5% ($K = 10, \varrho = .5, \mathcal{U} = (0, 1)$) and 84.26% ($K = 20, \varrho = .5, \mathcal{U} = (.5, 1)$) of all cases. More generally, additional clusters, lower intra-cluster dependence, and additional neighborhoods per cluster increased the power of the CRK test. □

**Example 4.2 (Quantile treatment effects, cont.).** For this experiment, I reuse the setup of Example 4.1 but replace the variable $X_{i,j,k}$ with a cluster-level treatment indicator $D_j$ that equals one if cluster $j$ received treatment and equals zero otherwise. I randomly assign $q_1 = \lfloor q/2 \rfloor$ clusters to treatment and $q_0 = \lceil q/2 \rceil$ to control. The coefficient of interest is $\delta$ in

$$Q(u \mid D_j) = \beta_0(u) + \delta(u) D_j + \beta_2(u) Z_{i,j}.$$

I do not assume that pairings are predetermined and therefore use the adjusted $p$-values of the CRK test from Algorithm 3.13. For each Monte Carlo replication, I drew a collection $\mathcal{I}$ with $|\mathcal{I}| = 50$ from $\mathcal{H}$ without replacement. The CRK test with unknown cluster parings requires $\alpha = .05 > 1/2^{q_1 \wedge q_0 - 1}$ to have power, which is satisfied here as

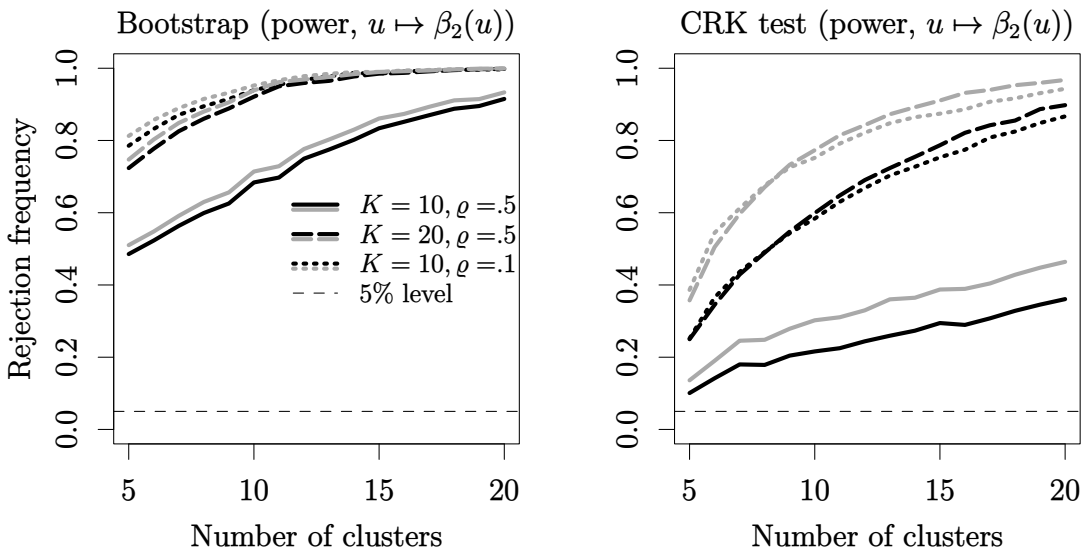


FIGURE 2. Rejection frequencies in Example 4.1 of false nulls $H_0\colon \beta_2(u) = 0$ for $u > .5$ (grey) and $H_0\colon \beta_2(u) = 0$ for all $u$ (black) as a function of the number of clusters for the bootstrap (left) and the CRK test (right) with (i) $K = 10$ neighborhoods per cluster with intra-neighborhood correlation $\varrho = .5$ (solid lines), (ii) $K = 20$ with $\varrho = .5$ (long-dashed), and (iii) $K = 10$ with $\varrho = .1$ (dotted).

long as $q \geqslant 12$. I therefore restrict $q$ to be between 12 and 20. All other parameters of the experiment are exactly as in Example 4.1.

The left panel of Figure 3 shows the rejection frequencies of a true null hypothesis $H_0\colon \delta(u) = 0$ for all $u$ in 5,000 Monte Carlo experiments per horizontal coordinate as $q$ increases. I again considered (i) $K = 10$ neighborhoods per cluster with intra-neighborhood correlation $\varrho = .5$ (solid lines), (ii) $K = 20$ with $\varrho = .5$ (long-dashed), and (iii) $K = 10$ with $\varrho = .1$ (dotted). As can be seen, adjusting the CRK test for unknown cluster pairings results in a markedly more conservative test relative to an unadjusted test from Figure 1. However, as the right panel of Figure 3 shows, this did not translate into poor power under the alternative. When I repeated the experiment with $\delta(u) \equiv .5$, the CRK test with identification across clusters had no problem detecting that neither $H_0\colon \delta(u)$ for all $u \in (0, 1)$ (black) nor $H_0\colon \delta(u)$ for $u > .5$ (grey) were true. Compared to Example 4.1, the alternative where $\mathcal{U} = (0, 1)$ rejects slightly more nulls because now every $u$ provides evidence against the null.

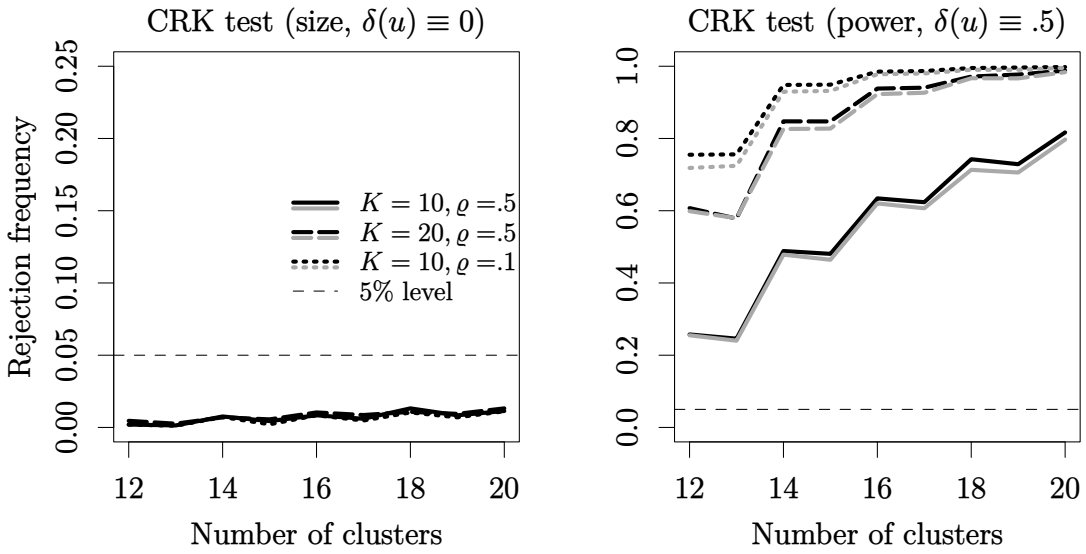


FIGURE 3. Rejection frequencies in Example 4.2 of a true null (left) $H_0\colon \delta(u) = 0$ for all $u$ and false nulls (right) $H_0\colon \delta(u) = 0$ for $u > .5$ (grey) and $H_0\colon \delta(u) = 0$ for all $u$ (black) as a function of the number of clusters for the CRK test when cluster pairings are not known with (i) $K = 10$ neighborhoods per cluster with intra-neighborhood correlation $\varrho = .5$ (solid lines), (ii) $K = 20$ with $\varrho = .5$ (long-dashed), and (iii) $K = 10$ with $\varrho = .1$ (dotted).

