Econometrics Journal (2017), volume **00**, pp. 1–35.
Article No. ectj??????

Nonparametric inference on (conditional) quantile differences and interquantile ranges, using L-statistics

Matt Goldman[†] and David M. Kaplan[‡]

[†] *Microsoft Research, Redmond, WA 98052, USA.*
E-mail: mattgold@microsoft.com
[‡] *Department of Economics, University of Missouri, 118 Professional Bldg, 909 University Ave, Columbia, MO 65211, USA.*
E-mail: kaplandm@missouri.edu
Received: December 2016

Summary We provide novel, high-order accurate methods for nonparametric inference on quantile differences between two populations in both unconditional and conditional settings. These quantile differences corresponds to (conditional) quantile treatment effects under (conditional) independence of a binary treatment and potential outcomes. Our methods use the probability integral transform and a Dirichlet (rather than Gaussian) reference distribution to pick appropriate L-statistics as confidence interval endpoints, achieving high-order accuracy. Using a similar approach, we also propose confidence intervals/sets for 1) vectors of quantiles, 2) interquantile ranges, and 3) differences of linear combinations of quantiles. In the conditional setting, when smoothing over continuous covariates, optimal bandwidth and coverage probability rates are derived for all methods. Simulations show the new confidence intervals to have a favourable combination of robust accuracy and short length compared with existing approaches. Detailed steps for confidence interval construction are provided in Supplemental Appendix E, and code for all methods, simulations, and empirical examples is provided.

Keywords: *Dirichlet distribution, Fractional order statistics, High-order accuracy, Inequality, Quantile treatment effect.*

1. INTRODUCTION

We consider inference on various quantile-based objects of interest used in empirical economics. The τ-quantile difference (τ-QD) is the difference between the τ-quantiles of two population distributions. Under certain asssumptions, the τ-QD corresponds to the τ-quantile treatment effect, i.e., the difference between the respective τ-quantiles of treated and untreated potential outcome distributions (Doksum, 1974; Lehmann, 1975). These quantile treatment effects capture heterogeneity and distributional impacts unseen in the average treatment effect. For example, such assumptions are satisfied in experimental settings like Gneezy and List (2006).[1] Similarly, the conditional (on covariates) τ-QD corresponds to the conditional τ-quantile treatment effect under certain assumptions (e.g., unconfoundedness), potentially revealing additional heterogeneity across covariate values. The (conditional) QDs also provide valuable summaries of income differences between two groups (conditional on covariates like education), for example.

Interquantile ranges (IQRs) are empirically valuable as robust, versatile measures of

[1] We revisit their data in Supplemental Appendix H; see also Björkman and Svensson (2009) and Charness and Gneezy (2009), among others.

spread and have been used to document trends in income inequality.[2] For example, Angrist, Chernozhukov, and Fernández-Val (2006, Table 1, p. 554) use three measures of U.S. wage inequality: the 90–10 IQR, i.e., the 0.9-quantile minus the 0.1-quantile (90th percentile minus 10th percentile); the 90–50 IQR; and the 50–10 IQR.[3] These show "inequality increasing in both the upper and lower halves of the wage distribution from 1980 to 1990, but in the top half only from 1990 to 2000" (p. 554). Similarly, Kopczuk, Saez, and Song (2010, Figure II, p. 106) examine trends over the period 1937–2004 in U.S. inequality using the 80–50 and 50–20 IQRs of log earnings.[4] Examining conditional (on education) IQRs, Angrist et al. (2006) note, "The increase in conditional inequality since 1990 has been much larger for college graduates than for high school graduates" (p. 554). Describing the 50–20 IQR conditional on sex, Kopczuk et al. (2010) write, "The series for men only is quite different... with an absolute minimum in 1969 followed by a sharp increase up to 1983... [versus] a secular and steady fall since World War II [for women]" (p. 107).

To construct a confidence interval (CI) for a quantile difference, for example, we use linear combinations of order statistics (i.e., L-statistics) as interval endpoints. To determine which L-statistics to use, we rely on the probability integral transform (Fisher, 1932; Neyman, 1937; Pearson, 1933): the distribution of order statistics (i.e., ordered sample values) relative to the population quantile values is analogous to the distribution of order statistics from a Unif$(0,1)$ distribution relative to the quantile indices.

The accuracy of each new unconditional CI is precisely characterised in terms of coverage error, defined as the difference between the true and nominal coverage probabilities. The foundation of our methods' high-order accuracy is the use of the probability integral transform and a Dirichlet (instead of Gaussian) approximation of the distribution of a linear combination of "fractional" order statistics (linearly interpolated between observed order statistics), formally studied in Goldman and Kaplan (2017). For our confidence set for a vector of quantiles, the Dirichlet approximation is the only source of error, so coverage error is $O(n^{-1})$. For a QD or IQR, nuisance parameters arise. Using our proposed bandwidth to nonparametrically estimate the nuisance parameters limits coverage error to $O(n^{-2/3}\log(n))$. Our two-sided CIs are equal-tailed rather than symmetric (like the usual ± 1.96SE interval), giving a more intuitive sense of uncertainty when the estimator's distribution is skewed, as is often true in the tails. In finite-sample simulations, our CI has more robust coverage than a variety of methods (normal, bootstrap, permutation) and usually the shortest length among CIs attaining the desired coverage probability.

Our quantile difference CI's coverage error is of the same theoretical order (up to the $\log(n)$) as the coverage error of the Edgeworth expansion-based method of Kaplan (2015), but our finite-sample coverage error (in simulations) is smaller across a variety

[2]This IQR is a different object of interest than the "interquantile range" of, e.g., Krewski (1976) or Sathe and Lingras (1981), whose CIs are intended to include the entire range $[Q(\tau_1), Q(\tau_2)]$ with probability $1-\alpha$, rather than the difference $Q(\tau_2) - Q(\tau_1)$. Taking the lengths of their "inner" and "outer" CIs forms an asymptotically conservative CI for the difference.

[3]Because they are interested in illustrating their results about quantile regression under misspecification, they actually average conditional quantiles to get overall average conditional IQR values, but the qualitative idea is similar.

[4]Equivalently, they describe these as log quantile ratios, e.g., (the log of) the 0.8-quantile divided by the 0.5-quantile. This is equivalent to the log 0.8-quantile minus the log 0.5-quantile. Because log is a continuous increasing function, this is equivalent to the 80–50 IQR of log earnings. That is, with $Q_Y(\cdot)$ the quantile function of earnings (Y), and $Q_{\log(Y)}(\cdot)$ the quantile function of log earnings, $\log(Q_Y(0.8)/Q_Y(0.5)) = \log(Q_Y(0.8)) - \log(Q_Y(0.5)) = Q_{\log(Y)}(0.8) - Q_{\log(Y)}(0.5)$.

of data generating processes. This finite-sample advantage derives in part from our CI length not being inversely proportional to a probability density function (PDF) estimate, unlike with a normality-based CI. Since nonparametric PDF estimates can be arbitrarily large or close to zero, the corresponding normality-based CI lengths fluctuate more from sample to sample. Section 3 has details.

We extend all our methods to a nonparametric conditional quantile model. If all covariates are discrete, then our methods are immediately applicable, like with subpopulations based on education or sex as in Angrist et al. (2006) or Kopczuk et al. (2010). If some covariates are continuous, then we propose looking at the "local" observations whose covariate vector is near the point of interest (like with local polynomial estimation) and applying our unconditional method to the corresponding outcome obsevations Y_i. We provide bandwidth rates that maximise our methods' coverage accuracy (which differ from those that maximise estimation precision), as well as plug-in bandwidths that work well in simulations; we are unaware of parallel results for local polynomial quantile regression.[5] Beside achieving good theoretical accuracy, our methods are accurate in finite-sample simulations, providing more robust coverage in some cases and shorter intervals in others. Our methods are also easy to use, with steps for construction in Supplemental Appendix E and available as functions in R.

The high-order accuracy achieved by our methods is important not only with modest sample sizes (e.g., for experiments) but also for nonparametric conditional analysis with small *local* sample sizes. For example, even with $n = 1024$ and just five binary covariates, the smallest local sample size cannot exceed $1024/2^5 = 32$.

Our new methods are not simple extensions of existing methods. Naively combining individual quantile CIs (like those in Goldman and Kaplan, 2017) is overly conservative for a QD or IQR, even asymptotically. Projecting from a $1-\alpha$ confidence set for the relevant quantiles also produces an asymptotically conservative CI. For example, Chu (1957) proposes an order statistic-based CI for the IQR, but it is asymptotically conservative because it is based on projection.[6] Our CI appears to be the first order statistic-based CI with even first-order exact asymptotic coverage of the QD or IQR, let alone high-order accuracy.

The main limitations of our conditional approach (with continuous covariates) are the iid sampling assumption and the use of a uniform kernel (which excludes boundary and higher-order kernels). Both seem necessary to link to the Dirichlet distribution and establish high-order accuracy. For inference on a single conditional quantile, Fan and Liu (2016) relax these assumptions, but only first-order accuracy is established, and no conditional QD or conditional IQR method is provided.[7] Qu and Yoon (2015) allow a local linear estimator (eqn. (4)) with a relatively general kernel function (Assumption 4), although still with iid sampling (Assumption 1(iii)), and they only show first-order accuracy (focusing instead on uniformity over the conditional quantile process). Their CI is included in our simulations, where it has much worse coverage error in some cases and significantly longer length in others. Since our method may be seen as using the residuals from a local constant estimator, it may be possible to extend our approach to use residuals from a local linear estimator, but this is left to future work.

Section 2 concerns fractional order statistic theory, including a new result. Sections 3

[5]For local polynomial (mean) regression, such results are in Calonico, Cattaneo, and Farrell (2017).
[6]There are other loose inequalities, like the first two in the proof of Lemma 1 on page 174.
[7]For further comparison in the single quantile case, see Goldman and Kaplan (2017).

and 4 respectively discuss unconditional and conditional confidence intervals. Section 5 contains an empirical example. Appendix A contains proof sketches; Appendix B has implementation details. The Supplemental Appendix (on the journal's website) includes fully detailed proofs, detailed steps to construct each CI we propose, and simulation results, among other material.

Acronyms used include those for [conditional] interquantile range ([C]IQR), [conditional] quantile difference ([C]QD), [conditional] quantile treatment effect ([C]QTE), confidence interval (CI), confidence set (CS), coverage probability (CP), coverage probability error (CPE), cumulative distribution function (CDF), mean squared error (MSE), mean value theorem (MVT), and probability density function (PDF), and GK is Goldman and Kaplan (2017). Notationally, $\text{Beta}(a, b)$ is a beta distribution with parameters a and b, or such a random variable if clear from context, and $\text{Dir}(k_1, k_2, \ldots)$ is a Dirichlet distribution (or random variable); $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF, respectively; $\doteq$ should be read as "is equal to, up to smaller-order terms"; $\asymp$ as "has exact (asymptotic) rate/order of"; and random and non-random (column) vectors are typeset as $\mathbf{Z} = (Z_1, Z_2, \ldots)'$ and $\mathbf{z} = (z_1, z_2, \ldots)'$, respectively, with random and non-random matrices $\underline{\mathbf{Z}}$ and $\underline{\mathbf{z}}$, and scalar random variables and values Z and z. For functions $f : \mathbb{R} \mapsto \mathbb{R}$, let $f(\mathbf{z}) \equiv (f(z_1), \ldots, f(z_J))'$.

2. FRACTIONAL ORDER STATISTICS

In this section, we introduce notation for fractional order statistics, state prior results, and contribute a new result.

Let $X_i \overset{iid}{\sim} F_X(\cdot)$, a continuous, unknown CDF with corresponding quantile function $Q_X(\tau) \equiv \inf\{x : F_X(x) \geq \tau\}$. If additionally $F_X(\cdot)$ is strictly increasing, which we assume it is in a neighbourhood of the quantile(s) of interest, then $Q_X(\cdot) = F_X^{-1}(\cdot)$. Let $X_{n:k}$ denote the kth order statistic in a sample of size n, so $X_{n:1} < X_{n:2} < \cdots < X_{n:n}$. For readers new to (fractional) order statistics, assuming $X_i \overset{iid}{\sim} \text{Unif}(0, 1)$ upon first reading may help with intuition since then $Q_X(\tau) = \tau$ and $F_X(x) = x$ for $x \in [0, 1]$.

The primary Dirichlet results that follow are from Wilks (1962, pp. 236–238). By the probability integral transform (Fisher, 1932; Neyman, 1937; Pearson, 1933) and continuity of $F_X(\cdot)$,

$$U_i \equiv F_X(X_i) \overset{iid}{\sim} \text{Unif}(0, 1), \tag{2.1}$$

and $F_X(Q_X(u)) = u$ for any $u \in (0, 1)$. Using the distribution of standard uniform order statistics,

$$U_{n:k} \equiv F_X(X_{n:k}) \sim \text{Beta}(k, n + 1 - k). \tag{2.2}$$

Moreover, the vector

$$(U_{n:1}, U_{n:2} - U_{n:1}, \ldots, U_{n:n} - U_{n:n-1}, 1 - U_{n:n}) \sim \text{Dir}(1, \ldots, 1), \tag{2.3}$$

a Dirichlet distribution with each of the $n + 1$ parameters equal to one. Generally, a Dirichlet distribution is supported on the unit simplex, i.e., vectors whose components sum to one; the distribution in (2.3) is uniform over the unit simplex (i.e., constant PDF). A Dirichlet's univariate marginals are beta distributions; in (2.3), these are $U_{n:k+1} - U_{n:k} \sim \text{Beta}(1, n)$, a highly right-skewed distribution on $[0, 1]$ with mode equal to zero and mean $1/(n + 1)$.

The Wilks (1962) results determine coverage probabilities. For the τ-quantile,

$$\mathrm{P}(X_{n:k} < Q_X(\tau)) = \mathrm{P}(F_X(X_{n:k}) < F_X(Q_X(\tau))) = \mathrm{P}(F_X(X_{n:k}) < \tau), \tag{2.4}$$

which from (2.2) is the Beta$(k, n+1-k)$ CDF evaluated at τ. This provides the exact, finite-sample coverage probability (CP) of any confidence interval for $Q_X(\tau)$ whose endpoint is an order statistic, $X_{n:k}$.

For given n and desired CP $1-\alpha$, it is unlikely that any integer k exactly solves $\mathrm{P}(F_X(X_{n:k}) < \tau) = 1-\alpha$. To proceed, one must either choose a different α, randomise, or interpolate. In this paper, we interpolate. The following definitions are helpful.

Definition 2.1. *The linearly (L) interpolated kth fractional order statistic for X is*

$$\hat{X}^L_{n:k} \equiv (1-\epsilon)X_{n:\lfloor k \rfloor} + \epsilon X_{n:\lfloor k \rfloor+1}, \quad \epsilon \equiv k - \lfloor k \rfloor, \tag{2.5}$$

where ϵ is the interpolation weight and $\lfloor \cdot \rfloor$ is the floor function. With reference to (2.1) *and* (2.2)*, the idealised (I) kth fractional order "statistic" for U is*

$$\begin{aligned} \tilde{U}^I_{n:k} &\equiv (1-C)U_{n:\lfloor k \rfloor} + CU_{n:\lfloor k \rfloor+1} \sim \mathit{Beta}(k, n+1-k), \\ C &\sim \mathit{Beta}(\epsilon, 1-\epsilon), \quad C \perp\!\!\!\perp \{U_{n:k}\}_{k=1}^n, \end{aligned} \tag{2.6}$$

using Jones (2002, eqn. (2)). If k is an integer, then $\epsilon = C = 0$, so $\hat{X}^L_{n:k} = X_{n:k} = Q_X(\tilde{U}^I_{n:k})$.

The construction of $\tilde{U}^I_{n:k}$ can be extended to multiple[8] k as in Jones (2002), with joint distribution similar to (2.3):

$$(\tilde{U}^I_{n:k_1}, \tilde{U}^I_{n:k_2} - \tilde{U}^I_{n:k_1}, \ldots, 1 - \tilde{U}^I_{n:k_J}) \sim \mathrm{Dir}(k_1, k_2 - k_1, \ldots, n+1-k_J). \tag{2.7}$$

Notationally, we write $\tilde{U}^I$ instead of $\hat{U}^I$ as a reminder that (unless k is an integer) these are randomised statistics (relying on C), not purely functions of sample values. In contrast, $\hat{X}^L_{n:k}$ is a (non-randomised) linear combination of observed order statistics, i.e., an L-statistic.

There is always a (fractional) k^* that exactly solves $\mathrm{P}(\tilde{U}^I_{n:k^*} < \tau) = 1-\alpha$ because the beta CDF is continuous. If the distribution of $F_X(\hat{X}^L_{n:k^*})$ is well approximated by $\tilde{U}^I_{n:k^*} \sim \mathrm{Beta}(k^*, n+1-k^*)$ from (2.6), then $\hat{X}^L_{n:k^*}$ is the endpoint of an approximate $1-\alpha$ confidence interval for $Q_X(\tau)$.

For inference on a quantile difference, we will need to approximate the distribution of the sample quantile difference that uses fractional order statistics. Theorem 2.1 is a new result providing such an approximation; a more general version (also new) is stated in Theorem A.2 and subsequently proved. For IQR inference, the needed result is in Theorem 2 of Goldman and Kaplan (2017).

Assumption A2.1. *For sampling, $X_i \overset{iid}{\sim} F_X(\cdot)$ with sample size n_x (or just n). If applicable, $Y_i \overset{iid}{\sim} F_Y(\cdot)$ with sample size n_y, samples are independent, and sample sizes have the same asymptotic rate: $n_x/n_y = \delta^2 + O(n^{-1})$ for constant $0 < \delta < \infty$.*

Assumption A2.2. *At any quantile index of interest τ, (a) $f_X(Q_X(\tau)) > 0$, and (if*

[8]Originally, Stigler (1977) proposed a fractional order statistic process (for a continuum of $k \in (0, n+1)$) following a Dirichlet process (Ferguson, 1973), but we use only finite-dimensional distributions.

applicable) $f_Y(Q_Y(\tau)) > 0$*; (b)* $f'_X(\cdot)$ *is continuous in a neighbourhood of* $Q_X(\tau)$*, and (if applicable)* $f'_Y(\cdot)$ *is continuous in a neighbourhood of* $Q_Y(\tau)$*.*

Assumption A2.2 implies that the first three derivatives of the quantile function are uniformly bounded in a neighbourhood of τ, and that the first derivative is uniformly bounded away from zero. The PDF having two derivatives is also required for a two-term Edgeworth expansion of a sample quantile; e.g., see Theorem 13.2 of DasGupta (2008, p. 189–190). Our confidence intervals still have exact asymptotic coverage (just not higher-order accuracy) without the PDF derivatives.

Theorem 2.1. *Let Assumptions A2.1 and A2.2 hold. Let* $L_0 \equiv Q_Y(u) - Q_X(u)$*. Uniformly over* $u = \tau + o(1)$*,*

$$\sup_{K\in\mathbb{R}} \big| \mathrm{P}(\hat{Y}^L_{n_y:(n_y+1)u} - \hat{X}^L_{n_x:(n_x+1)u} < L_0 + n^{-1/2}K) - \mathrm{P}(Q_Y(\tilde{U}^I_{n_y:(n_y+1)u}) - Q_X(\tilde{U}^I_{n_x:(n_x+1)u}) < L_0 + n^{-1/2}K) \big| = O(n^{-1}).$$

3. UNCONDITIONAL INFERENCE

In this section, we develop new confidence intervals (CIs) for the unconditional objects of interest and present their theoretical properties. For all methods, detailed steps for construction are given in Supplemental Appendix E, and code is available in the replication files on the journal's (or latter author's) website.

For quantile index $\tau_j \in (0,1)$ and confidence level $1-\alpha$, let $k^h_j(\alpha)$ and $k^l_j(\alpha)$ be defined to satisfy (superscript "h" for high, "l" for low)

$$\alpha = \mathrm{P}(\tilde{U}^I_{n:k^h_j(\alpha)} < \tau_j), \quad \alpha = \mathrm{P}(\tilde{U}^I_{n:k^l_j(\alpha)} > \tau_j), \tag{3.1}$$

where $\tilde{U}^I_{n:k} \sim \mathrm{Beta}(k, n+1-k)$ from (2.6). (For QD inference, we omit the j subscript.) These are equivalent to (7) and (8) in Hutson (1999).

For a single quantile $Q_X(\tau_j)$, one-sided CI endpoints are then $\hat{X}^L_{n:k^h_j(\alpha)}$ or $\hat{X}^L_{n:k^l_j(\alpha)}$. An equal-tailed two-sided CI has endpoints $\hat{X}^L_{n:k^h_j(\alpha/2)}$ and $\hat{X}^L_{n:k^l_j(\alpha/2)}$. Such CIs were proposed by Hutson (1999) and shown to have coverage probability error (CPE) of order $O(n^{-1})$ by Goldman and Kaplan (2017).

With multiple quantiles or samples, our general approach is to construct an individual CI for each quantile and then combine the CIs into an appropriate overall CI. However, using the nominal α in (3.1) to get pointwise $1-\alpha$ CP for the individual CIs leads to over-coverage or under-coverage, even asymptotically. Additional arguments are required to determine and estimate the properly calibrated pointwise CP $1-\tilde{\alpha}$ to achieve overall $1-\alpha$ CP for each particular object of interest.

3.1. Joint inference on multiple quantiles

We construct a confidence set (CS) for the vector

$$Q_X(\boldsymbol{\tau}) \equiv (Q_X(\tau_1), \ldots, Q_X(\tau_J))',$$

where each $\tau_j \in (0,1)$. Given $\tilde{\alpha}$, we construct a nominal $1-\tilde{\alpha}$ CI for each $Q_X(\tau_j)$ and take the Cartesian product to be the CS for $Q_X(\boldsymbol{\tau})$. With two-sided individual CIs, using

(3.1), the CS for $Q_X(\boldsymbol{\tau})$ is the Cartesian product

$$\prod_{j=1}^{J}[\hat{X}^L_{n:k_j^l(\tilde{\alpha}/2)}, \hat{X}^L_{n:k_j^h(\tilde{\alpha}/2)}], \tag{3.2}$$

We use the same $\tilde{\alpha}$ at each quantile for simplicity and to achieve equal pointwise CP of each $Q_X(\tau_j)$.[9]

The value of $\tilde{\alpha}$ must be precisely calibrated to achieve optimal coverage accuracy. Since there is positive but not perfect dependence across quantiles, $\tilde{\alpha} = 1-(1-\alpha)^{1/J}$ is too small and $\tilde{\alpha} = \alpha$ is too big; neither yields asymptotically exact CP, let alone high-order accuracy. To achieve high-order accuracy, $\tilde{\alpha}$ is calibrated using (2.7) such that

$$1-\alpha = \mathrm{P}\left(\left\{\bigcap_{j=1}^{J}\{\tilde{U}^I_{n:k_j^h(\tilde{\alpha}/2)} > \tau_j\}\right\}\bigcap\left\{\bigcap_{j=1}^{J}\{\tilde{U}^I_{n:k_j^l(\tilde{\alpha}/2)} < \tau_j\}\right\}\right), \tag{3.3}$$

where $\bigcap$ denotes the intersection of events and can be read as "and." Solving (3.3) for $\tilde{\alpha}$ numerically is easy because 1) it is a one-dimensional search over $\tilde{\alpha} \in (0,1)$ and 2) the right-hand side is a strictly decreasing function of $\tilde{\alpha}$.

Theorem 3.1. *Under Assumptions A2.1 and A2.2, the CS in* (3.2) *has CPE of order* $O(n^{-1})$.

The rectangular CS in (3.2) has some advantages over the elliptical CS based on asymptotic normality, although being larger is a disadvantage. First, it will be a helpful intermediate step in constructing the QD and IQR CIs, which are not larger than normality-based CIs. Second, the CS can be used to test the family of hypotheses $H_{0j} : Q_X(\tau_j) = Q_0(\tau_j)$ for $j = 1, \ldots, J$, rejecting H_{0j} if and only if $Q_0(\tau_j)$ lies outside the $1-\tilde{\alpha}$ CI for $Q_X(\tau_j)$. Because the CIs have *joint* $1-\alpha$ CP, this procedure has strong control of the familywise error rate at level α as defined in, e.g., Lehmann and Romano (2005, §9.1): the probability of falsely rejecting at least one true H_{0j} is below α, regardless of which H_{0j} are true. Goldman and Kaplan (2016) extend this approach to distributional inference where $J = n$, including stepdown and pre-test procedures to improve power.

3.2. Inference on interquantile ranges

The object of interest is, for $0 < \tau_1 < \tau_2 < 1$,

$$\mathrm{IQR} = Q_X(\tau_2) - Q_X(\tau_1). \tag{3.4}$$

As in Section 3.1, $1-\tilde{\alpha}$ CIs are constructed for $Q_X(\tau_1)$ and $Q_X(\tau_2)$, respectively, but now a different $\tilde{\alpha}$ must be chosen to achieve asymptotically exact CP.[10] The CI for the IQR contains all values $q_2 - q_1$ such that q_1 and q_2 are in the respective $1-\tilde{\alpha}$ CIs for $Q_X(\tau_1)$ and $Q_X(\tau_2)$.

Naively using $\tilde{\alpha} = \alpha$ causes first-order CPE: even asymptotically, the CI is too wide.

[9]The related idea of even pointwise type I error rates is applied to inference on the entire distribution by Goldman and Kaplan (2016).

[10]If one instead constructs a $1-\tilde{\alpha}_1$ CI for the τ_1-quantile and a $1-\tilde{\alpha}_2$ CI for the τ_2-quantile, for the corresponding hypothesis test, there is no first-order tradeoff in power among combinations $(\tilde{\alpha}_1, \tilde{\alpha}_2)$ that control size, as seen in the proof of Theorem 3.2(c), so in this sense nothing is lost by setting $\tilde{\alpha}_1 = \tilde{\alpha}_2 = \tilde{\alpha}$ as we do. This is partly because the object of interest is a scalar, unlike in Section 3.1.

The intuition is the same as with a difference of normal random variables whose positive covariance is unaccounted for.[11] Alternatively, using the $\tilde{\alpha} < \alpha$ that generates a $1-\alpha$ confidence set for $(Q_X(\tau_1), Q_X(\tau_2))$ is even worse (i.e., even wider). Instead, we determine the $\tilde{\alpha} > \alpha$ that achieves exact asymptotic CP and minimises CPE.

The CPE-optimal $\tilde{\alpha}$ is often much larger than the naive choice $\tilde{\alpha} = \alpha$. For example, with $n = 100$, $Q_X(u) = u$ (uniform), and $(\tau_1, \tau_2) = (0.25, 0.75)$ (interquartile range), the CPE-optimal $\tilde{\alpha}$ given $\alpha = 0.1$ is $\tilde{\alpha} = 0.34$ (rounded). In a simulation with 10 000 replications, the corresponding CI had CP 0.9058, extremely close to the nominal $1-\alpha = 0.9$. In contrast, the CI using $\tilde{\alpha} = \alpha$ was far too wide, with 0.9954 CP. Projecting from a $1-\alpha$ confidence set for $(Q_X(\tau_1), Q_X(\tau_2))$ entails $\tilde{\alpha} = 0.053$ (rounded), which is even more conservative, with 0.9991 CP.

Unlike in Section 3.1, the CPE-optimal $\tilde{\alpha}$ is not distribution-free. Now that the object of interest is a linear combination of quantiles, we need to work with linear approximations of $Q_X(\cdot)$. This reduces the dimension of the nuisance parameter from infinity to two, by using only the scalars $Q'_X(\tau_1)$ and $Q'_X(\tau_2)$ instead of the function $Q_X(\cdot)$, but it does not eliminate the nuisance parameter. Ignoring the linear approximation error, and assuming k_1 and k_2 are integers for now,

$$\begin{aligned}\mathrm{P}(X_{n:k_2} - X_{n:k_1} > Q_X(\tau_2) - Q_X(\tau_1)) &= \mathrm{P}(Q_X(U_{n:k_2}) - Q_X(\tau_2) - (Q_X(U_{n:k_1}) - Q_X(\tau_1)) > 0) \\ &\approx \mathrm{P}((U_{n:k_2} - \tau_2)Q'_X(\tau_2) - (U_{n:k_1} - \tau_1)Q'_X(\tau_1) > 0).\end{aligned}$$

Equivalently, this can be written in terms of a single nuisance parameter by dividing through by $Q'_X(\tau_2)$:

$$\mathrm{P}\left((U_{n:k_2} - \tau_2) - (U_{n:k_1} - \tau_1)\frac{Q'_X(\tau_1)}{Q'_X(\tau_2)} > 0\right).$$

That is, after taking linear approximations of the quantile function at τ_1 and τ_2, what matters asymptotically is the ratio of the slopes. Either way, using (2.7), the joint distribution of $(U_{n:k_1}, U_{n:k_2})$ is known since

$$(U_{n:k_1}, U_{n:k_2} - U_{n:k_1}, 1 - U_{n:k_2}) \sim \mathrm{Dir}(k_1, k_2 - k_1, n + 1 - k_2),$$

but $Q'_X(\tau_1)$ and $Q'_X(\tau_2)$ must be estimated. More generally, Theorem 2 of Goldman and Kaplan (2017) lets us use fractional k_1 and k_2 with only smaller-order error from interpolation; the proof of our Theorem 3.2 also rigourously treats the error from linearisation that was ignored above, as well as the nuisance parameter estimation error.

The nuisance parameters $Q'_X(\tau_1)$ and $Q'_X(\tau_2)$ are also known as "sparsities" since they can be written as the inverse of a density: $Q'_X(\tau) = 1/f_X(Q_X(\tau))$.[12] We use a "quantile spacing" estimator first proposed by Siddiqui (1960). For $j = 1, 2$, given smoothing parameter m_j,

$$\widehat{Q'_X}(\tau_j) = \frac{n}{2m_j}(X_{n:\lfloor (n+1)\tau_j \rfloor + m_j} - X_{n:\lfloor (n+1)\tau_j \rfloor - m_j}), \tag{3.5}$$

[11] For example, if $X \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$ are jointly normal with covariance $\sigma_{XY} \geq 0$ (since order statistics are positively correlated), then $Y - X \sim \mathrm{N}(0, 1 + 1 - 2\sigma_{XY})$: the variance is at most 2, when $\sigma_{XY} = 0$. The "CI endpoint" $Y - X + 1.96\sqrt{2} \approx Y - X + 2.8$ thus has at least 95% probability of being above zero. Naively combining the individual 95% "CIs" $Y + 1.96$ and $X - 1.96$ instead yields $Y - X + 3.9$, much wider still.

[12] Although the inverse density is a more common presentation, we prefer the quantile derivative notation as a more explicit reminder of its origin in a linear expansion, and because it can be directly estimated (whereas the inverse density requires first estimating $\hat{Q}_X(\tau)$ as the point of evaluation).

with $\widehat{Q'_X}$ (instead of $\hat{Q}'_X$) indicating the estimator of a derivative (instead of the derivative of an estimator).

We choose m_j in (3.5) to make the CI for the IQR as accurate as possible, i.e., to minimise CPE. Interestingly, this CPE-optimal rate is different than the rate $m_j \asymp n^{4/5}$ that minimises mean squared error (MSE) of the nuisance parameter estimator $\widehat{Q'_X}(\tau_j)$ (analogous to the more familiar $m_j/n \asymp n^{-1/5}$ MSE-optimal kernel bandwidth rate). Instead, $m_j \asymp n^{2/3}$ minimises two-sided CPE, echoing results from Hall and Sheather (1988, p. 384) for inference on a single quantile.[13] Estimation details, including our plug-in bandwidth (for m_j), are in Appendix B.1.

The lower one-sided CI for the IQR is, using notation from (2.5) and (3.1),

$$(-\infty, \hat{X}^L_{n:k_2^h(\tilde{\alpha})} - \hat{X}^L_{n:k_1^l(\tilde{\alpha})}), \tag{3.6}$$

where $\tilde{\alpha}$ implicitly depends on $\widehat{Q'_X}(\tau_1)$ and $\widehat{Q'_X}(\tau_2)$ and satisfies

$$1-\alpha = \mathrm{P}(\widehat{Q'_X}(\tau_2)(\tilde{U}^I_{n:k_2^h(\tilde{\alpha})} - \tau_2) - \widehat{Q'_X}(\tau_1)(\tilde{U}^I_{n:k_1^l(\tilde{\alpha})} - \tau_1) > 0). \tag{3.7}$$

As shown in the proof of Theorem 3.2, the actual CP of such a CI is $1-\alpha+T_h+E_h+O(n^{-1})$, where T_h is the remainder from the first-order Taylor expansion ("T" for Taylor), E_h is from estimation error in $\widehat{Q'_X}(\tau_1)$ and $\widehat{Q'_X}(\tau_2)$, and $O(n^{-1})$ is from Goldman and Kaplan (2017, Thm. 2). The rate-limiting term turns out to be $T_h = O(n^{-1/2}\log(n))$.

For an upper one-sided CI for the IQR, the analogues of (3.6) and (3.7) are, respectively,

$$(\hat{X}^L_{n:k_2^l(\tilde{\alpha})} - \hat{X}^L_{n:k_1^h(\tilde{\alpha})}, \infty), \tag{3.8}$$

$$1-\alpha = \mathrm{P}(\widehat{Q'_X}(\tau_2)(\tilde{U}^I_{n:k_2^l(\tilde{\alpha})} - \tau_2) - \widehat{Q'_X}(\tau_1)(\tilde{U}^I_{n:k_1^h(\tilde{\alpha})} - \tau_1) < 0). \tag{3.9}$$

CP is $1-\alpha+T_l+E_l+O(n^{-1})$, where similarly $T_l = O(n^{-1/2}\log(n))$ dominates.

An equal-tailed two-sided $1-\alpha$ CI for the IQR is the intersection of upper and lower one-sided $1-\alpha/2$ CIs. With $m_j \asymp n^{2/3}$, $E_h + E_l = O(n^{-2/3}\log(n))$ is the dominant CPE term since $T_h + T_l = O(n^{-1}(\log(n))^2)$.

The following theorem collects results on CPE of the CIs and power of the corresponding hypothesis tests. As stated and proved in the appendix (Theorem A.3), the same rates hold when generalising the object of interest to any linear combination of quantiles, and our code implements the more general method.

Theorem 3.2. *Let Assumptions A2.1 and A2.2 hold.*

(a) The one-sided lower and upper CIs in (3.6) *and* (3.8) *have* $O(n^{-1/2}\log(n))$ *CPE if* $\widehat{Q'_X}(\tau_1)$ *and* $\widehat{Q'_X}(\tau_2)$ *are estimated*[14] *by* (3.5) *with smoothing parameters* m_1 *and* m_2 *having rate larger than* $n^{1/2}$ *and smaller than* $n^{3/4}$.