A noteworthy feature of the right panel of Figure 3 is the "zig-zag" pattern in the rejection frequencies. The reason for this pattern is the treatment assignment mechanism. If $q = 12$, then $q_1 = 6$ clusters receive treatment and $q_0 = 6$ do not. If $q = 13$, then again $6 = \lfloor 13/2 \rfloor$ clusters receive treatment but now $7 = \lceil 13/2 \rceil$ do not. Algorithm 3.13 uses a large number of potential pairings of treatment to control for inference but effectively reduces the number of clusters to $\min\{q_1, q_0\}$. In this experiment, inference with $6 + 7$ clusters is therefore effectively the same as inference with $6 + 6$ clusters, which explains the similar performance of the test at $q$ and $q - 1$ when $q$ is odd.

I also experimented with alternative methods for combining the $p$-values in Algorithm 3.13. For this, I repeated the experiment in the right panel of Figure 3 (not shown) but replaced the left-hand side of (3.10) with either the standard Bonferroni correction or the geometric average of the $p$-values times $\exp(1)$. The latter method

is due to Mattner (2012) and discussed in detail in Vovk and Wang (2020). At $q = 12$ with $K = 20$ and $\varrho = .5$, Bonferroni rejected the false null $H_0 \colon \delta(u) = 0$ for all $u$ in none of the 5,000 simulations, Mattner's method rejected in 39.44% of all simulations, and Algorithm 3.13 rejected in 60.76% of all simulations in the same data. At $q = 20$, Mattner's method and Algorithm 3.13 rejected in about 99% of all cases. Bonferroni improved substantially and rejected in 96% of all cases. I conducted a large number of additional experiments but the results remained the same. The Bonferroni method had by far the lowest power. Relative to Algorithm 3.13, Mattner's method had significantly lower power when the number of clusters was small but both alternatives caught up as this number increased. Neither Bonferroni nor Mattner's method improved the power of Algorithm 3.13 in any of my experiments.

Finally, I compare Algorithm 3.13 to the ideal situation where an equal number of treated and control clusters are paired through a pilot experiment or pre-analysis plan. For this, I repeated the experiment in Figure 3 with a single, randomly chosen pairing. At $q_1 = q_0 = 8$, $K = 10$, and $\varrho = .5$, a test with pre-specified pairs rejected a false $H_0 \colon \delta(u) = 0$ for all $u$ in 82.16% of all cases as opposed to the 63.40% achieved by Algorithm 3.13. In the same experiment with $q_1 = q_0 = 10$, the test with pre-specified pairs rejected 91.90% of all false nulls and Algorithm 3.13 rejected 81.70% of all false nulls. However, there are $8! = 40{,}320$ potential ways of paring $q_1 = 8$ treated clusters and $q_0 = 8$ control clusters. If $q_1 = q_0 = 10$, there are 3,628,800 ways. Each separate set of pairs could be potentially selected for the test. If a researcher did not pre-specify pairs and instead searched over potential pairs to discover a significant result, the test would quickly lose size control. At $q_1 = q_0 = 6$, $K = 10$, and $\varrho = .5$, a correct null hypothesis was erroneously rejected in 8.14% of all cases when the lowest $p$-value among three potential pairings was used. If the researcher chooses the lowest $p$-value among ten potential matches, then a non-existent effect showed up as statistically significant in 13.96% of all cases in a test with 5% nominal level. Algorithm 3.13 does not choose among these tests and completely avoids this loss of size control. □

**Example 4.3 (Placebo interventions in Project STAR).** In this example, I revisit a challenging placebo exercise of Hagemann (2017, Experiment 5.1) in data from the first year of the Tennessee *Student/Teacher Achievement Ratio* experiment, known as Project STAR. Details about the data can be found in Word et al. (1990) and Graham (2008). I only provide a brief summary.

In 1985, incoming kindergarten students in 79 project schools were randomly assigned to small classes (13-17 students) or regular-size classes (22-25 students) with or without a teacher's aide. Each of the project schools was required to have at least one of each kindergarten class type. The outcome is standardized student performance on the *Stanford Achievement Test* (SAT) in mathematics and reading administered at the end of the school year. The raw test scores are standardized as in Krueger (1999). He finds across several mean regression models that students in small classes perform about five percentage points better on average than students in regular classrooms. (Assigning teachers aides had no effect uniformly across specifications and I do not consider such classes in the following.) Jackson and Page (2013) and Hagemann (2017) document similar effects in quantile regressions but show that the effects are smaller for students near the bottom and the very top of the conditional outcome distribution and larger near the center of the distribution. For example, in the model

$$Q_{Y_{i,j}}(u \mid X_{i,j}) = \beta_0(u) + \delta(u) small_{i,j} + \beta_2(u)^T Z_{i,j} \tag{4.2}$$

where the treatment dummy *small* indicates whether the student was assigned to a small class and $Z$ contains school dummies, the effect of being in a small class relative to a regular class varies between 2.78 percentage points at the 10th percentile to 7.23 percentage points at the 60th percentile. Jackson and Page (2013) hypothesize that this heterogeneity could be attributed to varying levels of student motivation to take advantage of increased individual attention from a teacher.

For the placebo experiment, I removed all small classes from the sample and only kept the 16 schools that had two regular-size classes without aide. In each of these

TABLE 1. Rejection frequencies of $H_0\colon \delta(u) = 0$ for all $u$ in placebo interventions in Project STAR for the CRK test and the wild gradient bootstrap at 5% level

| | size | power | | | | | |
|---|---|---|---|---|---|---|---|
| | $\delta = 0$ | $\delta = 2$ | $\delta = 3$ | $\delta = 4$ | $\delta = 5$ | $\delta = 6$ | $\delta = 7$ |
| CRK test | .043 | .122 | .161 | .212 | .318 | .379 | .478 |
| Bootstrap | .091 | .233 | .316 | .428 | .580 | .691 | .814 |

16 schools, I then randomly assigned one of the regular-size classes the treatment indicator $small = 1$. This mimics the random assignment of class sizes within schools in the original sample, even though in this case no student actually attended a small class. I clustered at the classroom level and applied the CRK test as in Algorithm 3.12 by running 16 separate quantile regressions, one for each school, on a constant and *small* to get 16 separate estimates of $\delta$. The fixed effects as in (4.2) are not needed here because the constant can vary freely by school in these quantile regressions. Algorithm 3.12 applies here because each school in this experiment has only one class with $small = 1$ and one class with $small = 0$. (If multiple small classes per school were available, then Algorithm 3.13 could be used instead.) For the wild gradient bootstrap, I reran the QR in (4.2) in the placebo data and again clustered at the classroom level. For both methods, I tested at the 5% level the correct null hypothesis that $H_0\colon \delta(u) = 0$ jointly at $u \in \{.1, .2, \dots, .9\}$ against the alternative that $\delta$ is positive.

The rejection frequencies in 'size' column in Table 1 show the outcome of repeating the placebo assignment 1,000 times. As can be seen, the CRK test provided a nearly exact test but the bootstrap over-rejected somewhat. The over-rejection for the bootstrap here was documented by Hagemann (2017) and can be attributed to the very small number of clusters available in the placebo sample vis-à-vis the large number of clusters needed for the consistency of the bootstrap.

I also investigated power by increasing the percentile scores of all students in the randomly drawn small classes of the placebo experiment by $\delta \in \{2, 3, 4, 5, 6, 7\}$ percentage points. These increases are of the same or smaller magnitude as the estimated quantile treatment effects in the actual sample. Then I tested the incorrect

hypothesis $H_0\colon \beta_1(u) = 0$ for all $u$ with the same experimental setup as before. The results are shown in 'power' column of Table 1. As can be seen, the CRK test was able to reliably detect effects for moderate deviations from the null hypothesis. The wild gradient bootstrap rejected more often, but this was likely caused by its tendency to over-reject in this data set. □

## 5. Conclusion

I introduce a generic method for inference on quantile and regression quantile processes in the presence of a finite number of large and arbitrarily heterogeneous clusters. The method asymptotically controls size by generating statistics that exhibit enough distributional symmetry such that randomization tests can be applied. This randomization test can even be asymptotically similar in empirically relevant situations. The test does not require ex-ante matching of clusters, is free of user-chosen parameters, and performs well at conventional significance levels with as few as five clusters. The main focus on the paper is inference on quantile treatment effects and quantile difference in differences but the method applies more broadly. Numerical examples and an empirical application are provided.

## References

Adler, R. J. and J. E. Taylor (2007). *Random Fields and Geometry.* Springer, New York.

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.

Callaway, B. and T. Li (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics 10*, 1579–1618.

Cameron, A. C. and D. L. Miller (2015). A practitoner's guide to cluster robust inference. *Journal of Human Resources 50*, 317–372.

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.

Chen, L., L.-J. Wei, and M. I. Parzen (2003). Quantile regression for correlated observations. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*. Lecture Notes in Statistics. Springer, New York.

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.

Fisher, R. A. (1935). "The coefficient of racial likeness" and the future of craniometry. *Journal of the Royal Anthropological Institute of Great Britain and Ireland 66*, 57–63.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*, 643–660.

Hagemann, A. (2017). Cluster-robust bootstrap inference in quantile regression models. *Journal of the American Statistical Association 112*, 446–456.

Hagemann, A. (2019). Permutation inference with a finite number of heterogeneous clusters. University of Michigan working paper, `arXiv:1907.01049`.

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.

Jackson, E. and M. E. Page (2013). Estimating the distributional effects of education reforms: A look at Project STAR. *Economics of Education Review 32*, 92–103.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics 114*, 497–532.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2022). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, forthcoming.

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

Mattner, L. (2012). Combining individually valid and arbitrarily dependent P-variables. In *Abstract Book of the Tenth German Probability and Statistics Days*, pp. 104. Institut für Mathematik, Johannes Gutenberg-Universität Mainz.

Meng, X.-L. (1993). Posterior predictive $p$-values. *Annals of Statistics 22*, 1142–1160.

Parente, P. M. and J. M. Santos Silva (2013). Quantile regression with clustered data. University of Essex Department of Economics Discussion Paper No. 728.

Rüschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability 14*, 623–632.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer New York, NY.

Vovk, V. and R. Wang (2020). Combining p-values via averaging. *Biometrika 107*, 791–808.

Wang, H. (2009). Inference on quantile regression for heteroscedastic mixed models. *Statistica Sinica 19*, 1247–1261.

Wang, H. and X. He (2007). Detecting differential expressions in genechip microarray studies: a quantile approach. *Journal of American Statistical Association 102*, 104–112.

Word, E., J. Johnston, H. P. Bain, B. D. Fulton, C. M. Achilles, M. N. Lintz, J. Folger, and C. Breda (1990). The state of Tennessee's student/teacher achievement ratio (STAR) project: Technical report 1985-1990. Report, Tennessee State University, Center of Excellence for Research in Basic Skills.

Yoon, J. and A. F. Galvao (2020). Cluster robust covariance matrix estimation in panel quantile regression with individual fixed effects. *Quantitative Economics 11*, 579–608.

## Appendix A. Proofs

*Proof of Theorem 2.1.* Denote the inverse element of $g \in \mathcal{G}$ by $g^{-1}$ and the identity element by id. The proof uses the inverse $g^{-1}$ of $g \in \mathcal{G}$ to clarify its role in the argument. However, note that inverting $g$ is redundant for this particular $\mathcal{G}$ because it satisfies $g^{-1} = g$. I first argue that $(T(gX))_{g\in\mathcal{G}}$ satisfies $(T(gX))_{g\in\mathcal{G}} \sim (T(g\tilde{g}^{-1}X))_{g\in\mathcal{G}}$, where $\tilde{g}$ is an arbitrary element of $\mathcal{G}$. For this, I show that both quantities must have the same distribution at continuity points and that $(T(gX))_{g\in\mathcal{G}}$ has a continuous distribution. I then argue that $(T(gX))_{g\in\mathcal{G}} \sim (T(g\tilde{g}^{-1}X))_{g\in\mathcal{G}}$ is enough for Hoeffding's (1952) argument to go through.

Take a finite grid of points $\mathcal{U}_m := \{i/m : i = 0, 1, \dots, m\} \cap \mathcal{U}$. Then every $u \in \mathcal{U}$ is a limit of a sequence in $\mathcal{U}_m$. Let $x \mapsto T_m(x) = \sup_{u\in\mathcal{U}_m} \sum_{j=1}^q x_j(u)/q$. Uniform continuity implies $T_m(x) \to T(x)$ and $(T_m(gX))_{g\in\mathcal{G}} \to (T(gX))_{g\in\mathcal{G}}$ almost surely and therefore also $(T_m(gX))_{g\in\mathcal{G}} \rightsquigarrow (T(gX))_{g\in\mathcal{G}}$. Independence and $P(X_j(u) = 0) = 0$ ensure that $\sum_{j=1}^q X_j(u)/q$ has a continuous distribution at every $u$. Because $X$ is separable, $T(gX) = \sup_{u\in\mathcal{U}\cap\mathbb{Q}} \sum_{j=1}^q X_j(u)/q$, where $\mathbb{Q}$ are the rationals. Conclude that $(T(gX))_{g\in\mathcal{G}}$ has a continuous distribution because for arbitrary $t_g \in \mathbb{R}$,

$$P\bigcap_{g\in\mathcal{G}}\{T(gX) = t_g\} \leqslant P\left(\sup_{u\in\mathcal{U}\cap\mathbb{Q}} \frac{1}{q}\sum_{j=1}^q X_j(u) = t_{\mathrm{id}}\right) \leqslant \bigcup_{u\in\mathcal{U}\cap\mathbb{Q}} P\left(\frac{1}{q}\sum_{j=1}^q X_j(u) = t_{\mathrm{id}}\right)$$