[13] Specifically, their result comes from minimising the higher-order terms in type I error of a Studentised quantile-based hypothesis test, using an Edgeworth expansion. Kaplan (2015, §5) instead controls type I error by using fixed-smoothing critical values and chooses m to maximise (higher-order) power, but he derives the same $m \asymp n^{2/3}$ rate, and even a very similar formula. Kaplan (2015, §5.3) also finds the $n^{2/3}$ rate optimal for QD testing.

[14] Absent a formal demonstration, we conjecture the same rates hold if the estimator in (3.5) is replaced with a kernel density estimator with bandwidth of order m_j/n. In practice, results are often numerically equivalent to multiple significant figures.

(b) *Two-sided CIs, formed by the intersection of upper and lower one-sided* $1-\alpha/2$ *CIs, have* $O(n^{-2/3}\log(n))$ *CPE if* $\widehat{Q'_X}(\tau_1)$ *and* $\widehat{Q'_X}(\tau_2)$ *are estimated*[14] *by* (3.5) *with* $m_1, m_2 \asymp n^{2/3}$.

(c) *The asymptotic probabilities of excluding* $D_n = Q_X(\tau_2) - Q_X(\tau_1) + \kappa n^{-1/2}$ *from lower one-sided (l), upper one-sided (u), and equal-tailed two-sided (t) CIs (i.e., asymptotic power of the corresponding hypothesis tests) are*

$$\mathcal{P}_n^l(D_n) \to \Phi(z_\alpha + S),\ \ \mathcal{P}_n^u(D_n) \to \Phi(z_\alpha - S),\ \ \mathcal{P}_n^t(D_n) \to \Phi(z_{\alpha/2} + S) + \Phi(z_{\alpha/2} - S),$$

where $S \equiv \kappa/\sqrt{\mathcal{V}}$ *and*

$$\mathcal{V} \equiv \frac{\tau_1(1-\tau_1)}{(f_X(Q_X(\tau_1)))^2} + \frac{\tau_2(1-\tau_2)}{(f_X(Q_X(\tau_2)))^2} - 2\frac{\tau_1(1-\tau_2)}{f_X(Q_X(\tau_1))f_X(Q_X(\tau_2))}, \tag{3.10}$$

the usual asymptotic variance of the (scaled) sample IQR.

3.3. Inference on two-sample quantile differences

QD inference is very similar to IQR inference, but somewhat easier: we need to approximate the distribution of a difference of order statistics, but now we assume they are independent. Consequently, this section largely parallels Section 3.2 but is briefer.

The object of interest is the τ-QD, $D = Q_Y(\tau) - Q_X(\tau)$. Notationally, certain variables defined previously will now have an additional subscript denoting the sample (x or y).

CI construction is similar to Section 3.2. One-sample $1-\tilde{\alpha}$ CIs are constructed for $Q_X(\tau)$ and $Q_Y(\tau)$. The CI for the QD contains all values $q_Y - q_X$ such that q_Y and q_X are in the respective $1-\tilde{\alpha}$ CIs for $Q_Y(\tau)$ and $Q_X(\tau)$. As in Section 3.2, naively using $\tilde{\alpha} = \alpha$ is asymptotically conservative, as is projecting from a $1-\alpha$ CS that has $\tilde{\alpha} < \alpha$; we find the $\tilde{\alpha} > \alpha$ achieving exact asymptotic CP and minimising higher-order CPE.

For intuition, ignoring interpolation and remainder terms and letting $n_x = n_y = n$, consider the CP of a one-sided CI with upper endpoint $Y_{n:k_2} - X_{n:k_1}$. Using the probability integral transform, write $Y_{n:k_2} = Q_Y(U^Y_{n:k_2})$ and $X_{n:k_1} = Q_X(U^X_{n:k_1})$, where U^Y and U^X are independent sets of uniform order statistics. Similar to Section 3.2,

$$\begin{aligned}\mathrm{P}(Y_{n:k_2} - X_{n:k_1} > Q_Y(\tau) - Q_X(\tau)) &= \mathrm{P}(Q_Y(U^Y_{n:k_2}) - Q_Y(\tau) - (Q_X(U^X_{n:k_1}) - Q_X(\tau)) > 0) \\ &\approx \mathrm{P}((U^Y_{n:k_2} - \tau)Q'_Y(\tau) - (U^X_{n:k_1} - \tau)Q'_X(\tau) > 0).\end{aligned}$$

Using (2.7) and A2.1, the joint distribution of $(U^Y_{n:k_2}, U^X_{n:k_1})$ is known to be

$$U^Y_{n:k_2} \sim \mathrm{Beta}(k_2, n+1-k_2), \quad U^X_{n:k_1} \sim \mathrm{Beta}(k_1, n+1-k_1), \quad U^Y_{n:k_2} \perp\!\!\!\perp U^X_{n:k_1},$$

but $Q'_Y(\tau)$ and $Q'_X(\tau)$ must be estimated. Similar to the IQR case, the proof of Theorem 3.3 shows how to use Theorem 2.1 to allow for fractional k_1 and k_2 and treats the errors from linearisation and nuisance parameter estimation.

The lower one-sided QD CI and calibration equation parallel the IQR versions in (3.6) and (3.7), respectively: using (2.7), (3.1), and (3.5),

$$(-\infty, \hat{Y}^L_{n_y:k_y^h(\tilde{\alpha})} - \hat{X}^L_{k_x^l(\tilde{\alpha})}), \tag{3.11}$$

$$1-\alpha = \mathrm{P}(\widehat{Q'_Y}(\tau)(\tilde{U}^{I,Y}_{n_y:k_y^h(\tilde{\alpha})} - \tau) - \widehat{Q'_X}(\tau)(\tilde{U}^{I,X}_{n_x:k_x^l(\tilde{\alpha})} - \tau) > 0). \tag{3.12}$$

For an upper one-sided CI, the analogues of (3.11) and (3.12) are

$$(\hat{Y}^L_{n_y:k^l_y(\tilde{\alpha})} - \hat{X}^L_{n_x:k^h_x(\tilde{\alpha})}, \infty), \tag{3.13}$$

$$1-\alpha = \mathrm{P}(\widehat{Q'_Y}(\tau)(\tilde{U}^{I,Y}_{n_y:k^l_y(\tilde{\alpha})} - \tau) - \widehat{Q'_X}(\tau)(\tilde{U}^{I,X}_{n_x:k^h_x(\tilde{\alpha})} - \tau) < 0). \tag{3.14}$$

An equal-tailed two-sided CI is the intersection of upper and lower one-sided $1-\alpha/2$ CIs.

The CPE rates in Theorem 3.3 are the same as in Theorem 3.2. Theorem A.6 (in the appendix) shows that the same rates hold more generally for CIs for differences of linear combinations of quantiles.

Theorem 3.3. *Let Assumptions A2.1 and A2.2 hold.*

(a) The one-sided lower and upper CIs in (3.11) *and* (3.13) *have* $O(n^{-1/2}\log(n))$ *CPE if* $\widehat{Q'_X}(\tau)$ *and* $\widehat{Q'_Y}(\tau)$ *are estimated*[15] *as in* (3.5) *with* $n^{1/2} \lesssim m_x, m_y \lesssim n^{3/4}$.

(b) Two-sided CIs, formed by the intersection of upper and lower one-sided $1-\alpha/2$ *CIs, have* $O(n^{-2/3}\log(n))$ *CPE if* $\widehat{Q'_X}(\tau)$ *and* $\widehat{Q'_Y}(\tau)$ *are estimated*[15] *as in* (3.5) *with* $m_x, m_y \asymp n^{2/3}$.

(c) The asymptotic probabilities of excluding $D_n = Q_Y(\tau) - Q_X(\tau) + \kappa n^{-1/2}$ *from lower one-sided (l), upper one-sided (u), and equal-tailed two-sided (t) CIs (i.e., asymptotic power of the corresponding hypothesis tests) are*

$$\mathcal{P}^l_n(D_n) \to \Phi(z_\alpha + S),\ \ \mathcal{P}^u_n(D_n) \to \Phi(z_\alpha - S),\ \ \mathcal{P}^t_n(D_n) \to \Phi(z_{\alpha/2} + S) + \Phi(z_{\alpha/2} - S),$$

$$S \equiv \kappa/\sqrt{\mathcal{V}_x + \mathcal{V}_y} = \frac{\kappa}{\sqrt{\tau(1-\tau)}\sqrt{(Q'_X(\tau))^2 + (Q'_Y(\tau))^2}}.$$

For additional intuition, we provide an analytic $\tilde{\alpha}$ based on a Gaussian rather than beta distribution, with $n_x = n_y$. For a two-sided CI, letting $\gamma = Q'_Y(\tau)/Q'_X(\tau)$,

$$\tilde{\alpha}/2 = \Phi(\Phi^{-1}(\alpha/2)/\theta^*), \quad \theta^* \equiv \frac{1+\gamma}{\sqrt{1+\gamma^2}}. \tag{3.15}$$

Since $\gamma \in [0, \infty)$, then $\theta^* \in [1, \sqrt{2}]$. The largest possible $\tilde{\alpha}$ is attained if $\theta^* = \sqrt{2}$, when $Q'_X(\tau) = Q'_Y(\tau)$ so that $\gamma = 1$, and the sample quantiles $\hat{Q}_X(\tau)$ and $\hat{Q}_Y(\tau)$ have the same asymptotic variance. The smallest possible value is $\tilde{\alpha} = \alpha$, as $Q'_Y(\tau) \to 0$, $\gamma \to 0$, and $\theta^* \to 1$. That is, as the distribution of Y collapses to a constant (locally), the problem reduces to one-sample inference (asymptotically), so $\tilde{\alpha} = \alpha$ is intuitive. More details are in Supplemental Appendix F.

Section 1 mentioned a finite-sample advantage of our method over normality-based CIs whose length is proportional to a (nonparametrically estimated) PDF-based standard error. This advantage can now be illustrated further. Consider a one-sided 90% CI for the median difference $Q_Y(0.5) - Q_X(0.5)$ between two Unif$(0,1)$ populations, with $n_x = n_y = 39$. The upper endpoint should be approximately $Y_{39:23} - X_{39:17}$ (simulated CP is 91.4%). Even with a worst-case coding error that sets $\widehat{Q'_X}(0.5)/\widehat{Q'_Y}(0.5) = 0$ in every sample, the CI would not change dramatically: it would be $Y_{39:24} - X_{39:16}$ and have 96.7% CP. This CI is somewhat too long, but it is much better than the normality-based CI with a worst-case error setting $\widehat{Q'_X}(0.5)$ or $\widehat{Q'_Y}(0.5)$ arbitrarily large: such a CI is

[15]The comments in Footnote 14 apply here, too.

arbitrarily long, with CP nearing 100%. At the opposite extreme, if instead $X = 0.5$ is a degenerate random variable while $Y \sim \text{Unif}(0,1)$, but we had a worst-case error setting $\widehat{Q'_X}(0.5)/\widehat{Q'_Y}(0.5) = 1$, our CI would have 83.2% CP. This is somewhat too short, but it is much better than the normality-based CI with a worst-case error setting $\widehat{Q'_X}(0.5) = \widehat{Q'_Y}(0.5) = 0$: such a CI has 0% CP. This illustrates how our approach improves finite-sample performance by ameliorating the effect of nonparametric nuisance parameter estimation on both CP and CI length variability.

4. CONDITIONAL INFERENCE

We now present CIs for conditional versions of the objects in Section 3. Theoretical properties are given, including the bandwidth rate that maximises coverage accuracy. Detailed steps for construction are given in Supplemental Appendix E, and code is available in the replication material on the journal's (or latter author's) website.[16]

Let $Q_{Y|\mathbf{W}}(u \mid \mathbf{w})$ be the conditional u-quantile function of scalar Y given $\mathbf{W} \in \mathcal{W} \subseteq \mathbb{R}^d$, evaluated at $\mathbf{W} = \mathbf{w}$. A sample of $\{Y_i, \mathbf{W}_i\}_{i=1}^n$ is drawn. If the conditional CDF $F_{Y|\mathbf{W}}(\cdot \mid \mathbf{w})$ is continuous at $Q_{Y|\mathbf{W}}(u \mid \mathbf{w})$, then $F_{Y|\mathbf{W}}(Q_{Y|\mathbf{W}}(u \mid \mathbf{w}) \mid \mathbf{w}) = u$. For a chosen $\mathbf{W} = \mathbf{w}_0$, quantile indices $\tau_j \in (0,1)$, and observed binary ("treatment") variable T_i, the objects of interest are

$$\begin{array}{ll} \text{Vector:} & (Q_{Y|\mathbf{W}}(\tau_1 \mid \mathbf{w}_0), \ldots, Q_{Y|\mathbf{W}}(\tau_J \mid \mathbf{w}_0)), \\ \text{CIQR:} & Q_{Y|\mathbf{W}}(\tau_2 \mid \mathbf{w}_0) - Q_{Y|\mathbf{W}}(\tau_1 \mid \mathbf{w}_0), \\ \text{CQD:} & Q_{Y|\mathbf{W},T}(\tau \mid \mathbf{w}_0, 1) - Q_{Y|\mathbf{W},T}(\tau \mid \mathbf{w}_0, 0). \end{array}$$

The CQD has a causal interpretation under conditional independence where $Y_0, Y_1 \perp\!\!\!\perp T \mid \mathbf{W}$ for potential outcomes Y_0 and Y_1, as in Assumption 1 of MaCurdy, Chen, and Hong (2011, p. 545). Then, the CQD is a conditional quantile treatment effect as on their page 547. In our appendix, we include results for more general objects of interest, including conditional IQR differences.

If $\mathbf{W}$ is discrete, we can take all observations with $\mathbf{W}_i = \mathbf{w}_0$ and compute the appropriate CI from the corresponding Y_i values. This achieves the same CPE rate as in the unconditional setting, although of course finite-sample CPE is greater. The CPE rate remains unchanged even with considerable dependence among the $\mathbf{W}_i$ as long as the Y_i are conditionally independent and the local sample size $\sum_{i=1}^n \mathbb{1}\{\mathbf{W}_i = \mathbf{w}_0\}$ is almost surely of order n. The key is that we have iid draws of Y_i from the same conditional quantile function, $Q_{Y|\mathbf{W}}(\cdot \mid \mathbf{w}_0)$, so the problem reduces to that of the unconditional setting.

If $\mathbf{W}$ is continuous, then we must use observations with $\mathbf{W}_i \neq \mathbf{w}_0$. As in Chaudhuri (1991), we use observations with $\mathbf{W}_i \in C_b$ for some set C_b depending on bandwidth b. Although we omit the subscript n, this is actually a deterministic *sequence* of bandwidths, b_n.

If $d = 1$, then C_b is the interval $[w_0 - b, w_0 + b]$. This is equivalent to using a uniform kernel or symmetric nearest-neighbour method. For $d > 1$, C_b is a hypercube centred at $\mathbf{w}_0$. Since the hypercube has the same width in each dimension, some normalisation of the $\mathbf{W}_i$ should be used in practice, although the asymptotic CPE rates are unaffected.

[16]Our code relies on the following contributions: Koenker (2016), Duong (2017), Hayfield and Racine (2008), Furrer, Nychka, and Sain (2012), and of course R Core Team (2017).

Definition 4.1. *Given bandwidth b and point of interest $\mathbf{w}_0$, let*

$$C_b \equiv \{\mathbf{w} : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w} - \mathbf{w}_0\|_\infty \leq b\}, \quad N_n \equiv \sum_{i=1}^{n} \mathbb{1}\{\mathbf{W}_i \in C_b\},$$

where $\|\mathbf{v}\|_\infty \equiv \max_{1\leq j\leq d}|v_j|$ is the max-norm and C stands for (hyper)cube. The "local sample" is

$$\{Y_i : \mathbf{W}_i \in C_b, 1 \leq i \leq n\},$$

with local sample size N_n. The u-quantile of Y conditional on $\mathbf{W} \in C_b$ is denoted $Q_{Y|\mathbf{W}}(u \mid C_b)$ and satisfies

$$u = \mathrm{P}(Y < Q_{Y|\mathbf{W}}(u \mid C_b) \mid \mathbf{W} \in C_b).$$

For mixed discrete and continuous components of $\mathbf{W}$, we can restrict attention to the subsample where the discrete components of $\mathbf{W}_i$ exactly match those of $\mathbf{w}_0$, and then smooth over the continuous components. The subsample size is still of order n, so asymptotic rates are unaffected by the presence of discrete covariates (although, again, the finite-sample difference may be important). Consequently, without loss of generality, we proceed with $\mathbf{W}_i$ containing only continuous components.