and the extreme right-hand side equals zero. Finite-dimensional distributional invariance implies that $(T_m(gX))_{g\in\mathcal{G}}$ and $(T_m(g\tilde{g}^{-1}X))_{g\in\mathcal{G}}$ have the same distribution for every $\tilde{g} \in \mathcal{G}$. Because $(T_m(gX))_{g\in\mathcal{G}} \rightsquigarrow (T(gX))_{g\in\mathcal{G}}$, it must also be true that $(T_m(g\tilde{g}^{-1}X))_{g\in\mathcal{G}} \rightsquigarrow (T(gX))_{g\in\mathcal{G}}$ and $(T_m(g\tilde{g}^{-1}X))_{g\in\mathcal{G}} \rightsquigarrow (T(g\tilde{g}^{-1}X))_{g\in\mathcal{G}}$. Conclude from continuity that $(T(gX))_{g\in\mathcal{G}}$ and $(T(g\tilde{g}^{-1}X))_{g\in\mathcal{G}}$ have the same distribution for

every $\tilde{g} \in \mathcal{G}$. These two random vectors are of the form

$$(T(X), \dots, T(gX), \dots, T(\tilde{g}X), \dots) \sim (T(\tilde{g}^{-1}X), \dots, T(g\tilde{g}^{-1}X), \dots, T(\mathrm{id}\, X), \dots).$$

Because $T^{1-\alpha}(X, \mathcal{G}\tilde{g}^{-1}) = T^{1-\alpha}(X, \mathcal{G}) = T^{1-\alpha}(\tilde{g}X, \mathcal{G})$, this implies $\varphi_\alpha(X, \mathcal{G}) \sim \varphi_\alpha(\tilde{g}X, \mathcal{G})$, where $\varphi_\alpha(\tilde{g}X, \mathcal{G})$ is the test function $\varphi_\alpha(X, \mathcal{G})$ computed with $\tilde{g}X$ instead of $X$. Because $\tilde{g} \in \mathcal{G}$ was arbitrary, conclude $\mathrm{E} \sum_{g \in \mathcal{G}} \varphi_\alpha(gX, \mathcal{G}) = \mathrm{E}\varphi_\alpha(X, \mathcal{G})|\mathcal{G}|$. The same argument as the finite-dimensional case now yields $\mathrm{E}\varphi_\alpha(X, \mathcal{G}) \leqslant \alpha$.

For the randomized test decision introduced in (2.5), the arguments so far also imply that $\phi_\alpha(X, \mathcal{G}) \sim \phi_\alpha(gX, \mathcal{G})$ for every $g \in \mathcal{G}$. Use the definition on $a(X)$ to see that $\sum_{g \in \mathcal{G}} \phi_\alpha(gX, \mathcal{G}) = |\{g \in \mathcal{G} : T(gX) > T^{1-\alpha}(X, \mathcal{G})| + a(X)|\{g \in \mathcal{G} : T(gX) = T^{1-\alpha}(X, \mathcal{G})| = \alpha|\mathcal{G}|$ and therefore

$$P(\phi_\alpha(X, \mathcal{G}) \geqslant V) = \mathrm{E}\phi_\alpha(X, \mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathrm{E}\phi_\alpha(gX, \mathcal{G}) = \frac{1}{|\mathcal{G}|} \mathrm{E} \sum_{g \in \mathcal{G}} \phi_\alpha(gX, \mathcal{G}) = \alpha,$$

as desired. □

*Proof of Theorem 2.2.* For $x, x' \in \ell^\infty(\mathcal{U})^q$ and every $g \in \mathcal{G}$, sub-additivity and monotonicity give

$$T(gx) - T(gx') \leqslant \sup_{u \in \mathcal{U}} \frac{1}{q} \left( \sum_{j=1}^q g_j \big( x_j(u) - x'_j(u) \big) \right) \leqslant \sup_{u \in \mathcal{U}} \frac{1}{q} \sum_{j=1}^q \big| x_j(u) - x'_j(u) \big|.$$

The far right of the display is at most $|x - x'|_{\mathcal{U}} / \sqrt{q}$. Reverse the roles of $x$ and $x'$ to conclude $|T(gx) - T(gx')|^2 \leqslant |x - x'|_{\mathcal{U}}^2 / q$ for every $g \in \mathcal{G}$ and therefore

$$\Big| \big( T(gx) - T(gx') \big)_{g \in \mathcal{G}} \Big| \leqslant \sqrt{2^q / q} |x - x'|_{\mathcal{U}}.$$

Let $|x - x'|_{\mathcal{U}} \to 0$ to deduce that $x \mapsto (T(gx))_{g \in \mathcal{G}}$ a continuous map from $\ell^\infty(\mathcal{U})^q$ to $\mathbb{R}^{|\mathcal{G}|}$ with respect to the sup-norm. Because $X_n \rightsquigarrow X$, the continuous mapping theorem implies $(T(gX_n))_{g \in \mathcal{G}} \rightsquigarrow (T(gX))_{g \in \mathcal{G}}$.

Order $\mathcal{G}$ so that the identity action $g = (1, \ldots, 1)$ is the first element. Define

$$B_\alpha = \left\{(t_1, t_2, \ldots, t_{|\mathcal{G}|}) : |\{2 \leqslant i \leqslant |\mathcal{G}| : t_i < t_1\}| \geqslant \lceil (1-\alpha)|\mathcal{G}| \rceil \right\} \tag{A.1}$$

as the set of all vectors where the first element of the vector exceeds at least $\lceil (1-\alpha)|\mathcal{G}| \rceil$ of the remaining elements. Because only the relative ranking of $(T(gX))_{g\in\mathcal{G}}$ enters the test decision, the test rejects if and only if $(T(gX))_{g\in\mathcal{G}} \in B_\alpha$. Conclude that $P(T(X) > T^\alpha(X, \mathcal{G})) = P((T(gX))_{g\in\mathcal{G}} \in B_\alpha)$. The boundary $\partial B_\alpha$ of $B_\alpha$ can be expressed as

$$\partial B_\alpha = \bigcup_{j \geqslant 1} \left\{(t_1, t_2, \ldots, t_{|\mathcal{G}|}) : |t_1 = t_i| = j, |\{2 \leqslant i \leqslant |\mathcal{G}| : t_i < t_1\}| = \lceil (1-\alpha)|\mathcal{G}| \rceil - j \right\}$$

and therefore $\partial B_\alpha \subset \bigcup_{j\geqslant 1} \{(t_1, t_2, \ldots, t_{|\mathcal{G}|}) : |t_1 = t_i| = j\}$. By the portmanteau lemma, $P((T(gX_n))_{g\in\mathcal{G}} \in B_\alpha) \to P((T(gX))_{g\in\mathcal{G}} \in B_\alpha)$ as long $\partial B_\alpha$ satisfies $P((T(gX))_{g\in\mathcal{G}} \in \partial B_\alpha) = 0$. The goal is therefore to show that

$$P\left((T(gX))_{g\in\mathcal{G}} \in \bigcup_{j\geqslant 1} \{(t_1, t_2, \ldots, t_{|\mathcal{G}|}) : |t_1 = t_i| = j\}\right) = 0,$$

i.e., $(T(gX))_{g\in\mathcal{G}}$ has no ties with probability one.

The main difficulty here is that each component of $(T(gX))_{g\in\mathcal{G}}$ is dependent, so the preceding display does not follow from smoothness of the marginals of $(T(gX))_{g\in\mathcal{G}}$. Instead, for $u, u' \in \mathcal{U}$ and $g \neq g'$, write

$$\sum_{j=1}^q g_j X_j(u) - \sum_{j=1}^q g'_j X_j(u') = (g, -g')^T (X(u), X(u'))$$

Because $X$ is a Gaussian process, it follows that $(X(u), X(u'))$ is a jointly Gaussian vector and therefore $(g, -g')^T(X(u), X(u'))$ is a normally distributed random variable.

If $u = u'$ or $u \neq u'$ but $X(u) = X(u')$, then $g \neq g'$ guarantees that $\sum_{j=1}^q g_j X_j(u) - \sum_{j=1}^q g'_j X_j(u) = \sum_{j=1}^q (g_j - g'_j) X_j(u)$ has non-zero variance. Hence, suppose $u \neq u'$ and $X(u) \neq X(u')$. Let $c(u, u') = \mathrm{E}X(u)X(u')$ be the covariance function and note that $(g, -g')^T(X(u), X(u'))$ is zero with positive probability if and only if

$(g, -g')^T c(u, u')(g, -g') = 0$. Because the elements of $X$ are independent, the covariance function satisfies

$$(g, -g')^T c(u, u')(g, -g') = \sum_{j=1}^{n} c_{jj}(u, u) + \sum_{j=1}^{n} c_{jj}(u', u') - 2\sum_{j=1}^{n} g_j g_j' c_{jj}(u, u').$$

Apply the Cauchy-Schwarz inequality to the right-hand side to deduce

$$0 = (g, -g')^T c(u, u')(g, -g') \geqslant \sum_{j=1}^{n} \Big( c_{jj}(u, u) - c_{jj}(u', u') \Big)^2,$$

which implies $\operatorname{Var} X_j(u) = \operatorname{Var} X_j(u')$ for $1 \leqslant j \leqslant q$. It follows that

$$0 = \sum_{j=1}^{n} \Big( c_{jj}(u, u) - g_j g_j' c_{jj}(u, u') \Big)$$

Apply the Cauchy-Schwarz inequality again to see that every covariance must be non-zero because $c_{jj}(u, u) > 0$ and either $c_{jj}(u, u') = c_{jj}(u, u)$ or $c_{jj}(u, u') = -c_{jj}(u, u)$. This implies that either $X_j(u) = X_j(u')$ or $X_j(u) = -X_j(u')$. Because $g \neq g'$, $X(u) = X(u')$ is impossible and at least one $j$ must satisfy $X_j(u) = -X_j(u')$, which is ruled out by assumption. Conclude

$$\sum_{j=1}^{q} g_j X_j(u) \neq \sum_{j=1}^{q} g_j' X_j(u')$$

almost surely for all $u, u' \in \mathcal{U}$ and all $g \neq g'$. Because $\mathcal{U}$ is compact and $X$ has continuous sample paths, this ensures

$$T(gX) - T(g'X) = \max_{u \in \mathcal{U}} \sum_{j=1}^{q} g_j X_j(u) - \max_{u \in \mathcal{U}} \sum_{j=1}^{q} g_j' X_j(u) \neq 0$$

for almost every sample path unless $g = g'$. □

*Proof of Theorem 3.4.* If $H_0$ is true, then scale invariance implies $\varphi_\alpha(\hat{\delta} - \delta_0 1_q, \mathcal{G}) = \varphi_\alpha(X_n, \mathcal{G})$. Assumption 3.3 and Theorem 2.2 yield $\mathrm{E}\varphi_\alpha(\hat{\delta} - \delta_0 1_q, \mathcal{G}) \to \mathrm{E}\varphi_\alpha(X, \mathcal{G})$. $\mathrm{E}\varphi_\alpha(X, \mathcal{G}) \leqslant \alpha$ holds because $X$ satisfies the conditions of Theorem 2.1. □

*Proof of Theorem 3.5.* Suppose $\delta = \delta_0 + \lambda/\sqrt{n}$ so that $X_n + \lambda 1_q \overset{\delta}{\rightsquigarrow} X + \lambda 1_q$. As in the proof of Theorem 3.4, joint continuity of the map $x \mapsto (T(gx))_{g\in\mathcal{G}}$ implies $(T(g(X_n + \lambda 1_q)))_{g\in\mathcal{G}} \overset{\delta}{\rightsquigarrow} (T(g(X + \lambda 1_q)))_{g\in\mathcal{G}}$. With $B_\alpha$ as defined in (A.1), I only have to show that $P((T(g(X+\lambda)))_{g\in\mathcal{G}} \in \partial B_\alpha) = 0$ to conclude $P(T(X_n+\lambda) > T^\alpha(X_n+\lambda,\mathcal{G})) \to P(T(X+\lambda) > T^\alpha(X+\lambda,\mathcal{G}))$.

The boundary has probability zero if $(T(g(X+\lambda)))_{g\in\mathcal{G}}$ has no ties. For $u, u' \in \mathcal{U}$ and $g \neq g'$, write

$$\left[\sum_{j=1}^q g_j X_j(u) - \sum_{j=1}^q g'_j X_j(u')\right] + \lambda(u)\sum_{j=1}^q g_j - \lambda(u')\sum_{j=1}^q g'_j,$$

to see from the proof of Theorem 3.4 that the term in square brackets is nonzero almost surely for all $u, u' \in \mathcal{U}$ and all $g \neq g'$. Because the expression in square brackets is normally distributed with mean zero, it cannot take on any fixed nonzero value with positive probability. The remainder of the preceding display is constant. Conclude that the preceding display is nonzero almost surely for all $u, u' \in \mathcal{U}$ and all $g \neq g'$. As in the proof of Theorem 3.4, this implies $T(g(X+\lambda)) \neq T(g'(X+\lambda))$ almost surely unless $g \neq g'$.