Assumption A4.1. *Sampling of $\{Y_i, \mathbf{W}_i\}$ or $\{Y_i, \mathbf{W}_i, T_i\}$ is iid, for continuous scalar Y_i, continuous vector $\mathbf{W}_i \in \mathcal{W} \subseteq \mathbb{R}^d$, and (for CQD inference) binary indicator T_i with $0 < \mathrm{P}(T_i = 1) < 1$. The point of interest $\mathbf{W} = \mathbf{w}_0$ is in the interior of $\mathcal{W}$, and the quantile indices of interest are $\tau_j \in (0, 1)$.*

Assumption A4.2. *The marginal density of $\mathbf{W}$, denoted $f_{\mathbf{W}}(\cdot)$, satisfies $0 < f_{\mathbf{W}}(\mathbf{w}_0) < \infty$ and has at least one continuous derivative. For CQD inference, this applies to both $f_{\mathbf{W}|T}(\mathbf{w}_0 \mid T = 0)$ and $f_{\mathbf{W}|T}(\mathbf{w}_0 \mid T = 1)$.*

Assumption A4.3. *For each τ_j, for all u in a neighbourhood of τ_j and all $\mathbf{w}$ in a neighbourhood of $\mathbf{w}_0$, $Q_{Y|\mathbf{W}}(u \mid \mathbf{w})$ has at least two continuous derivatives in $\mathbf{w}$. For CQD inference, this applies to both $Q_{Y|\mathbf{W},T}(u \mid \mathbf{w}, 0)$ and $Q_{Y|\mathbf{W},T}(u \mid \mathbf{w}, 1)$.*

Assumption A4.4. *For each τ_j, for all u in a neighbourhood of τ_j and all $\mathbf{w}$ in a neighbourhood of $\mathbf{w}_0$, $f_{Y|\mathbf{W}}(Q_{Y|\mathbf{W}}(\tau_j \mid \mathbf{w}_0) \mid \mathbf{w}_0)$ is uniformly bounded away from zero. For CQD inference, this applies to both $f_{Y|\mathbf{W},T}(Q_{Y|\mathbf{W},T}(\tau_j \mid \mathbf{w}_0, 0) \mid \mathbf{w}_0, 0)$ and $f_{Y|\mathbf{W},T}(Q_{Y|\mathbf{W},T}(\tau_j \mid \mathbf{w}_0, 1) \mid \mathbf{w}_0, 1)$.*

Assumption A4.5. *For each τ_j, for all y in a neighbourhood of $Q_{Y|\mathbf{W}}(\tau_j \mid \mathbf{w}_0)$ and all $\mathbf{w}$ in a neighbourhood of $\mathbf{w}_0$, $f_{Y|\mathbf{W}}(y \mid \mathbf{w})$ has a second derivative in its first argument (y) that is uniformly bounded and continuous in y. For CQD inference, this applies to both $f_{Y|\mathbf{W},T}(y \mid \mathbf{w}, 0)$ and $f_{Y|\mathbf{W},T}(y \mid \mathbf{w}, 1)$, for y in a neighbourhood of any $Q_{Y|\mathbf{W},T}(\tau_j \mid \mathbf{w}_0, 0)$ or $Q_{Y|\mathbf{W},T}(\tau_j \mid \mathbf{w}_0, 1)$, respectively.*

Assumption A4.6. *As $n \to \infty$, (a) $b \to 0$, (a') $b^{2+d/2}\sqrt{n} \to 0$, (b) $nb^d/(\log(n))^2 \to \infty$. For CQD inference, this applies to both b_0 and b_1 (bandwidths for the $T_i = 0$ and $T_i = 1$ subsamples), and $b_0 \asymp b_1$.*

All assumptions except A4.6(b) are used in the bias calculation discussed below. As-

sumptions A4.1 and A4.4–A4.6 help satisfy Assumption A2.2 when applying the unconditional methods to the data in C_b. Assumptions A4.1, A4.2, and A4.6 help establish that N_n is almost surely of order nb^d. Given b and n, Assumption A4.1 implies the local sample of Y_i values are drawn iid from the conditional distribution of $Y \mid \mathbf{W} \in C_b$, as in the unconditional setup. (Local sampling is not purely iid since C_b changes with n, as in a row-wise iid triangular array.)

From A4.6(i), asymptotically C_b is entirely contained within the neighbourhoods in A4.3 and A4.5. To get $N_n \xrightarrow{a.s.} \infty$, A4.6(b) is a primitive condition. This in turn allows us to restrict attention to only local neighbourhoods around the conditional quantiles of interest since the CI endpoints converge to the true value at a $\sqrt{N_n}$ rate.

Our smoothness assumptions are quite mild. Since we use a uniform kernel, which is second-order, there is no bias reduction benefit from assuming additional derivatives. Furthermore, any type of conditional heteroskedasticity is permitted.

For joint inference over m different values of $\mathbf{w}_0$, the usual α/m Bonferroni adjustment can be used. Given the iid sampling in A4.1, if the bandwidth windows (C_b) are mutually exclusive, this can be refined slightly to $1-(1-\alpha)^{1/m}$.

There are two sources of CPE: one from applying an unconditional method, and one from the bias, $Q_{Y|\mathbf{W}}(\tau \mid C_b) - Q_{Y|\mathbf{W}}(\tau \mid \mathbf{w}_0)$. For example, for the CQD confidence interval $\widehat{\mathrm{CI}}$, we can decompose the CP for $D = Q_{Y|\mathbf{W},T}(\tau \mid \mathbf{w}_0, 1) - Q_{Y|\mathbf{W},T}(\tau \mid \mathbf{w}_0, 0)$ into

$$\begin{aligned}
\mathrm{P}(D \in \widehat{\mathrm{CI}}) &= 1 - \alpha + \mathrm{CPE}_{\mathrm{U}} + \mathrm{CPE}_{\mathrm{Bias}},\\
\mathrm{CPE}_{\mathrm{U}} &= \mathrm{P}((Q_{Y|\mathbf{W},T}(\tau \mid C_b, 1) - Q_{Y|\mathbf{W},T}(\tau \mid C_b, 0)) \in \widehat{\mathrm{CI}}) - (1-\alpha),\\
\mathrm{CPE}_{\mathrm{Bias}} &= \mathrm{P}(D \in \widehat{\mathrm{CI}}) - \mathrm{P}((Q_{Y|\mathbf{W},T}(\tau \mid C_b, 1) - Q_{Y|\mathbf{W},T}(\tau \mid C_b, 0)) \in \widehat{\mathrm{CI}}).
\end{aligned}$$

If $\mathbf{W}$ were discrete, then $C_b = \{\mathbf{w}_0\}$, so there would be no bias: $\mathrm{CPE}_{\mathrm{Bias}} = 0$. Replacing n with N_n in results from Section 3 gives $\mathrm{CPE}_{\mathrm{U}}$ (as more rigourously justified in the proofs).

Under A4.2–A4.6, Lemma 5 in Goldman and Kaplan (2017) has

$$Q_{Y|\mathbf{W}}(\tau \mid C_b) - Q_{Y|\mathbf{W}}(\tau \mid \mathbf{w}_0) = O(b^2).$$

Further, they show that the CPE from this bias for a one-sided CI for $Q_{Y|\mathbf{W}}(\tau \mid \mathbf{w}_0)$ is $O(b^2 N_n^{1/2})$. This is essentially the $O(b^2)$ bias times the $O(N_n^{1/2})$ PDF of the CI endpoint (by applying the mean value theorem). Since our new methods involve a fixed number of quantiles, the order of the CPE due to bias remains $O(b^2 N_n^{1/2})$.

The $\mathrm{CPE}_{\mathrm{U}}$ and $\mathrm{CPE}_{\mathrm{Bias}}$ terms have high-level parallels to high-order terms in the Edgeworth expansion for Studentised local polynomial (mean) regression estimators, as in Calonico et al. (2017). Consider Theorem S.II.1(a) in their Supplement (pp. 58–59).[17] Recall that for an unconditional mean, the usual one-sided CI has $O(n^{-1/2})$ CPE, while a two-sided CI has $O(n^{-1})$ CPE since the $n^{-1/2}$ terms from the two endpoints cancel; see Theorem 13.3 of DasGupta (2008, p. 191), for example. For the conditional mean, analogous to the idea of our $\mathrm{CPE}_{\mathrm{U}}$ term, the unconditional $n^{-1/2}$ rate becomes an $N_n^{-1/2} \asymp (nb)^{-1/2}$ term (with scalar W) for a one-sided CI, and the n^{-1} term becomes N_n^{-1}. Our $\mathrm{CPE}_{\mathrm{U}}$ is paralleled by their $(nb)^{-1/2}$ term for a one-sided CI, or by their $(nb)^{-1}$

[17] http://www.tandfonline.com/doi/suppl/10.1080/01621459.2017.1285776/suppl_file/uasa_a_1285776_sm5085.pdf

term for a two-sided CI (since the $(nb)^{-1/2}$ terms cancel). Our CPE_{Bias} is also paralleled in Calonico et al. (2017): their "scaled bias" term has rate $b^2\sqrt{nb}$ (for a local constant or local linear estimator), i.e., $b^2\sqrt{N_n}$, the same rate as our CPE_{Bias} term. Even more clear are the parallels between Theorem 2(a) of Calonico et al. (2017) and the CPE for a single conditional quantile: Goldman and Kaplan (2017, p. 345) have leading CPE terms with rates N_n^{-1}, b^2, and $N_n b^4$, identical to the three leading higher-order terms in Theorem 2(a) of Calonico et al. (2017) for the two-sided local linear-based CI for the conditional mean.[18] In all cases, the CPE-optimal bandwidth rate can be derived by minimising the leading higher-order terms; see Corollary 5 in Calonico et al. (2017), for example.

Our results focus on the "CPE-optimal" bandwidth rate that minimises $\text{CPE}_{\text{U}} + \text{CPE}_{\text{Bias}}$. This is different than the usual "MSE-optimal" bandwidth rate that minimises a point estimator's mean squared error (MSE), which leads to very poor $O(1)$ CPE. The usual asymptotic argument in the literature is that any amount of under-smoothing (i.e., $b \to 0$ faster) leads to first-order coverage accuracy because the bias is then asymptotically negligible (compared to the asymptotic standard deviation); e.g., see Assumption (H) in Fan and Liu (2016, p. 202). Instead, we minimise CPE to attain the best high-order accuracy.[19] This difference is particularly important in small local samples where severe under-coverage is more likely. The tradeoff is that our CI lengths are generally longer. In practice, we suggest a bandwidth that drifts from the CPE-optimal rate in small samples (to maximise coverage accuracy) towards the MSE-optimal rate in large samples (to increase precision).

CPE-optimal bandwidth rates and overall CPE rates are collected in Theorem 4.1. As stated and proved in the appendix, the same rates hold when generalising CIQR to any linear combination of conditional quantiles and generalising CQD to differences of linear combinations of conditional quantiles; our code implements these more general methods for $d = 1$.

Theorem 4.1. *Under Definition 4.1 and Assumptions A4.1–A4.6, the following bandwidth and CPE rates are CPE-optimal (up to* $\log(n)$ *terms). (a) Joint inference, one-sided or two-sided:* $b^* \asymp n^{-3/(4+3d)}$, $\text{CPE} = O(n^{-4/(4+3d)})$*; (b) CIQR or CQD inference, one-sided:* $b^* \asymp n^{-1/(2+d)}$, $\text{CPE} = O(n^{-1/(2+d)}\log(n))$*; (c) CIQR or CQD inference, two-sided:* $b^* \asymp n^{-7/(12+7d)}$, $\text{CPE} = O(n^{-8/(12+7d)}\log(n))$.

Appendix B.2 contains details on plug-in bandwidths.

5. EMPIRICAL APPLICATION

The following results can be replicated using the replication materials from the journal's (or latter author's) website. An additional application to the "gift exchange" experiment of Gneezy and List (2006) is presented in Supplemental Appendix H, also included in the replication package.

Simulation results for both unconditional and conditional methods may be found in

[18] From page 27, $\eta_{\text{us}} \asymp \sqrt{nb}b^2$ for a local linear (or local constant) estimator, so $\eta_{\text{us}}/\sqrt{nb} \asymp b^2$ and $\eta_{\text{us}}^2 \asymp (\sqrt{nb}b^2)^2 = N_n b^4$.

[19] This approach of minimising CPE instead of MSE has been used for bandwidth selection by, among others, Hall and Sheather (1988) and Kaplan (2015) for unconditional quantiles and quantile differences, and by Calonico et al. (2017) with kernel density and local polynomial regression estimators, using Edgeworth expansions in each case.

Supplemental Appendix G, including a DGP based on the following empirical application; all may be replicated with the replication materials from the journal's (or latter author's) website. In the unconditional simulations, compared with bootstrap, permutation, normality, and Edgeworth expansion based methods, our new CIs achieve the desired coverage probability in a wide variety of DGPs (unlike most other methods) while usually being the shortest among those CIs also achieving the nominal coverage probability. In the conditional simulations, compared with a bias-corrected local linear CI, our CI has more robust coverage and often (14/30 cases) shorter length, too. Compared with a bootstrapped local cubic CI, our CI has similar coverage (sometimes better, sometimes worse) and usually shorter length when both CIs attain the correct coverage probability. Away from the median, the bootstrap CIs tend to be shorter since they are symmetric, whereas ours are equal-tailed; i.e., in the tails there is a trade-off between being short and being equal-tailed. Overall, the L-statistic approach seems to provide accurate, robust, equal-tailed, short confidence intervals.

We extend the analysis of Deaton and Paxson (1998) to quantile Engel curve differences. They consider the theoretical prediction that larger households benefit from economies of scale for "public" (within a household) goods like housing and utilities, consequently shifting expenditure into private goods like food. Instead of conditional means, our object of interest is, for example, the difference in median food budget share between two-adult and one-adult households (with no children) conditional on a value of log total per capita expenditure (PCE). We use inflation-adjusted[20] data from the 2001–2012 U.K. Living Costs and Food Surveys (Office of National Statistics, 2012).[21] We compare two-adult ($n = 25\,648$) and one-adult ($n = 20\,833$) households with no children, as well as two-adult, two-child ($n = 7095$) and one-adult, one-child ($n = 2490$) households, examining food, alcohol, and housing/utilities budget shares.

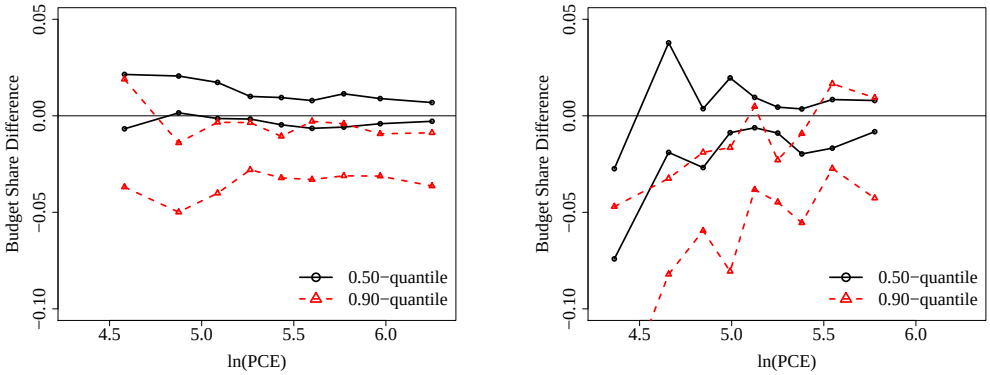

Figure 1. Pointwise 90% confidence intervals (connected for visual ease) for conditional quantile differences in food budget share between households of different size, conditional on real log total per capita expenditure. Left: two adults minus one adult, no children; right: two adults/two children minus one adult/one child.

[20] `https://www.ons.gov.uk/generator?format=csv&uri=/economy/inflationandpriceindices/timeseries/d7bt/mm23/previous/v1`

[21] This is a successor of the Family Expenditure Survey, which Deaton and Paxson (1998) used.

Figure 1 shows food budget share conditional quantile difference CIs. The right panel, comparing two-adult/two-child households with one-adult/one-child households, shows mostly negative differences, consistent with the negative conditional mean differences found by Deaton and Paxson (1998). Food being a private good seems to be outweighed by other factors, like the reduced cost of cooking food at home. The left panel compares two-adult and one-adult childless households, which Deaton and Paxson (1998, p. 910) noted to be an exception where food share was slightly *higher* in two-adult than one-adult households in the UK. Here, the conditional median CIs are consistent with such a pattern, but the conditional 0.9-quantile difference is still negative. Differences across PCE are mostly small. There appears to be compression of the upper half of the food budget share distribution, which we formally examine below. The compression is apparent in the right panel, too.

Figure 1 also shows that unlike with childless households, there are significant differences across PCE when comparing two-adult/two-child and one-adult/one-child households. Specifically, the larger households have much lower food budget shares at lower levels of total PCE, but this difference attenuates as PCE increases, becoming statistically indistinguishable from zero at the highest PCE. This could be explained by the single parents (at any PCE) lacking the time and/or family size to make cooking at home worthwhile, while low-PCE two-adult/two-child households take more advantage of the lower costs of home food production.

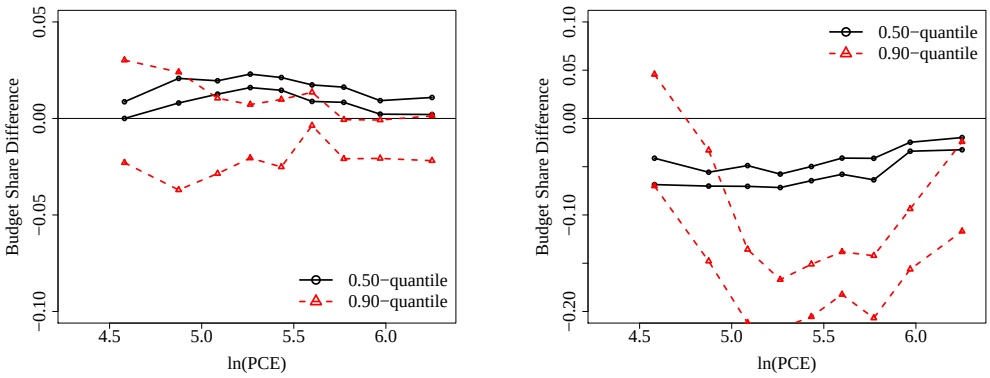

Figure 2. Pointwise 90% confidence intervals (connected for visual ease) for conditional quantile differences in alcohol (left) or housing/utilities (right; note different vertical scaling) budget share between two-adult and one-adult childless households, conditional on real log total per capita expenditure.

Figure 2 is similar to the left panel of Figure 1, but for alcohol (left) and housing and utilities (right) instead of food. In line with theory, at any PCE, the larger households generally have smaller budget shares of housing and utilities (public goods) due to economies of scale. Consequently, they have room for larger budget shares of alcohol, a private good. However, there are significant differences across both PCE and quantile in both graphs.

Figure 3 directly examines differences in conditional interquantile ranges (0.9-quantile minus median) in food budget share, motivated by the pattern observed in Figure 1. Re-

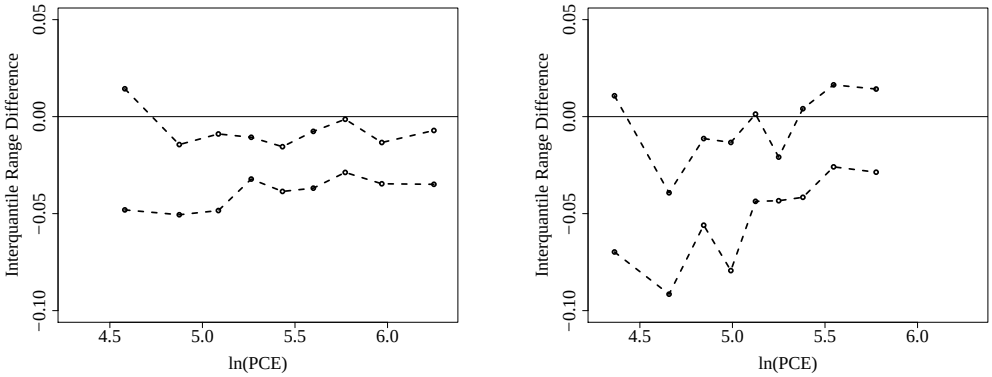

Figure 3. Pointwise 90% confidence intervals (connected for visual ease) for conditional interquantile range (0.9-quantile minus median) differences in food budget share between households of different size, conditional on real log total per capita expenditure. Left: two adults minus one adult, no children; right: two adults/two children minus one adult/one child.

call that naively differencing conditional IQR CIs would produce too conservative (wide) CIs compared with ours shown here. With exceptions at some PCE levels, most of the CIs are well below zero, showing that the upper part of the conditional distributions is indeed compressing. This may be driven partly by a "regression to the mean" phenomenon. The same compression is seen for alcohol and housing/utilities in Supplemental Appendix Figure 5.

6. CONCLUSION

For inference on various quantile-based objects of economic interest, we have proposed confidence intervals that are implemented as functions in R, following the steps detailed in Supplemental Appendix E. We have characterised the theoretical properties of these new nonparametric, L-statistic-based, equal-tailed confidence intervals for quantile differences and interquantile ranges (and differences of linear combinations), as well as confidence sets for vectors of quantiles, in both unconditional and conditional settings. Simulations reflect the theoretical accuracy and robustness, showing a favourable combination of coverage accuracy and length compared with existing methods.

Future research could use our framework to derive confidence intervals with shorter length rather than the equal-tailed/median-unbiased property. Related work on inference on distributions and quantile marginal effects is found in Goldman and Kaplan (2016) and Kaplan (2014), respectively. A formal extension to regression discontinuity as in the setup of Calonico, Cattaneo, and Titiunik (2014) may also be of interest, as well as developing closer connections to the literature on nonseparable models. It may also be possible to apply our unconditional methods to residuals from a local polynomial estimator, instead of a local constant estimator as in the current paper (implicitly); this could help reduce bias and boundary effects. Alternatively, one could keep the "local constant" approach but try to explicitly correct for bias, as Calonico et al. (2017) show can be more accurate in a local polynomial (mean) regression setting.

ACKNOWLEDGEMENTS

We thank Yixiao Sun along with reviewers and editors for helpful comments and references. Thanks also to Brendan Beare, Patrik Guggenberger, Karen Messer, Andres Santos, and active participants at seminars and conferences.

REFERENCES

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica 74*(2), 539–563.

Björkman, M. and J. Svensson (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda. *Quarterly Journal of Economics 124*(2), 735–769.

Bloch, D. A. and J. L. Gastwirth (1968). On a simple estimate of the reciprocal of the density function. *Annals of Mathematical Statistics 39*(3), 1083–1085.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2017). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association XX*(XX), XXX–XXX. Forthcoming.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Charness, G. and U. Gneezy (2009). Incentives to exercise. *Econometrica 77*(3), 909–931.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics 19*(2), 760–777.

Chu, J. T. (1957). Some uses of quasi-ranges. *Annals of Mathematical Statistics 28*(1), 173–180.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. New York: Springer.

Deaton, A. and C. Paxson (1998). Economies of scale, household size, and the demand for food. *Journal of Political Economy 106*(5), 897–930.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Annals of Statistics 2*(2), 267–277.

Duong, T. (2017). *ks: Kernel smoothing*. R package version 1.10.6.

Fan, Y. and R. Liu (2016). A direct approach to inference in nonparametric and semiparametric quantile models. *Journal of Econometrics 191*(1), 196–216.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics 1*(2), 209–230.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th ed.). Edinburgh: Oliver and Boyd.

Furrer, R., D. Nychka, and S. Sain (2012). *fields: Tools for spatial data*. R package version 6.6.3.

Gneezy, U. and J. A. List (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica 74*(5), 1365–1384.

Goldman, M. and D. M. Kaplan (2016). Comparing distributions by multiple testing across quantiles. Working Paper WP 16-19, Department of Economics, University of Missouri, available at `http://faculty.missouri.edu/~kaplandm`.

Goldman, M. and D. M. Kaplan (2017). Fractional order statistic approximation for nonparametric conditional quantile inference. *Journal of Econometrics 196*(2), 331–346.

Hall, P. and S. J. Sheather (1988). On the distribution of a Studentized quantile. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 50*(3), 381–391.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5), 1–32.

Hutson, A. D. (1999). Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *Journal of Applied Statistics 26*(3), 343–353.

Jones, M. C. (2002). On fractional uniform order statistics. *Statistics and Probability Letters 58*(1), 93–96.

Kaplan, D. M. (2014). Nonparametric inference on quantile marginal effects. Working Paper WP 14-13, Department of Economics, University of Missouri, available at `http://faculty.missouri.edu/~kaplandm`.

Kaplan, D. M. (2015). Improved quantile inference via fixed-smoothing asymptotics and Edgeworth expansion. *Journal of Econometrics 185*(1), 20–32.

Koenker, R. (2016). *quantreg: Quantile Regression*. R package version 5.29.

Kopczuk, W., E. Saez, and J. Song (2010). Earnings inequality and mobility in the United States: Evidence from Social Security data since 1937. *Quarterly Journal of Economics 125*(1), 91–128.

Krewski, D. (1976). Distribution-free confidence intervals for quantile intervals. *Journal of the American Statistical Association 71*(354), 420–422.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (3rd ed.). Springer Texts in Statistics. Springer.

MaCurdy, T., X. Chen, and H. Hong (2011). Flexible estimation of treatment effect parameters. *American Economic Review (Papers and Proceedings) 101*(3), 544–551.

Neyman, J. (1937). »Smooth test» for goodness of fit. *Skandinavisk Aktuarietidskrift 20*(3–4), 149–199.

Office for National Statistics and Department for Environment, Food and Rural Affairs (2012). Living Costs and Food Survey. 2nd Edition. Colchester, Essex: UK Data Archive. `https://doi.org/10.5255/UKDA-SN-7472-2`.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika 25*, 379–410.

Qu, Z. and J. Yoon (2015). Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics 185*(1), 1–19.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sathe, Y. S. and S. R. Lingras (1981). Bounds for the confidence coefficients of outer and inner confidence intervals for quantile intervals. *Journal of the American Statistical Association 76*(374), 473–475.

Siddiqui, M. M. (1960). Distribution of quantiles in samples from a bivariate population. *Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics 64B*(3), 145–150.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Volume 26 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall/CRC.

Stigler, S. M. (1977). Fractional order statistics, with applications. *Journal of the American Statistical Association 72*(359), 544–550.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley.

A. PROOFS AND PROOF SKETCHES

We abbreviate Goldman and Kaplan (2017) as GK. For one-sample results, we omit the X subscript, e.g., writing $F(\cdot)$ instead of $F_X(\cdot)$.

First, we introduce notation. The following is one-sample notation; for two-sample notation, subscript x or y is added to indicate the sample. Vectors are always column vectors (unless otherwise noted) and in bold, e.g., $\mathbf{u} = (u_1, \ldots, u_J)'$.

Instead of (2.5), let the linearly interpolated fractional order statistics be

$$\hat{Q}_X^L(k/(n+1)) \equiv \hat{X}_{n:k}^L.$$

Let the idealised fractional order statistics be

$$\tilde{Q}_X^I(k/(n+1)) \equiv \tilde{X}_{n:k}^I, \quad \tilde{Q}_U^I(k/(n+1)) \equiv \tilde{U}_{n:k}^I.$$

Also, let

$$\tilde{Q}_X^I(\cdot) \equiv Q_X(\tilde{Q}_U^I(\cdot)).$$

Instead of (3.1), we use

$$u_j^h(\alpha) \equiv k_j^h(\alpha)/(n+1), \quad u_j^l(\alpha) \equiv k_j^l(\alpha)/(n+1). \tag{A.1}$$

From A2.1, $X_i \overset{iid}{\sim} F$, so $U_i \equiv F(X_i) \overset{iid}{\sim} \text{Unif}(0,1)$, with order statistics $U_{n:k}$. Let $\mathbf{u} = (u_1, \ldots, u_J)'$ be a generic vector with all $u_j \in (0,1)$. For convenience let $u_0 \equiv 0$ and $u_{J+1} \equiv 1$. Given $\mathbf{u}$, for all $j \in \{1, 2, \ldots, J\}$,

$$k_j \equiv \lfloor (n+1)u_j \rfloor, \qquad \epsilon_j \equiv (n+1)u_j - k_j,$$

where the $\epsilon_j \in [0,1)$ are interpolation weights. Let $\Delta\mathbf{k}$ denote the $(J+1)$-vector with elements $\Delta k_j = k_j - k_{j-1}$. Let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_J)'$ be a fixed weight vector, and

$$\begin{aligned}
Y_j^{\mathbf{u}} &\equiv U_{n:k_j} \sim \text{Beta}(k_j, n+1-k_j), & \mathbf{Y}^{\mathbf{u}} &\equiv (Y_1^{\mathbf{u}}, \ldots, Y_J^{\mathbf{u}})', \\
\Delta\mathbf{Y}^{\mathbf{u}} &\equiv (Y_1, Y_2 - Y_1, \ldots, 1 - Y_J)' \sim \text{Dirichlet}(\Delta\mathbf{k}), & & \\
\Lambda_j^{\mathbf{u}} &\equiv U_{n:k_j+1} - U_{n:k_j} \sim \text{Beta}(1, n), & \boldsymbol{\Lambda}^{\mathbf{u}} &\equiv (\Lambda_1^{\mathbf{u}}, \ldots, \Lambda_J^{\mathbf{u}})', \\
\mathbb{W}^{\mathbf{u}} &\equiv \sqrt{n}\left(\sum_{j=1}^J \psi_j Q(Y_j^{\mathbf{u}}) - \sum_{j=1}^J \psi_j Q(u_j)\right), & & \\
\mathbb{W}_{\boldsymbol{\epsilon},\boldsymbol{\Lambda}}^{\mathbf{u}} &\equiv \mathbb{W}^{\mathbf{u}} + n^{1/2} \sum_{j=1}^J \epsilon_j \psi_j \Lambda_j^{\mathbf{u}} (Q'(u_j) + Q''(u_j)(Y_j^{\mathbf{u}} - u_j)), & &
\end{aligned} \tag{A.2}$$

where $Q(\cdot) = F^{-1}(\cdot)$ is the quantile function of interest, with first and second derivatives $Q'(\cdot)$ and $Q''(\cdot)$, and now $\tilde{\alpha}$ is explicitly written as a function of $\hat{\boldsymbol{\gamma}}$. For random variables with a $\mathbf{u}$ superscript like $\mathbf{Y}^{\mathbf{u}}$ and $\mathbb{W}^{\mathbf{u}}$, if the vector $\mathbf{u}$ is clear from context, then the corresponding superscript may be omitted. The values and distributions of the preceding variables are all understood to vary with n.

By construction, $\boldsymbol{\Delta\mathbf{k}}$ is a $(J+1)$-vector of natural numbers such that $\sum_{j=1}^{J+1} \Delta k_j = n+1$ and $k_j = \sum_{i=1}^{j} \Delta k_i$. In our applications to quantile inference, $\min_j\{\Delta k_j\} \to \infty$, and moreover all $\Delta k_j \asymp n$.

Define Condition $\star$ as satisfied by any value $\mathbf{y}$ if and only if

$$\max_j\{n\Delta k_j^{-1/2}|\Delta y_j - \Delta k_j/n|\} \leq 2\log(n). \qquad \text{Condition } \star$$

From GK Lemma 7(i), this implies the same $O(\log(n))$ bound when centring at the mode or mean of ΔY_j. GK Lemma 7(iv,v) shows that Condition $\star$ is violated with $O(n^{-2})$ probability, which is negligible for all our results. Essentially, Condition $\star$ plays the role of a cruder, weaker (but more broadly applicable in our case) law of iterated logarithm.

The following lemma is also useful, approximating the CI endpoint indices by standard normal quantiles.

Lemma A.1. *Let $z_{1-\alpha}$ denote the $(1-\alpha)$-quantile of a standard normal distribution. From the definitions in* (3.1) *and* (A.1)*, the values $u_j^l(\alpha)$ and $u_j^h(\alpha)$ can be approximated as*

$$u_j^l(\alpha) = \tau_j - n^{-1/2} z_{1-\alpha}\sqrt{\tau_j(1-\tau_j)} - \frac{2\tau_j - 1}{6n}(z_{1-\alpha}^2 + 2) + O(n^{-3/2}),$$
$$u_j^h(\alpha) = \tau_j + n^{-1/2} z_{1-\alpha}\sqrt{\tau_j(1-\tau_j)} - \frac{2\tau_j - 1}{6n}(z_{1-\alpha}^2 + 2) + O(n^{-3/2}).$$

Proof: The proof is in GK. □

We also use the following result from Theorem 2(ii) of GK, reproduced here for convenience. With $L_0 \equiv \sum_{j=1}^J \psi_j Q_X(u_j)$, uniformly over $\mathbf{u} = \boldsymbol{\tau} + o(1)$,

$$\sup_{K\in\mathbb{R}} \left| \mathrm{P}\left(\sum_{j=1}^J \psi_j \hat{X}^L_{n:(n+1)u_j} < L_0 + n^{-1/2}K \right) - \mathrm{P}\left(\sum_{j=1}^J \psi_j \tilde{X}^I_{n:(n+1)u_j} < L_0 + n^{-1/2}K \right) \right| = O(n^{-1}). \tag{A.3}$$

A.1. More general version of Theorem 2.1, with proof

Theorem A.2. *Let $L_0 = \sum_{j=1}^J \psi_j (Q_X(u_j) + Q_Y(u_j))$ and*

$$L_X^L \equiv \sum_{j=1}^J \psi_j \hat{X}^L_{n:(n+1)u_j}, \quad L_X^I \equiv \sum_{j=1}^J \psi_j \tilde{X}^I_{n:(n+1)u_j}. \tag{A.4}$$

Defining L_Y^L and L_Y^I similarly, under Assumptions A2.1 and A2.2, uniformly over $\mathbf{u} = \boldsymbol{\tau} + o(1)$,

$$\sup_{K\in\mathbb{R}} \left| \mathrm{P}(L_X^L + L_Y^L < L_0 + n^{-1/2}K) - \mathrm{P}(L_X^I + L_Y^I < L_0 + n^{-1/2}K) \right| = O(n^{-1}).$$

Proof: As in the proof in GK, we assume that the realised values of all random variables satisfy Condition $\star$. By application of GK Lemma 7(iv,v) this induces at most $O(n^{-2})$ error in our calculations, which is asymptotically negligible.