I will now develop a lower bound on $P(T(X+\lambda 1_q) > T^\alpha(X+\lambda 1_q,\mathcal{G}))$. Because the original statistic cannot exceed the largest order statistic, monotonicity implies

$$\begin{aligned} P\Big(T(X+\lambda 1_q) > T^{1-\alpha}(X+\lambda 1_q,\mathcal{G})\Big) &\geqslant P\Big(T(X+\lambda 1_q) > T^{(|\mathcal{G}|-1)}(X+\lambda 1_q,\mathcal{G})\Big) \\ &= P\Big(T(X+\lambda 1_q) = \max_{g\in\mathcal{G}} T\Big(g(X+\lambda 1_q)\Big)\Big) \end{aligned}$$

and the right-hand side is at most

$$P\left(\left\{T(X+\lambda 1_q) = \max_{g\in\mathcal{G}} T\Big(g(X+\lambda 1_q)\Big)\right\}, \bigcap_{j=1}^q \left\{\inf_{u\in\mathcal{U}}(X_j(u)+\lambda(u)) \geqslant 0\right\}\right).$$

If $\inf_{u\in\mathcal{U}}(X_j(u)+\lambda(u)) \geqslant 0$ for $1 \leqslant j \leqslant q$, then $T(X+\lambda 1_q) = \max_{g\in\mathcal{G}} T(g(X+\lambda 1_q))$ because $T$ cannot be increased by making large negative values positive through multiplication by $-1$. By independence and symmetry of the Gaussian processes,

conclude that the preceding display equals

$$\prod_{j=1}^{q} P\Big(\inf_{u\in\mathcal{U}}\big(X_j(u)+\lambda(u)\big)\geqslant 0\Big)=\prod_{j=1}^{q} P\Big(\sup_{u\in\mathcal{U}}\big(X_j(u)-\lambda(u)\big)\leqslant 0\Big).$$

Because $\sup(f-f')\geqslant \sup f-\sup f'$ for arbitrary $f,f'$, this cannot exceed

$$\prod_{j=1}^{q} P\Big(\sup_{u\in\mathcal{U}} X_j(u)\leqslant \sup_{u\in\mathcal{U}}\lambda(u)\Big)\geqslant \prod_{j=1}^{q}\Big(1-e^{-[\sup_u\lambda(u)-\mathrm{E}\sup_u X_j(u)]^2/2\sup_u \mathrm{E}X_j^2(u)}\Big)$$

by the Borell-TIS inequality as long as $\sup_u\lambda(u)>\mathrm{E}\sup_u X_j(u)$. In that case, the right-hand side of the preceding display is strictly positive, as required.

Suppose $\delta=\delta_0+\lambda 1_q$. We have $\hat\delta-\delta_0=X_n/\sqrt{n}+\lambda 1_q$ with $X_n/\sqrt{n}\rightsquigarrow 0$, and by arguments as in the proof of Theorem 3.3, the continuous mapping theorem yields $(T(g(\hat\delta-\delta_0 1_q)))_{g\in\mathcal{G}}\rightsquigarrow (T(g\lambda))_{g\in\mathcal{G}}$. Monotonicity implies

$$\mathrm{E}\varphi_{1-\alpha}(\hat\delta-\delta_0 1_q,\mathcal{G})\geqslant P\Big(T(\hat\delta-\delta_0 1_q)>T^{(|\mathcal{G}|-1)}(\hat\delta-\delta_0 1_q,\mathcal{G})\Big)$$

As before, use a set of the form

$$B=\Big\{(t_1,t_2,\ldots,t_{|\mathcal{G}|}):|\{2\leqslant i\leqslant|\mathcal{G}|:t_i<t_1\}|\geqslant|\mathcal{G}|-1\Big\}$$

to write $P(T(\lambda)>T^{(|\mathcal{G}|-1)}(\lambda,\mathcal{G}))=P((T(g\lambda))_{g\in\mathcal{G}}\in B)=1\{(T(g\lambda))_{g\in\mathcal{G}}\in B\}$. The boundary $\partial B$ is contained in the set

$$\bigcup_{j\geqslant 1}\Big\{(t_1,t_2,\ldots,t_{|\mathcal{G}|}):|t_1=t_i|=j\Big\}.$$

Because $T(g\lambda)=\sup_{u\in\mathcal{U}}\lambda(u)\sum_{j=q}^{q}g_j/q$ with $\lambda\geqslant 0$ and $\sup_{u\in\mathcal{U}}\lambda(u)>0$, we have $T(\lambda)>T(g\lambda)$ for all $g\neq \mathrm{id}:=(1,\ldots,1)$. Hence, there are no ties with the first element of $(T(g\lambda))_{g\in\mathcal{G}}$ and $1\{(T(g\lambda))_{g\in\mathcal{G}}\in\partial B\}=0$. Conclude from the portmanteau lemma that $T(\hat\delta-\delta_0 1_q)-T^{(|\mathcal{G}|-1)}(\hat\delta-\delta_0 1_q,\mathcal{G})\rightsquigarrow\sup_{u\in\mathcal{U}}\lambda(u)-\sup_{u\in\mathcal{U}}\lambda(u)\sum_{j=q}^{q}g_j/q$ for some $g\neq\mathrm{id}$. Because this limit is strictly positive,

$$\mathrm{E}\varphi_{1-\alpha}(\hat\delta-\delta_0,\mathcal{G})\geqslant P\Big(T(\hat\delta-\delta_0)>T^{(|\mathcal{G}|-1)}(\hat\delta-\theta_0,\mathcal{G})\Big)\to 1,$$

as required. □

*Proof of Theorem 3.8.* I can work with $X_{n,[h]} = \sqrt{n}(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}})$ instead of $\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}}$ because $x \mapsto p(x, \mathcal{G})$ is scale invariant. In the following I make repeated use of the fact that the map $x \mapsto (x_{[h]})_{h\in\mathcal{H}}$, the map $x_{[h]} \mapsto (T(gx_{[h]}))_{g\in\mathcal{G}}$, and their composition are continuous.

Suppose $q_1 \leqslant q_0$. The case $q_1 > q_0$ requires only notational changes. The components of $X_{n,[h]}$ are of the form $\sqrt{n}(\hat{\delta}_{j,h(j)} - \delta_0) = \sqrt{n}(\hat{\delta}_{j,h(j)} - \delta)$ under the null hypothesis. By Assumption 3.7, these components converge in distribution to $X_{[h]} := (X_{1,h(1)}, \ldots, X_{q_1,h(q_1)})$ jointly in $h$. The same arguments as in the proof of Theorem 3.4 imply that $T(gX_{n,[h]})$ converges in distribution, jointly in $h$ and $g$, to $T(gX_{[h]})$. For the same reasons as in the proof of Theorem 3.4, for a given $h$, $(T(gX_{[h]}))_{g\in\mathcal{G}\setminus\mathrm{id}}$ has no ties $T(X_{[h]})$ with probability 1, provided Assumption 3.7 holds.

Consider

$$|\mathcal{G}| \sum_{h\in\mathcal{H}} p(X_{n,[h]}, \mathcal{G}) = \sum_{h\in\mathcal{H}} \sum_{g\in\mathcal{G}} 1\{T(gX_{n,[h]}) \geqslant T(X_{n,[h]})\}.$$

This function jumps discretely if, for some $h$ and $g$, $T(gX_{n,[h]}) = T(X_{n,[h]})$. The continuous mapping theorem applies to this function if the probability of hitting these jumps is zero, i.e., $P(T(gX_{n,[h]}) = T(X_{n,[h]}) \text{ for some } g \in \mathcal{G}, h \in \mathcal{H}) = 0$. The union bound implies that this probability cannot exceed $\sum_{h\in\mathcal{H}} \sum_{g\in\mathcal{G}} P(T(gX_{n,[h]}) = T(X_{n,[h]})) = 0$ because $(T(gX_{[h]}))_{g\in\mathcal{G}}$ has no ties almost surely. Conclude that the preceding display converges in distribution to $\sum_{h\in\mathcal{H}} \sum_{g\in\mathcal{G}} 1\{T(gX_{[h]}) \geqslant T(X_{[h]})\}$ and therefore

$$P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,[h]}, \mathcal{G}) \leqslant \alpha\right) \to P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{[h]}, \mathcal{G}) \leqslant \alpha\right)$$

if $\alpha$ is a continuity point of the right-hand side. Because $|\mathcal{G}| \sum_{h\in\mathcal{H}} p(X_{[h]}, \mathcal{G})$ is integer-valued, non-integer values of $\alpha H|\mathcal{G}|/2$ are continuity points.