For any probability distribution $G(\cdot)$ and function $h(\cdot)$,

$$\left| \int h(x)\, dG(x) \right| \leq \sup_x |h(x)|,$$

so, using $L_X^L \perp\!\!\!\perp L_Y^L$ from A2.1,

$$\mathrm{P}(L_X^L + L_Y^L < K) = \int_{\mathbb{R}} \mathrm{P}(L_Y^L < K - x)\, dF_{L_X^L}(x)$$

$$
\begin{aligned}
&= \int_{\mathbb{R}} \mathrm{P}(L_Y^I < K - x)\, dF_{L_X^L}(x) + \overbrace{\int_{\mathbb{R}} \overbrace{(\mathrm{P}(L_Y^L < K - x) - \mathrm{P}(L_Y^I < K - x))}^{\text{uniformly } O(n^{-1}) \text{ by (A.3)}}\, dF_{L_X^L}(x)}^{\text{uniformly } O(n^{-1})} \\
&= \mathrm{P}(L_Y^I + L_X^L < K) + O(n^{-1}) \\
&= \int_{\mathbb{R}} \mathrm{P}(L_X^I < K - x)\, dF_{L_Y^I}(x) \\
&\quad + \overbrace{\int_{\mathbb{R}} \overbrace{(\mathrm{P}(L_X^L < K - x) - \mathrm{P}(L_X^I < K - x))}^{\text{uniformly } O(n^{-1}) \text{ by (A.3)}}\, dF_{L_Y^I}(x)}^{\text{uniformly } O(n^{-1})} + O(n^{-1}) \\
&= \mathrm{P}(L_X^I + L_Y^I < K) + O(n^{-1})
\end{aligned}
$$

uniformly over $K \in \mathbb{R}$ and $\mathbf{u} = \boldsymbol{\tau} + o(1)$. □

A.2. Proof of Theorem 3.1

Proof: Applying (A.3) to get the first equality below, actual CP is

$$
\begin{aligned}
&\mathrm{P}\left(\left\{\bigcap_{j=1}^{J}\{\hat{Q}_X^L(u_j^h(\tilde{\alpha}/2)) > Q(\tau_j)\}\right\}\bigcap\left\{\bigcap_{j=1}^{J}\{\hat{Q}_X^L(u_j^l(\tilde{\alpha}/2)) < Q(\tau_j)\}\right\}\right) \\
&= \overbrace{\mathrm{P}\left(\left\{\bigcap_{j=1}^{J}\{\tilde{Q}_X^I(u_j^h(\tilde{\alpha}/2)) > Q(\tau_j)\}\right\}\bigcap\left\{\bigcap_{j=1}^{J}\{\tilde{Q}_X^I(u_j^l(\tilde{\alpha}/2)) < Q(\tau_j)\}\right\}\right)}^{\text{use definition of } \tilde{Q}_X^I(\cdot)} + O(n^{-1}) \\
&= \overbrace{\mathrm{P}\left(\left\{\bigcap_{j=1}^{J}\{\tilde{Q}_U^I(u_j^h(\tilde{\alpha}/2)) > \tau_j\}\right\}\bigcap\left\{\bigcap_{j=1}^{J}\{\tilde{Q}_U^I(u_j^l(\tilde{\alpha}/2)) < \tau_j\}\right\}\right)}^{=1-\alpha \text{ by (3.3)}} + O(n^{-1}) \\
&= 1 - \alpha + O(n^{-1}).
\end{aligned}
$$

The application of (A.3) above follows from the Cramér–Wold device. □

A.3. Proof of Theorem 3.2 and Theorem A.3 (from main appendix)

We first state (and then prove) Theorem A.3, which is a more general version of Theorem 3.2. The X subscript is dropped for notational simplicity. Let $Q(\boldsymbol{\tau}) \equiv (Q(\tau_1), \ldots, Q(\tau_J))'$, and similarly for $Q'(\boldsymbol{\tau})$, $\widehat{Q'}(\boldsymbol{\tau})$, etc.

For quantile index vector $\boldsymbol{\tau} \in (0,1)^J$ and weights $\boldsymbol{\psi} \in \mathbb{R}^J$, we construct a CI for $D = \sum_{j=1}^{J} \psi_j Q(\tau_j)$. The theorem stated in the main text is for the special case with $\boldsymbol{\psi} = (-1, 1)'$ and $\boldsymbol{\tau} = (0.25, 0.75)'$.

Using (A.1), let

$$
\begin{aligned}
u_j^H(\alpha) &\equiv \mathbb{1}\{\psi_j > 0\} u_j^h(\alpha) + \mathbb{1}\{\psi_j < 0\} u_j^l(\alpha), \\
u_j^L(\alpha) &\equiv \mathbb{1}\{\psi_j > 0\} u_j^l(\alpha) + \mathbb{1}\{\psi_j < 0\} u_j^h(\alpha).
\end{aligned} \tag{A.5}
$$

In the notation of (A.5), the lower one-sided CI for D is

$$\left(-\infty, \sum_{j=1}^{J} \psi_j \hat{Q}_X^L(u_j^H(\tilde{\alpha}))\right), \tag{A.6}$$

where $\tilde{\alpha}$ implicitly depends on the $\widehat{Q'}(\tau_j)$ and satisfies

$$1-\alpha = \mathrm{P}\left(\sum_{j=1}^{J} \psi_j \widehat{Q'}(\tau_j)(\tilde{Q}_U^I(u_j^H(\tilde{\alpha})) - \tau_j) > 0\right). \tag{A.7}$$

For an upper one-sided CI, the analogues of (A.6) and (A.7) are

$$\left(\sum_{j=1}^{J} \psi_j \hat{Q}_X^L(u_j^L(\tilde{\alpha})), \infty\right), \quad 1-\alpha = \mathrm{P}\left(\sum_{j=1}^{J} \psi_j \widehat{Q'}(\tau_j)(\tilde{Q}_U^I(u_j^L(\tilde{\alpha})) - \tau_j) < 0\right). \tag{A.8}$$

Theorem A.3. *Let A2.1 and A2.2 hold.*

(a) The one-sided lower and upper CIs in (A.6) *and* (A.8) *have CPE of order* $O(n^{-1/2}\log(n))$ *if all* $\widehat{Q'}(\tau_j)$ *are estimated by* (3.5) *with smoothing parameters* m_j *having rate larger than* $n^{1/2}$ *and smaller than* $n^{3/4}$.

(b) The two-sided CI formed by the intersection of upper and lower one-sided $1-\alpha/2$ *CIs has CPE of order* $O(n^{-2/3}\log(n))$ *if all* $\widehat{Q'}(\tau_j)$ *are estimated by* (3.5) *with* $m_j \asymp n^{2/3}$.

(c) The asymptotic probabilities of excluding $D_n = \boldsymbol{\psi}'(Q(\boldsymbol{\tau}) + \boldsymbol{\kappa} n^{-1/2})$ *from lower one-sided (l), upper one-sided (u), and equal-tailed two-sided (t) CIs (i.e., asymptotic power of the corresponding hypothesis tests) are*

$$\mathcal{P}_n^l(D_n) \to \Phi(z_\alpha + S), \ \ \mathcal{P}_n^u(D_n) \to \Phi(z_\alpha - S), \ \ \mathcal{P}_n^t(D_n) \to \Phi(z_{\alpha/2} + S) + \Phi(z_{\alpha/2} - S),$$

where $S \equiv \boldsymbol{\psi}'\boldsymbol{\kappa}/\sqrt{\mathcal{V}_{\boldsymbol{\psi}}}$ *and*

$$\mathcal{V}_{\boldsymbol{\psi}} \equiv \sum_{i=1}^{J}\sum_{j=1}^{J} \psi_i \psi_j \frac{\min\{\tau_i, \tau_j\} - \tau_i\tau_j}{f(Q(\tau_i))f(Q(\tau_j))}. \tag{A.9}$$

Proof: We focus on the lower one-sided CI first; the upper one-sided results are entirely parallel.

To be more explicit about the dependence of $\tilde{\alpha}$ on the nuisance parameter, for a general value $\mathbf{g} = (g_1, \ldots, g_J)'$, let $\tilde{\alpha}(\mathbf{g})$ satisfy

$$1-\alpha = \mathrm{P}\left(\sum_{j=1}^{J} \psi_j g_j(\tilde{Q}_U^I(u_j^H(\tilde{\alpha}(\mathbf{g}))) - \tau_j) > 0\right). \tag{A.10}$$

This is (A.7) with g_j replacing $\widehat{Q'}(\tau_j)$. Correspondingly, we write $\hat{\mathbf{u}}$ for the vector of quantile indices selected to form CI endpoints given estimates

$$\hat{\boldsymbol{\gamma}} \equiv \widehat{Q'}(\boldsymbol{\tau}) \equiv (\widehat{Q'}(\tau_1), \ldots, \widehat{Q'}(\tau_J))', \tag{A.11}$$

and $\mathbf{u}_0^H$ is the vector of quantile indices that would be selected if the true

$$\boldsymbol{\gamma} \equiv Q'(\boldsymbol{\tau}) \tag{A.12}$$

were known. Note that (A.10) is invariant to scaling $\mathbf{g}$ by a constant scalar, so we could also divide by the first element in the vector to normalise the first element to be one, e.g., $\boldsymbol{\gamma} = (1, Q'(\tau_2)/Q'(\tau_1), \ldots, Q'(\tau_J)/Q'(\tau_1))'$, which will be used later. For the lower one-sided case,

$$\hat{\mathbf{u}}^H \equiv \{u_j^H(\tilde{\alpha}(\hat{\boldsymbol{\gamma}}))\}_{j=1}^J \text{ and } \mathbf{u}_0^H \equiv \{u_j^H(\tilde{\alpha}(\boldsymbol{\gamma}))\}_{j=1}^J,$$

and objects with L superscripts can be defined similarly for the upper one-sided case.

The CP of the lower one-sided CI can be decomposed into different components. We use notation from (A.2). We also use an implication of GK Lemma 8(i). Translating to our present context, that lemma states

$$|\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}_U^I(\hat{\mathbf{u}}^H)) - Q(\hat{\mathbf{u}}^H)) - \mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H}| = O(n^{-3/2}(\log(n))^3),$$

so for any $t \in \mathbb{R}$,

$$\begin{aligned}
&\mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}_U^I(\hat{\mathbf{u}}^H)) - Q(\hat{\mathbf{u}}^H)) > t \text{ and } \mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H} < t) \\
&= \overbrace{\mathrm{P}(t - O(n^{-3/2}(\log(n))^3) < \mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H} < t)}^{\text{use MVT}} \\
&\leq O(n^{-3/2}(\log(n))^3) \overbrace{\sup_{w \in [t - O(n^{-3/2}(\log(n))^3), t]} f_{\mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H}}(w)}^{=O(1) \text{ by GK Lemma 8(ii)}} \\
&= O(n^{-3/2}(\log(n))^3),
\end{aligned}$$

and switching the $<$ and $>$ leaves the rate unchanged. Thus,

$$\begin{aligned}
&|\mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}_U^I(\hat{\mathbf{u}}^H)) - Q(\hat{\mathbf{u}}^H)) > t) - \mathrm{P}(\mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H} > t)| \\
&\leq \overbrace{|\mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}_U^I(\hat{\mathbf{u}}^H)) - Q(\hat{\mathbf{u}}^H)) > t \text{ and } \mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H} < t)|}^{=O(n^{-3/2}(\log(n))^3)} \\
&\quad + \overbrace{|\mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}_U^I(\hat{\mathbf{u}}^H)) - Q(\hat{\mathbf{u}}^H)) < t \text{ and } \mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H} > t)|}^{=O(n^{-3/2}(\log(n))^3)} \\
&= O(n^{-3/2}(\log(n))^3). \qquad \text{(A.13)}
\end{aligned}$$

The lower one-sided CP is

$$\begin{aligned}
&\mathrm{P}(\boldsymbol{\psi}'\hat{Q}_X^L(\hat{\mathbf{u}}^H) > \boldsymbol{\psi}'Q(\boldsymbol{\tau})) \\
&= \overbrace{\mathrm{P}(\boldsymbol{\psi}'\tilde{Q}_X^I(\hat{\mathbf{u}}^H) > \boldsymbol{\psi}'Q(\boldsymbol{\tau})) + O(n^{-1})}^{\text{by (A.3)}} \\
&= \mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}_U^I(\hat{\mathbf{u}}^H)) - Q(\hat{\mathbf{u}}^H)) > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H))) + O(n^{-1}) \\
&= \mathrm{P}(\mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H))) \\
&\quad + (\mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}_U^I(\hat{\mathbf{u}}^H)) - Q(\hat{\mathbf{u}}^H)) > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H))) \\
&\qquad - \mathrm{P}(\mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H)))) \\
&\quad + O(n^{-1}) \\
&= \mathrm{P}(\mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\mathbf{u}_0^H} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)))
\end{aligned}$$

$$
\begin{aligned}
&+ \overbrace{\mathrm{P}(\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H))) - \mathrm{P}(\mathbb{W}^{\mathbf{u}_0^H}_{\mathbf{C},\boldsymbol{\Lambda}} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)))}^{E_h} \\
&+ \overbrace{(O(n^{-3/2}(\log(n))^3))}^{\text{by (A.13)}} + O(n^{-1}) \\
&= \overbrace{\mathrm{P}\left(\sum_{j=1}^{J} \psi_j \gamma_j (\tilde{Q}^I_U(u^H_{0,j}) - \tau_j) > 0\right)}^{=1-\alpha \text{ by (A.10)}} + T_h + E_h + O(n^{-1}), \qquad \text{(A.14)}
\end{aligned}
$$

where

$$
\begin{aligned}
T_h &= \mathrm{P}(\mathbb{W}^{\mathbf{u}_0^H}_{\mathbf{C},\boldsymbol{\Lambda}} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H))) - \mathrm{P}\left(\sum_{j=1}^{J} \psi_j \gamma_j (\tilde{Q}^I_U(u^H_{0,j}) - \tau_j) > 0\right) \\
&= \overbrace{\mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}^I_U(\mathbf{u}_0^H)) - Q(\mathbf{u}_0^H)) > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H))) + O(n^{-3/2}(\log(n))^3)}^{\text{by (A.13)}} \\
&\quad - \mathrm{P}\left(\sum_{j=1}^{J} \psi_j \gamma_j (\tilde{Q}^I_U(u^H_{0,j}) - \tau_j) > 0\right) \\
&= \mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(Q(\tilde{Q}^I_U(\mathbf{u}_0^H)) - Q(\boldsymbol{\tau})) > 0) - \mathrm{P}\left(\sum_{j=1}^{J} \psi_j \gamma_j (\tilde{Q}^I_U(u^H_{0,j}) - \tau_j) > 0\right) \\
&\quad + O(n^{-3/2}(\log(n))^3) \\
&= \mathrm{P}\left(\sum_{j=1}^{J} \psi_j (Q(\tilde{Q}^I_U(u^H_{0,j})) - Q(\tau_j)) > 0\right) - \mathrm{P}\left(\sum_{j=1}^{J} \psi_j Q'(\tau_j)(\tilde{Q}^I_U(u^H_{0,j}) - \tau_j) > 0\right) \\
&\quad + O(n^{-3/2}(\log(n))^3), \qquad \text{(A.15)}
\end{aligned}
$$

$$
E_h = E[\mathrm{P}(\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H)) \mid \hat{\boldsymbol{\gamma}}) - \mathrm{P}(\mathbb{W}^{\mathbf{u}_0^H}_{\mathbf{C},\boldsymbol{\Lambda}} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)) \mid \hat{\boldsymbol{\gamma}})]. \qquad \text{(A.16)}
$$

The term T_h captures the error in the first-order Taylor approximation of $Q(\tilde{Q}^I_U(u^H_{0,j})) - Q(\tau_j)$, and E_h captures estimation error in $\hat{\boldsymbol{\gamma}}$. The upper one-sided derivation yields similar terms, denoted T_l and E_l.

The proof of part (a) follows by applying Lemmas A.4 and A.5, which respectively have $T_h = O(n^{-1/2}\log(n))$ and $E_h = O(m^{-1}\log(n) + (m/n)^2)$ for common smoothing parameter rate m (so $m_j \asymp m$ for all j), and similarly for T_l and E_l, which correspond to the upper one-sided CI. Plugging these into (A.14) gives one-sided CPE equal to $O(n^{-1/2}\log(n)) + O(m^{-1}\log(n) + (m/n)^2)$. As long as $n^{1/2} \lesssim m \lesssim n^{3/4}$, the dominant CPE term is order $O(n^{-1/2}\log(n))$.

The proof of part (b) also follows by applying Lemmas A.4 and A.5, which also give $T_h + T_l = O(n^{-1}(\log(n))^2)$. Thus, CPE is $O(n^{-1}(\log(n))^2) + O(m^{-1}\log(n) + (m/n)^2)$. Now, the second term dominates, and it is minimised by $m \asymp n^{2/3}$, leaving CPE of order $O(n^{-2/3}\log(n))$.