For integer $\alpha H|\mathcal{G}|/2$, find an $\varepsilon > 0$ such that $(\alpha+\varepsilon)H|\mathcal{G}|/2$ is not an integer but $\alpha + \varepsilon \leqslant 1$. Monotonicity and weak convergence imply

$$\limsup_{n\to\infty} P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,[h]},\mathcal{G}) \leqslant \alpha\right) \leqslant P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{[h]},\mathcal{G}) \leqslant \alpha+\varepsilon\right).$$

By the Rüschendorf (1982) inequality, the right-hand side cannot exceed $\alpha+\varepsilon$. Now let $\varepsilon \searrow 0$ to obtain the desired result. □

*Proof of Theorem 3.9.* As in the proof of Theorem 3.8, let $X_{n,[h]} = \sqrt{n}(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}})$ and $q_1 \leqslant q_0$ without loss of generality. Consider fixed alternatives $\delta = \delta_0 + \lambda$. The components of $\hat{\delta}_{[h]} - \delta_0 1_{q_1}$ are of the form

$$\hat{\delta}_{j,h(j)} - \delta_0 = \sqrt{n}(\hat{\delta}_{j,k} - \delta)/\sqrt{n} + \lambda \rightsquigarrow \lambda$$

by uniform continuity. Deduce that for every $g$ and $h$, $T(X_{n,[h]}/\sqrt{n}) - T(gX_{n,[h]}/\sqrt{n})$ converges in probability to $T(\lambda_{[h]}) - T(g\lambda_{[h]})$. For $g \neq \mathrm{id}$, this limit equals

$$\sup_{u\in\mathcal{U}} \lambda(u) - \sup_{u\in\mathcal{U}} \lambda(u) \sum_{j=q}^{q} g_j/q > 0.$$

Zero is therefore a continuity point of the (degenerate) limiting distribution of $T(X_{n,[h]}/\sqrt{n}) - T(gX_{n,[h]}/\sqrt{n})$, which implies

$$P\Big(T(gX_{n,[h]}/\sqrt{n}) \geqslant T(X_{n,[h]}/\sqrt{n})\Big) \to 0$$

and $1\{T(gX_{n,[h]}/\sqrt{n}) \geqslant T(X_{n,[h]}/\sqrt{n})\} \to 0$ for every $g \neq \mathrm{id}$ and $h$. Conclude that

$$\frac{1}{H}\sum_{h\in\mathcal{H}} p(X_{n,[h]},\mathcal{G}) = \frac{1}{|\mathcal{G}|H}\sum_{h\in\mathcal{H}}\sum_{g\in\mathcal{G}} 1\{T(gX_{n,[h]}) \geqslant T(X_{n,[h]})\} \rightsquigarrow \frac{1}{|\mathcal{G}|}$$

and therefore

$$P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(\hat{\delta}_{[h]} - \delta_0 1_{q_1},\mathcal{G}) \leqslant \alpha\right) \to 1\{2 \leqslant \alpha|\mathcal{G}|\}$$

as long as $\alpha|\mathcal{G}| \neq 2$ to guarantee that convergence occurs at a continuity point.

Now consider local alternatives $u \mapsto \delta(u) = \delta_0(u) + c\lambda(u)/\sqrt{n}$ with $c$ constant. As in the proof of Theorem 3.4, continuity of the maps $x \mapsto x_{[h]}$ and $x_{[h]} \mapsto (T(gx_{[h]}))_{g\in\mathcal{G}}$ implies $(T(g(X_{n,[h]} + c\lambda 1_{q_1})))_{g\in\mathcal{G}} \rightsquigarrow (T(g(X_{[h]} + c\lambda 1_{q_1})))_{g\in\mathcal{G}}$ jointly in $h \in \mathcal{H}$. For a given $h$, $(T(g(X_{[h]} + c\lambda 1_{q_1})))_{g\in\mathcal{G}}$ again has no ties with probability 1. As before, deduce

$$P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,[h]} + c\lambda 1_{q_1}, \mathcal{G}) \leqslant \alpha\right) \to P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{[h]} + c\lambda 1_{q_1}, \mathcal{G}) \leqslant \alpha\right)$$

if $\alpha$ is a continuity point of the right-hand side. Because $|\mathcal{G}|\sum_{h\in\mathcal{H}} p(X_{[h]} + c\lambda 1_{q_1}, \mathcal{G})$ is integer-valued, non-integer values of $\alpha H|\mathcal{G}|/2$ are continuity points. For integer $\alpha H|\mathcal{G}|/2$, find an $\varepsilon > 0$ such that $(\alpha - \varepsilon)H|\mathcal{G}|/2$ is not an integer but $\alpha - \varepsilon > 0$.

$$\liminf_{n\to\infty} P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,[h]} + c\lambda 1_{q_1}, \mathcal{G}) \leqslant \alpha\right) \geqslant P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{[h]} + c\lambda 1_{q_1}, \mathcal{G}) \leqslant \alpha - \varepsilon\right)$$

by monotonicity. Let $\varepsilon \searrow 0$ to see that the limit inferior is bounded below by

$$P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{[h]} + c\lambda 1_{q_1}) < \alpha\right).$$

The same bound holds trivially for non-integer $\alpha H|\mathcal{G}|/2$.

For the analysis as $c \to \infty$, consider

$$\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{[h]} + c\lambda 1_{q_1}) = \frac{2}{|\mathcal{G}|H}\sum_{h\in\mathcal{H}}\sum_{g\in\mathcal{G}} 1\left\{\frac{T(g(X_{[h]} + c\lambda 1_{q_1}))}{T(c\lambda 1_{q_1})} \geqslant \frac{T(X_{[h]} + c\lambda 1_{q_1})}{T(c\lambda 1_{q_1})}\right\}.$$

For $g = \mathrm{id}$, the indicator function in the preceding display equals $q$. Consider $g \neq \mathrm{id}$. Because $T(c\lambda 1_{q_1}) = cT(\lambda 1_{q_1}) > 0$ and $T(gX_{[h]})/T(c\lambda 1_{q_1}) \to 0$ almost surely for every $g \in \mathcal{G}$ as $c \to \infty$, it follows from the subadditivity of suprema that $T(g(X_{[h]} + c\lambda 1_{q_1}))/T(c\lambda 1_{q_1}) \to T(g\lambda 1_{q_1})/T(\lambda 1_{q_1})$ almost surely and therefore $(T(X_{[h]} + c\lambda 1_{q_1}) - T(g(X_{[h]} + c\lambda 1_{q_1})))/T(c\lambda 1_{q_1}) \to 1 - (T(g\lambda 1_{q_1})/T(\lambda 1_{q_1}))$ almost surely. That last limit is a strictly positive constant for every $g \neq \mathrm{id}$ and there is one id for every $h$. Conclude from the continuous mapping theorem that the preceding

display converges almost surely to $2/|\mathcal{G}|$ as $c \to \infty$. If $\alpha|\mathcal{G}| \neq 2$, it follows that

$$\lim_{c\to\infty} P\left( \frac{2}{H} \sum_{h \in \mathcal{H}} p(X_{[h]} + c\lambda 1_q) < \alpha \right) = 1\{2 < \alpha|\mathcal{G}|\},$$

as required. □

*Proof of Proposition 3.11.* If $\mathcal{I}$ is fixed, then the proof of Theorems 3.8 and 3.9 goes through without any changes. For random $\mathcal{I}$, work conditional on $\mathcal{I}$ to see that Theorem 3.8 implies

$$\limsup_{n\to\infty} P\left( \frac{2}{|\mathcal{I}|} \sum_{h \in \mathcal{I}} p(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}}, \mathcal{G}) \leqslant \alpha \mid \mathcal{I} \right) \leqslant \alpha$$

almost surely. Apply expectations to conclude from the (reverse) Fatou lemma that

$$\begin{aligned} &\limsup_{n\to\infty} P\left( \frac{2}{|\mathcal{I}|} \sum_{h \in \mathcal{I}} p(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}}, \mathcal{G}) \leqslant \alpha \right) \\ &\quad \leqslant \mathrm{E} \limsup_{n\to\infty} P\left( \frac{2}{|\mathcal{I}|} \sum_{h \in \mathcal{I}} p(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}}, \mathcal{G}) \leqslant \alpha \mid \mathcal{I} \right) \leqslant \alpha \end{aligned}$$

as needed. Similarly, Fatou's lemma implies

$$\begin{aligned} &\liminf_{n\to\infty} P\left( \frac{2}{|\mathcal{I}|} \sum_{h \in \mathcal{I}} p(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}}, \mathcal{G}) \leqslant \alpha \right) \\ &\quad \geqslant \mathrm{E} \liminf_{n\to\infty} P\left( \frac{2}{|\mathcal{I}|} \sum_{h \in \mathcal{I}} p(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}}, \mathcal{G}) \leqslant \alpha \mid \mathcal{I} \right). \end{aligned}$$

Now apply the first part of Theorem 3.9 for a given $\mathcal{I}$ to get the result for fixed alternatives. For local alternatives, the proof of Theorem 3.9 implies

$$\begin{aligned} &\liminf_{n\to\infty} P\left( \frac{2}{|\mathcal{I}|} \sum_{h \in \mathcal{I}} p(\hat{\delta}_{[h]} - \delta_0 1_{\min\{q_1,q_0\}}, \mathcal{G}) \leqslant \alpha \mid \mathcal{I} \right) \\ &\quad \geqslant P\left( \frac{2}{|\mathcal{I}|} \sum_{h \in \mathcal{I}} p(X_{[h]} + c\lambda 1_{q_1}) < \alpha \mid \mathcal{I} \right) \to 1 \end{aligned}$$

almost surely as $c \to \infty$, as required. □

*Proof of Proposition 3.10.* Limits are as $m \to \infty$ unless noted otherwise. Consider a process $X_n$ possibly depending on $n$ and recall that $T(X_n) > T^{1-\alpha}(X_n, \mathcal{G}_m)$ if and only if $\hat{p}_m := p(X_n, \mathcal{G}_m) \leqslant \alpha$. Let $p := p(X_n, \mathcal{G})$ and notice that $\mathrm{E}(\hat{p} \mid X_n) = p$. For almost every realization of $X_n$, $\hat{p}_m$ is an average of bounded iid random variables that satisfies $P(|\hat{p}_m - p| > \varepsilon \mid X_n) \to 0$ almost surely for every $\varepsilon > 0$. Conclude from dominated convergence that this convergence also holds unconditionally and therefore $\hat{p}_m \rightsquigarrow p$. Because $p$ can only vary at the points $j/|\mathcal{G}|$, $1 \leqslant j \leqslant |G|$, $P(\hat{p}_m \leqslant \alpha) \to P(p \leqslant \alpha)$ as long as $\alpha \neq j/|\mathcal{G}|$. If $\alpha$ equals $j/|\mathcal{G}|$ for some $j$, use $0 < \varepsilon < 1/|\mathcal{G}|$ and monotonicity to see that $P(\hat{p}_m \leqslant \alpha - \varepsilon) \leqslant P(\hat{p}_m \leqslant \alpha) \leqslant P(\hat{p}_m \leqslant \alpha + \varepsilon)$ must satisfy

$$P(p \leqslant \alpha - \varepsilon) \leqslant \liminf_{m\to\infty} P(\hat{p}_m \leqslant \alpha) \leqslant \limsup_{m\to\infty} P(\hat{p}_m \leqslant \alpha) \leqslant P(p \leqslant \alpha + \varepsilon).$$

Let $\varepsilon \searrow 0$ to see that the extreme right-hand side can be decreased to $P(p \leqslant \alpha)$.

For Theorem 3.4, apply this result to obtain

$$\limsup_{m\to\infty} P\Big(T(X_n) > T^{1-\alpha}(X_n, \mathcal{G}_m)\Big) \leqslant P\Big(p(X_n, \mathcal{G}) \leqslant \alpha\Big) = \mathrm{E}\varphi_\alpha(X_n, \mathcal{G}).$$

Now apply limits as $n \to \infty$.

For Theorem 3.8, consider stochastic processes $X_{n,h}$ indexed by $h$ and $n$. The continuous mapping theorem implies $2\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G}_m)/H \xrightarrow{\mathrm{P}} 2\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G})/H$ and therefore $2\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G}_m)/H \rightsquigarrow 2\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G})/H$. Using the same argument as before gives

$$\limsup_{m\to\infty} P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G}_m) \leqslant \alpha\right) \leqslant P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G}) \leqslant \alpha\right)$$

For Theorem 3.5, if $\alpha > 1/2^q$, there is a $\varepsilon > 0$ such that $\alpha - \varepsilon > 1/2^q$. Then

$$\liminf_{m\to\infty} P\Big(T(X_n) > T^{1-\alpha}(X_n, \mathcal{G}_m)\Big) \geqslant P\Big(p(X_n, \mathcal{G}) \leqslant \alpha - \varepsilon\Big) = \mathrm{E}\varphi_{\alpha-\varepsilon}(X_n, \mathcal{G})$$

and Theorem 3.5 applies directly to the extreme right-hand side.

For Theorem 3.9, there is a $\varepsilon > 0$ such that $\alpha - \varepsilon > 1/2^{q-1}$. Then

$$\liminf_{m\to\infty} P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G}_m) \leqslant \alpha\right) \geqslant P\left(\frac{2}{H}\sum_{h\in\mathcal{H}} p(X_{n,h}, \mathcal{G}) \leqslant \alpha - \varepsilon\right)$$

and Theorem 3.9 can be applied to the extreme right-hand side. □

University of Michigan Ross School of Business, 701 Tappan Ave, Ann Arbor, MI 48109, USA. Tel.: +1 (734) 764-2355. Fax: +1 (734) 764-2769. E-mail: hagem@umich.edu.