The proof of part (c) remains. One-sided power against $H_0 : D_n = \boldsymbol{\psi}'(Q(\boldsymbol{\tau}) + \boldsymbol{\kappa}n^{-1/2})$ with $\boldsymbol{\psi}'\boldsymbol{\kappa} > 0$ is the probability that D_n is not contained in the lower one-sided CI. Below, $\tilde{u}_j$ comes from the mean value theorem and lies between τ_j and u_j^H. Since $u_j^H \to \tau_j$

by Lemma A.1, $\tilde{u}_j \to \tau_j$, so for large enough n, all $\tilde{u}_j$ lie within an arbitrarily small neighbourhood of τ_j and thus A2.2 uniformly bounds $Q''(\tilde{u}_j) = O(1)$. The CI exclusion probability is

$$\begin{aligned}
\mathcal{P}_n^l(D_n) &= \mathrm{P}\left(\sum_{j=1}^{J} \psi_j(\hat{Q}_X^L(u_j^H(\tilde{\alpha}_j)) - Q(\tau_j)) < n^{-1/2}\boldsymbol{\psi}'\boldsymbol{\kappa}\right) \\
&= \mathrm{P}(\boldsymbol{\psi}'\hat{Q}_X^L(\mathbf{u}^H(\tilde{\boldsymbol{\alpha}})) - \boldsymbol{\psi}'Q(\mathbf{u}^H(\tilde{\boldsymbol{\alpha}})) < n^{-1/2}\boldsymbol{\psi}'\boldsymbol{\kappa} - \boldsymbol{\psi}'(Q(\mathbf{u}^H(\tilde{\boldsymbol{\alpha}})) - Q(\boldsymbol{\tau}))) \\
&= \mathrm{P}(\sqrt{n}\boldsymbol{\psi}'(\hat{Q}_X^L(\mathbf{u}^H(\tilde{\boldsymbol{\alpha}})) - Q(\mathbf{u}^H(\tilde{\boldsymbol{\alpha}}))) \\
&\qquad < \boldsymbol{\psi}'\boldsymbol{\kappa} - \sqrt{n}\sum_{j=1}^{J} \psi_j(Q'(\tau_j)(u_j^H - \tau_j) + (1/2)\overbrace{Q''(\tilde{u}_j)}^{=O(1)}\ \overbrace{(u_j^H - \tau_j)^2}^{=O(n^{-1}) \text{ by Lemma A.1}}\ )) \\
&\qquad \overbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}^{\text{by GK Lemma 8}} \\
&= \Phi\left(\frac{\sum_{j=1}^{J} \psi_j\kappa_j - \psi_j Q'(\tau_j)\overbrace{\sqrt{n}(u_j^H(\tilde{\alpha}_j) - \tau_j)}^{\text{apply Lemma A.1}} + O(n^{-1/2})}{\sqrt{\hat{\mathcal{V}}_{\boldsymbol{\psi}}}}\right) + O(n^{-1/2}(\log(n))^3) \\
&= \Phi\left(\frac{\boldsymbol{\psi}'\boldsymbol{\kappa}}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}}}} - \overbrace{\frac{1}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}}}}\sum_{j=1}^{J} \psi_j Q'(\tau_j) z_{1-\tilde{\alpha}_j}\sqrt{\tau_j(1-\tau_j)}}^{\to z_{1-\alpha} \text{ to control size when } \boldsymbol{\kappa}=\mathbf{0}} + O(n^{-1/2})\right) + O(n^{-1/2}(\log(n))^3) \\
&\to \Phi\left(\frac{\boldsymbol{\psi}'\boldsymbol{\kappa}}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}}}} - z_{1-\alpha}\right) = \Phi\left(\frac{\boldsymbol{\psi}'\boldsymbol{\kappa}}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}}}} + z_{\alpha}\right),
\end{aligned}$$

where

$$\hat{\mathcal{V}}_{\boldsymbol{\psi}} \equiv \sum_{i=1}^{J}\sum_{j=1}^{J} \psi_i\psi_j \frac{\min\{u_i^H(\tilde{\alpha}_i), u_j^H(\tilde{\alpha}_j)\} - u_i^H(\tilde{\alpha}_i)u_j^H(\tilde{\alpha}_j)}{f(Q(u_i^H(\tilde{\alpha}_i)))f(Q(u_j^H(\tilde{\alpha}_j)))} \to \mathcal{V}_{\boldsymbol{\psi}}.$$

These results are invariant to choosing a single $\tilde{\alpha}$ or different $\tilde{\alpha}_j$ because the term involving the $\tilde{\alpha}_j$ must equal z_α in order to control size. In the special case $\psi = 1$ for a single quantile, then $\tilde{\alpha} = \alpha$, and the result reduces to the result in GK Theorem 4.

The upper one-sided case follows similarly.

For the two-sided case, since the two-sided CI is the intersection of the upper and lower one-sided $1 - \alpha/2$ CIs, the exclusion probability is

$$\begin{aligned}
\mathcal{P}_n^t(D_n) &= \mathrm{P}\left(D_n \notin \left[\sum_{j=1}^{J} \psi_j\hat{Q}_X^L(u_j^L(\tilde{\alpha}/2)), \sum_{j=1}^{J} \psi_j\hat{Q}_X^L(u_j^H(\tilde{\alpha}/2))\right]\right) \\
&= \overbrace{\mathrm{P}\left(\sum_{j=1}^{J} \psi_j\hat{Q}_X^L(u_j^H(\tilde{\alpha}/2)) < D_n\right)}^{\mathcal{P}_n^l(D_n) \text{ with } \alpha/2} + \overbrace{\mathrm{P}\left(\sum_{j=1}^{J} \psi_j\hat{Q}_X^L(u_j^L(\tilde{\alpha}/2)) > D_n\right)}^{\mathcal{P}_n^l(D_n) \text{ with } \alpha/2} \\
&\to \Phi\left(z_{\alpha/2} + \frac{\boldsymbol{\psi}'\boldsymbol{\kappa}}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}}}}\right) + \Phi\left(z_{\alpha/2} - \frac{\boldsymbol{\psi}'\boldsymbol{\kappa}}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}}}}\right). \qquad \square
\end{aligned}$$

A.3.1. CPE from Taylor Approximations: T_h, T_l

Lemma A.4. *Under the assumptions of Theorem A.3, the term T_h from* (A.14) *is of order $O(n^{-1/2}\log(n))$, and similarly $T_l = O(n^{-1/2}\log(n))$ for the corresponding upper one-sided term. Additionally, $T_h + T_l = O(n^{-1}(\log(n))^2)$.*

A sketch of the proof follows; the full proof is in the supplemental appendix. The intuition is that we are computing the remainder term from a linear approximation of the quantile function, as described in the main text. Because $u_{0,j}^H = \tau_j + O(n^{-1/2})$ by Lemma A.1, the quadratic remainder term can be shown to be nearly $O(n^{-1/2})$, using the Assumption A2.2 bound on quantile function derivatives uniformly over a neighbourhood of τ_j. For the two-sided result, the upper and lower (T_h and T_l) $n^{-1/2}$ terms cancel due to (approximate) symmetry: symmetry of the asymptotic normal distribution of certain random variables, and symmetry of the $n^{-1/2}$ term in Lemma A.1. This leaves a nearly $O(n^{-1})$ remainder.

Consider T_h, with $u_{0,j}^H = u_j^H(\tilde{\alpha}(\boldsymbol{\gamma}_0))$. The relevant Taylor expansion is

$$Q(\tilde{Q}_U^I(u_{0,j}^H)) - Q(\tau_j) = Q'(\tau_j)(\tilde{Q}_U^I(u_{0,j}^H) - \tau_j) + \frac{1}{2}Q''(\tau_j)(\tilde{Q}_U^I(u_{0,j}^H) - \tau_j)^2 + \frac{1}{6}\overbrace{\underbrace{Q'''(\tilde{u}_j)}_{=O(1)\text{ uniformly by A2.2}}\overbrace{(\tilde{Q}_U^I(u_{0,j}^H) - \tau_j)^3}^{=O(n^{-3/2}(\log(n))^{3/2})}}^{=O(n^{-3/2}(\log(n))^{3/2})}, \tag{A.17}$$

where $\tilde{u}_j \to \tau_j$ since $u_{0,j}^H \to \tau_j$ and thus A2.2 applies. The $\log(n)$ terms arise from applying Condition $\star$ to get $O(\cdot)$ instead of $O_p(\cdot)$ bounds on terms; the probability of the corresponding $O(\cdot)$ bounds not uniformly holding is negligible. In sum, the linear approximation captures the first term in (A.17), while the third is smaller-order, so the question becomes: what is the effect of ignoring the quadratic term?

Using (A.17), we can (eventually) decompose

$$\begin{aligned}
T_h &= T_{H,1} - T_{H,2} + O(n^{-1}(\log(n))^{3/2}),\\
T_{H,1} &\equiv \mathrm{P}\left(\sum_{j=1}^J \psi_j Q'(\tau_j)(\Delta_j^H + D_j^H) > -\frac{n^{-1/2}}{2}\sum_{j=1}^J \psi_j Q''(\tau_j)(\Delta_j^H + D_j^H)^2\right),\\
T_{H,2} &\equiv \mathrm{P}\left(\sum_{j=1}^J \psi_j Q'(\tau_j)(\Delta_j^H + D_j^H) > 0\right),\\
\Delta_j^H &\equiv \sqrt{n}(\tilde{Q}_U^I(u_j^H(\tilde{\alpha})) - u_j^H(\tilde{\alpha})),\\
D_j^H &\equiv \sqrt{n}(u_j^H(\tilde{\alpha}) - \tau_j).
\end{aligned}$$

The overall T_h captures the effect of the additional $n^{-1/2}$ term after the $>$ in $T_{H,1}$. From equation (A.4) in GK Lemma 7(iii), the PDFs of the corresponding vectors $\boldsymbol{\Delta}^H \equiv (\Delta_1^H, \ldots, \Delta_J^H)'$ and $\boldsymbol{\Delta}^L \equiv (\Delta_1^L, \ldots, \Delta_J^L)'$ are asymptotically normal, which makes the probabilities easier to simplify and compare. Further, since both $u_j^H \to \tau_j$ and $u_j^L \to \tau_j$, the two vectors have the same first-order asymptotic distribution. It also helps to restrict attention to the case where all $|\Delta_j| < \sqrt{2\log(n)}$, which has negligible probability of being violated.

In the one-sided case, T_h is essentially the probability that a (non-degenerate) nor-

mal random variable falls in an interval of length $O(n^{-1/2}\log(n))$, which is the same $O(n^{-1/2}\log(n))$.

The two-sided result requires more precision. We solve for the roots of the quadratic inside $T_{H,1}$, seeing it as a function of Δ_1^H conditional on other Δ_j^H, $j = 2, \ldots, J$. One root is so large that the probability of exceeding it is negligible, so it reduces to the probability of a single inequality. The conditional probability (given the other Δ_j^H) can be written out, and then integrated over the multivariate normal joint distribution of $(\Delta_2^H, \ldots, \Delta_J^H)$. Parallel steps apply to $T_{L,1}$, yielding a very similar integral. When adding $T_h + T_l$, the symmetry from Lemma A.1 and the symmetry of the normal PDF cause the $n^{-1/2}$ terms to cancel, leaving only the nearly $O(n^{-1})$ remainder terms.

A.3.2. CPE from nuisance parameter estimation error: E_h, E_l

Lemma A.5. *Under the assumptions of Theorem A.3, $E_h = O(m^{-1}\log(n) + (m/n)^2)$ in* (A.14)*, and similarly for the corresponding upper one-sided term, $E_l = O(m^{-1}\log(n) + (m/n)^2)$, where m is the common rate of smoothing parameters, $m_j \asymp m$ for all j.*

A sketch of the proof follows; the full proof is in the supplemental appendix. The general idea is to apply the mean value theorem repeatedly, where we can compute bounds on the derivatives (and other terms) and eventually pull out $\widehat{Q'}(\boldsymbol{\tau}) - Q'(\boldsymbol{\tau})$, the nuisance parameter estimation error. The eventual result is essentially the sum of the bias and variance of the sparsity estimator, $\widehat{Q'}(\boldsymbol{\tau})$, which are given in Bloch and Gastwirth (1968). The smoothing parameter m affects the bias-variance tradeoff; larger m (more smoothing) decreases variance but increases bias, and vice-versa.

First, decompose

$$
\begin{aligned}
E_h &= E_{\hat{\boldsymbol{\gamma}}}[\mathrm{P}(\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H)) \mid \hat{\boldsymbol{\gamma}}) - \mathrm{P}(\mathbb{W}^{\mathbf{u}_0^H}_{\mathbf{C},\boldsymbol{\Lambda}} > \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)) \mid \hat{\boldsymbol{\gamma}})] \\
&= E_{\hat{\boldsymbol{\gamma}}}[\mathrm{P}(\mathbb{W}^{\mathbf{u}_0^H}_{\mathbf{C},\boldsymbol{\Lambda}} < \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)) \mid \hat{\boldsymbol{\gamma}}) - \mathrm{P}(\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}} < \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H)) \mid \hat{\boldsymbol{\gamma}})] \\
&= \overbrace{E_{\hat{\boldsymbol{\gamma}}}[\mathrm{P}(\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}} < \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)) \mid \hat{\boldsymbol{\gamma}}) - \mathrm{P}(\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}} < \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H)) \mid \hat{\boldsymbol{\gamma}})]}^{E_h^1} \\
&\quad + \overbrace{E_{\hat{\boldsymbol{\gamma}}}[\mathrm{P}(\mathbb{W}^{\mathbf{u}_0^H}_{\mathbf{C},\boldsymbol{\Lambda}} < \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)) \mid \hat{\boldsymbol{\gamma}}) - \mathrm{P}(\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}} < \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)) \mid \hat{\boldsymbol{\gamma}})]}^{E_h^2}.
\end{aligned}
$$

It can be shown that E_h^2 is smaller-order (i.e., not the rate-limiting term), so here we focus on E_h^1. The intuition for E_h^2 being small is that $\hat{u}_j^H$ and $u_{0,j}^H$ are very close to each other, as will be seen below.

For E_h^1, GK Lemma 8(ii) is helpful: uniformly over any $\mathbf{u} = \boldsymbol{\tau} + o(1)$, which includes all possible $\hat{\mathbf{u}}^H = \boldsymbol{\tau} + O(n^{-1/2})$, the PDF of $\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}}$ is approximately (up to a multiplicative error) that of a mean-zero normal distribution with variance $\mathcal{V}_{\boldsymbol{\psi}}^{\hat{\mathbf{u}}^H}$, which is the same as $\mathcal{V}_{\boldsymbol{\psi}}$ in the statement of the theorem but with $\boldsymbol{\tau}$ replaced by $\hat{\mathbf{u}}^H$. This $\mathcal{V}_{\boldsymbol{\psi}}^{\hat{\mathbf{u}}^H}$ is then approximated by $\mathcal{V}_{\boldsymbol{\psi}}^{0,H}$, based on $\mathbf{u}_0^H$. Using this PDF approximation and other approximations, given a value of $\hat{\mathbf{u}}^H$, the mean value theorem gives

$$
E_h^1(\hat{\mathbf{u}}^H) = \int_{\sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H))}^{\sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H))} f_{\mathbb{W}^{\hat{\mathbf{u}}^H}_{\mathbf{C},\boldsymbol{\Lambda}}}(w)\, dw
$$

$$
\begin{aligned}
&= (\sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\mathbf{u}_0^H)) - \sqrt{n}\boldsymbol{\psi}'(Q(\boldsymbol{\tau}) - Q(\hat{\mathbf{u}}^H)))f_{\mathbb{W}_{\mathbf{C},\boldsymbol{\Lambda}}^{\hat{\mathbf{u}}^H}}(\tilde{w}) \\
&= \sqrt{n}\boldsymbol{\psi}'(Q(\hat{\mathbf{u}}^H) - Q(\mathbf{u}_0^H))\overbrace{\phi_{\mathcal{V}_{\boldsymbol{\psi}}^{0,H}}(\tilde{w})(1 + O(n^{-1/2}(\log(n))^3))}^{\text{normal PDF approximation}} \\
&= \sqrt{n}\boldsymbol{\psi}'\overbrace{(\widehat{Q'}(\boldsymbol{\tau}) - Q'(\boldsymbol{\tau}))O(1)O(n^{-1/2})O(1)}^{\text{by (A.18) and (A.19)}}\phi_{\mathcal{V}_{\boldsymbol{\psi}}^{0,H}}(\tilde{w})(1 + O(n^{-1/2}(\log(n))^3)) \\
&= \boldsymbol{\psi}'(\widehat{Q'}(\boldsymbol{\tau}) - Q'(\boldsymbol{\tau}))\phi_{\mathcal{V}_{\boldsymbol{\psi}}^{0,H}}(\tilde{w})(1 + O(n^{-1/2}(\log(n))^3))
\end{aligned}
$$

The following approximations were used above. Using mean value expansions where $\tilde{a}$ is between $\tilde{\alpha}(\widehat{Q'}(\boldsymbol{\tau}))$ and $\tilde{\alpha}(Q'(\boldsymbol{\tau}))$, and each element of $\tilde{Q}'(\boldsymbol{\tau})$ is between the corresponding elements of the true and estimated vectors, using $u_j^{h\prime}(\tilde{a}) \equiv \frac{d}{d\tilde{\alpha}}u_j^h(\tilde{\alpha}) = O(n^{-1/2})$,

$$
\begin{aligned}
\hat{u}_j^h - u_{0,j}^h &\equiv u_j^h(\tilde{\alpha}(\widehat{Q'}(\boldsymbol{\tau}))) - u_j^h(\tilde{\alpha}(Q'(\boldsymbol{\tau}))) \\
&= (\tilde{\alpha}(\widehat{Q'}(\boldsymbol{\tau})) - \tilde{\alpha}(Q'(\boldsymbol{\tau})))u_j^{h\prime}(\tilde{a}) \\
&= \overbrace{(\widehat{Q'}(\boldsymbol{\tau}) - Q'(\boldsymbol{\tau}))'}^{=O(m^{-1/2}\log(n)+m/n)}\overbrace{\tilde{\alpha}'(\tilde{Q}'(\boldsymbol{\tau}))}^{=O(1)}\overbrace{u_j^{h\prime}(\tilde{a})}^{=O(n^{-1/2})} \qquad \text{(A.18)} \\
&= O(m^{-1/2}n^{-1/2}\log(n) + mn^{-3/2}), \\
Q(\hat{u}_j^h) - Q(u_{0,j}^h) &= (\hat{u}_j^h - u_{0,j}^h)Q'(\tilde{u}) = O(m^{-1/2}n^{-1/2}\log(n) + mn^{-3/2}), \qquad \text{(A.19)}
\end{aligned}
$$

where $\tilde{u}_j$ is between $\hat{u}_j^h$ and $u_{0,j}^h$ and thus $\tilde{u}_j \to \tau_j$, so for large enough n, $Q'(\tilde{u})$ is uniformly bounded by A2.2; and similarly with $\hat{u}_j^l$ and $u_{0,j}^l$. The bound on $\widehat{Q'}(\boldsymbol{\tau}) - Q'(\boldsymbol{\tau})$ is under Condition $\star$. The bound on the derivative of $\tilde{\alpha}$ as a function of the sparsity estimator seems intuitive since $\tilde{\alpha}$ varies over a subset of $(0,1)$ while the argument varies over $(0,\infty)$, but it takes much work using the implicit function theorem to show formally. The bound on the derivative of u_j^h as a function of $\tilde{\alpha}$ is also intuitive since $u^h = \tau + n^{-1/2}\Phi^{-1}(1-\tilde{\alpha})\sqrt{\tau(1-\tau)} + O(n^{-1})$ from Lemma A.1.

With some additional work,

$$
E_h^1 = O(E[\overbrace{(1 + O(m^{-1/2}\log(n) + mn^{-1}))}^{A}\overbrace{(\widehat{Q'}(\boldsymbol{\tau}) - Q'(\boldsymbol{\tau}))}^{B}]).
$$

Now,

$$
\begin{aligned}
E[AB] &= \mathrm{Cov}(A,B) + E[A]E[B], \\
|\mathrm{Cov}(A,B)| &= |\mathrm{Corr}(A,B)\sqrt{\mathrm{Var}(A)\,\mathrm{Var}(B)}| \le \sqrt{\mathrm{Var}(A)\,\mathrm{Var}(B)}.
\end{aligned}
$$

From equations (2.5) and (2.6) in Bloch and Gastwirth (1968, p. 1084), $E[B] = O(m^2/n^2)$ and $\mathrm{Var}(B) = O(m^{-1})$. Also, $E[A] = O(1)$ and $\mathrm{Var}(A) = O(m^{-1}(\log(n))^2 + m^2n^{-2})$, so

$$
\begin{aligned}
|E[A])| &= \sqrt{O(m^{-1}(\log(n))^2 + m^2n^{-2})O(m^{-1})} + O(1)O(m^2/n^2) \\
&= O(m^{-1}\log(n) + m^2/n^2).
\end{aligned}
$$

Using $m \asymp n^{2/3}$ attains the (nearly) minimum rate of $O(n^{-2/3}\log(n))$. With any $n^{1/2} \lesssim m \lesssim n^{3/4}$, the rate is no greater than $T_h = O(n^{-1/2}\log(n))$.

A.4. Theorem for CI for difference of linear combination of quantiles (and QD)

We first state (and then prove) Theorem A.6, where the object of interest is

$$D = \sum_{j=1}^{J} \psi_j (Q_Y(\tau_j) - Q_X(\tau_j)).$$

In the special case $J = 1$, D is the τ-QD. If $\boldsymbol{\psi} = (-1, 1)'$, then D is a difference of IQRs.

For a lower one-sided CI, using (2.7) and (A.5), $\tilde{\alpha}$ satisfies

$$1-\alpha = \mathrm{P}\left(\sum_{j=1}^{J} \psi_j (\widehat{Q'_Y}(\tau_j)(\tilde{Q}^I_{U_y}(u^H_{y,j}(\tilde{\alpha})) - \tau_j) - \widehat{Q'_X}(\tau_j)(\tilde{Q}^I_{U_x}(u^L_{x,j}(\tilde{\alpha})) - \tau_j)) > 0 \right). \quad \text{(A.20)}$$

The $1 - \alpha$ CI is then

$$\left(-\infty, \sum_{j=1}^{J} \psi_j (\hat{Q}^L_Y(u^H_{y,j}(\tilde{\alpha})) - \hat{Q}^L_X(u^L_{x,j}(\tilde{\alpha}))) \right). \quad \text{(A.21)}$$

For an upper one-sided CI, the analogues of (A.20) and (A.21) are

$$1 - \alpha = \mathrm{P}\left(\sum_{j=1}^{J} \psi_j (\widehat{Q'_Y}(\tau_j)(\tilde{Q}^I_{U_y}(u^L_{y,j}(\tilde{\alpha})) - \tau_j) - \widehat{Q'_X}(\tau_j)(\tilde{Q}^I_{U_x}(u^H_{x,j}(\tilde{\alpha})) - \tau_j)) < 0 \right), \quad \text{(A.22)}$$

$$\left(\sum_{j=1}^{J} \psi_j (\hat{Q}^L_Y(u^L_{y,j}(\tilde{\alpha})) - \hat{Q}^L_X(u^H_{x,j}(\tilde{\alpha}))), \infty \right). \quad \text{(A.23)}$$

Theorem A.6. *Let Assumptions A2.1 and A2.2 hold.*

(a) The one-sided CIs in (A.21) *and* (A.23) *both have CPE of order* $O(n^{-1/2}\log(n))$ *if all* $\widehat{Q'_X}(\tau_j)$ *and* $\widehat{Q'_Y}(\tau_j)$ *are estimated by* (3.5) *with smoothing parameters* $m_{x,j}$ *and* $m_{y,j}$ *having rates larger than* $n^{1/2}$ *and smaller than* $n^{3/4}$.

(b) Two-sided CIs, formed by the intersection of upper and lower one-sided $1 - \alpha/2$ *CIs, have CPE of order* $O(n^{-2/3}\log(n))$ *if all* $\widehat{Q'_X}(\tau_j)$ *and* $\widehat{Q'_Y}(\tau_j)$ *are estimated by* (3.5) *with* $m_{x,j} \asymp n^{2/3}$ *and* $m_{y,j} \asymp n^{2/3}$.

(c) The asymptotic probabilities of excluding $D_n = \boldsymbol{\psi}'(Q_Y(\boldsymbol{\tau}) - Q_X(\boldsymbol{\tau}) + \boldsymbol{\kappa} n_y^{-1/2})$ *from lower one-sided (l), upper one-sided (u), and equal-tailed two-sided (t) CIs (i.e., asymptotic power of the corresponding hypothesis tests) are*

$$\mathcal{P}^l_n(D_n) \to \Phi(z_\alpha + S), \;\; \mathcal{P}^u_n(D_n) \to \Phi(z_\alpha - S), \;\; \mathcal{P}^t_n(D_n) \to \Phi(z_{\alpha/2} + S) + \Phi(z_{\alpha/2} - S),$$

where $S \equiv \boldsymbol{\psi}'\boldsymbol{\kappa}/\sqrt{\mathcal{V}_{\boldsymbol{\psi},x} + \delta^2 \mathcal{V}_{\boldsymbol{\psi},y}}$, *and* $\mathcal{V}_{\boldsymbol{\psi},x}$ *and* $\mathcal{V}_{\boldsymbol{\psi},y}$ *are as defined in Theorem A.3 for the* X *and* Y *population distributions, respectively.*

The proof closely parallels that of Theorem A.3, so we defer it to the supplemental appendix. The overall CPE is again decomposed into a term from using a linear Taylor approximation, a term from the estimation error of the nuisance parameter, and a smaller-order remainder; the same rates apply to each. Since the second sample is assumed independent of the first, there is no additional dependence structure to consider, just two

independent Dirichlet distributions. It is more cumbersome due to having more terms and different sample sizes, but the intuition and strategy are the same.

A.5. Proof of Theorem 4.1

Proof: This proof shares the same structure as that of GK Theorem 6, with four main components needed. To establish the order of the CPE term due to applying the unconditional method to the local sample (i.e., $\mathrm{CPE_U}$), first, it must be shown that $N_n \asymp nb^d$ almost surely, and second, A2.2 must be satisfied uniformly by the "local PDF" from which the local sample is drawn (which changes with n). Third, the order of the CPE due to bias (i.e., $\mathrm{CPE_{Bias}}$) is needed. Fourth, the sum $\mathrm{CPE_U} + \mathrm{CPE_{Bias}}$ can be minimised to derive the CPE-optimal bandwidth rate and corresponding CPE. Steps two and three can be taken directly from the proof of GK Theorem 6; we comment on them but refer to the other paper for details. As in Chaudhuri (1991), we consider a deterministic bandwidth sequence, leaving treatment of a random (data-dependent) bandwidth to future work.

First, although it is random, the local sample size N_n is almost surely of order nb^d (exactly, not just $O(nb^d)$), as shown in the proof of GK Theorem 6, following the argument in Chaudhuri (1991, proof of Thm. 3.1, p. 769). Specifically, using Bernstein's Inequality and the Borel–Cantelli Lemma, it can be shown that there exist constants c_1 and c_2 such that $c_1 nb^d \leq N_n \leq c_2 nb^d$ for large enough n with probability one. For joint or CIQR inference, because the same bandwidth is used at each quantile, there is a single local sample and single N_n, so the prior result in Chaudhuri (1991) applies directly. For CQD inference, there are two local samples, so some additional arguments are required. For the $T_i = 0$ subsample, define the event $A_{n0} \equiv \{c_{01} nb_0^d \leq N_{n0} \leq c_{02} nb_0^d\}$, and similarly let $A_{n1} \equiv \{c_{11} nb_1^d \leq N_{n1} \leq c_{12} nb_1^d\}$. Let $A_n \equiv A_{n0} \cap A_{n1}$. We want to show that with probability one, A_n occurs for all n larger than some value n_0, i.e., $\mathrm{P}(\liminf A_n) = 1$. The Borel–Cantelli Lemma gives this conclusion if $\sum_{n=1}^{\infty}(1 - \mathrm{P}(A_n)) < \infty$. Using probability/set identities and inequalities, writing A^{c} for the complement of event A,

$$\begin{aligned}\mathrm{P}(A_{n0} \cap A_{n1}) &= 1 - \overbrace{\mathrm{P}(A_{n0}^{\mathsf{c}} \cup A_{n1}^{\mathsf{c}})}^{\leq \mathrm{P}(A_{n0}^{\mathsf{c}}) + \mathrm{P}(A_{n1}^{\mathsf{c}})} \geq 1 - \mathrm{P}(A_{n0}^{\mathsf{c}}) - \mathrm{P}(A_{n1}^{\mathsf{c}}) \\ &= 1 - (1 - \mathrm{P}(A_{n0})) - (1 - \mathrm{P}(A_{n1})) \\ &= \mathrm{P}(A_{n0}) + \mathrm{P}(A_{n1}) - 1. \end{aligned} \tag{A.24}$$

The probabilities in the RHS of (A.24) are bounded by the application of Bernstein's Inequality in Chaudhuri (1991). The specific constants involved will change since $\mathrm{P}(T_i = 1)$ now enters the binomial probability parameter, but since $\mathrm{P}(T_i = 1)$ is fixed and strictly between zero and one (from A4.1), the rates are the same. So, there exist constants $c_{03}, c_{04}, c_{13}, c_{14} > 0$ such that for all n,

$$\begin{aligned} \mathrm{P}(A_{n0}) \geq 1 - c_{03}\exp(-c_{04} nb_0^d), \quad \mathrm{P}(A_{n1}) \geq 1 - c_{13}\exp(-c_{14} nb_1^d), \\ 1 - \mathrm{P}(A_{n0}) \leq c_{03}\exp(-c_{04} nb_0^d), \quad 1 - \mathrm{P}(A_{n1}) \leq c_{13}\exp(-c_{14} nb_1^d), \end{aligned} \tag{A.25}$$

Altogether,

$$\sum_{n=1}^{\infty}(1 - \mathrm{P}(A_n)) = \sum_{n=1}^{\infty}(1 - \overbrace{\mathrm{P}(A_{n0} \cap A_{n1})}^{\text{use (A.24)}})$$

$$
\begin{aligned}
&\leq \sum_{n=1}^{\infty}(1-(\mathrm{P}(A_{n0})+\mathrm{P}(A_{n1})-1))\\
&= \sum_{n=1}^{\infty}(\overbrace{1-\mathrm{P}(A_{n0})}^{\text{use (A.25)}}+\overbrace{1-\mathrm{P}(A_{n1})}^{\text{use (A.25)}})\\
&\leq \sum_{n=1}^{\infty} c_{03}\exp(-c_{04}nb_0^d)+\sum_{n=1}^{\infty} c_{13}\exp(-c_{14}nb_1^d)\\
&= c_{03}\sum_{n=1}^{\infty}\exp(-c_{04}\overbrace{nb_0^d}^{\gtrsim(\log(n))^2 \text{ by A4.6}})+c_{13}\sum_{n=1}^{\infty}\exp(-c_{14}\overbrace{nb_1^d}^{\gtrsim(\log(n))^2 \text{ by A4.6}})\\
&\leq c_{03}\sum_{n=1}^{\infty}\exp(-c_{04}(\log(n))^2)+c_{13}\sum_{n=1}^{\infty}\exp(-c_{14}(\log(n))^2),
\end{aligned}
$$

and both sums are finite by comparison with

$$\sum_{n=1}^{\infty}\exp(-2\log(n))=\sum_{n=1}^{\infty}\exp(\log(n^{-2}))=\sum_{n=1}^{\infty}n^{-2}=\pi^2/6.$$

This means the summability condition from the Borel–Cantelli Lemma is satisfied, so as desired $\mathrm{P}(\liminf A_n)=1$, and $N_{n0}\asymp nb_0^d$ and $N_{n1}\asymp nb_1^d$, where b_0 and b_1 are the same rate by assumption.

Second, in addition to having N_n instead of n, having a local distribution that changes with n is another difference with the unconditional setting. Specifically, these local PDFs must uniformly satisfy A2.2 for large enough n. This is shown to be true in the proof of GK Theorem 6, by using $b\to 0$ (and thus $C_b\to\{\mathbf{w}_0\}$) from A4.6 along with the assumed smoothness from A4.2–A4.5. For CQD inference, since the same assumptions hold conditional on $T=0$ and $T=1$ alike, the same argument applies. Consequently, the CPE due to application of the unconditional method to the local sample ($\mathrm{CPE}_{\mathrm{U}}$ in the main text) is obtained by replacing n with nb^d in the unconditional results. For example, two-sided QD or IQR CIs have unconditional CPE of order $O(n^{-2/3}\log(n))$; for CQD and CIQR, replacing n with $N_n\asymp nb^d$ leaves $O((nb^d)^{-2/3}\log(nb^d))$, where $b_0,b_1\asymp b$ is the common bandwidth rate for the CQD case.

Third, the other component of overall CPE is from bias. In the proof of GK Theorem 6, this is $O(N_n^{1/2}b^2)=O(n^{1/2}b^{2+d/2})$: the bias is $O(b^2)$, the CI endpoint PDF is proportional to $N_n^{1/2}$, and (using the MVT) their product gives the order of CPE due to bias. For two-sided inference on a single conditional quantile, GK show that some cancellation occurs to reduce the order of magnitude, but this does not seem to occur for CQD or CIQR inference. If such cancellation did occur, then CPE would be even better (smaller) than in the results given here.

Fourth, we derive the CPE-optimal bandwidth rates and optimal CPE rates for all conditional methods. For joint inference on multiple conditional quantiles, whether one-sided or two-sided, the CPE from Theorem 3.1 is $O(N_n^{-1})$, so setting $\mathrm{CPE}_{\mathrm{U}}=\mathrm{CPE}_{\mathrm{Bias}}$ gives

$$N_n^{-1}\asymp N_n^{1/2}b^2 \implies (nb^d)^{3/2}\asymp b^{-2} \implies b^*\asymp n^{-3/(4+3d)},$$

and the overall CPE is

$$O((n(b^*)^d)^{-1}) = O((n^{1-3d/(4+3d)})^{-1}) = O(n^{-4/(4+3d)}).$$

For one-sided CIQR or CQD inference (or more general linear combinations or differences thereof), the CPE from Theorem A.3 or Theorem A.6 is $O(N_n^{-1/2}\log(N_n))$. Ignoring the $\log(N_n)$ for simplicity,

$$N_n^{-1/2} \asymp N_n^{1/2}b^2 \implies (nb^d)^{-1} \asymp b^2 \implies b^* \asymp n^{-1/(2+d)},$$

and the overall CPE is $O((n(b^*)^d)^{-1/2}\log(n(b^*)^d)) = O(n^{-1/(2+d)}\log(n))$. For two-sided CIQR or CQD inference (or the more general versions), the CPE from Theorem A.3 or Theorem A.6 is $O(N_n^{-2/3}\log(N_n))$. Ignoring the $\log(N_n)$ for simplicity,

$$N_n^{-2/3} = N_n^{1/2}b^2 \implies b^* \asymp n^{-7/(12+7d)}, \text{CPE} = O(n^{-8/(12+7d)}\log(n)). \qquad \square$$

B. NUISANCE PARAMETER ESTIMATION AND PLUG-IN BANDWIDTH DETAILS

B.1. Nuisance parameter estimation

Selection of $\tilde{\alpha}$ in Sections 3.2 and 3.3 requires preliminary estimation of derivatives of the quantile function. We recommend the "quantile spacing" estimator first proposed by Siddiqui (1960), given earlier in (3.5). In practice, results are often very similar when using fractional order statistics in (3.5) instead of rounding to integers, or even using a kernel density estimator instead, but we do not explore those here. Below we derive a rule for bandwidth selection that ensures an optimal order of CPE, but our results are also not particularly sensitive to the bandwidth choice.

Suppressing the j subscript for simplicity, the smoothing parameter rate $m \asymp n^{2/3}$ gives the most accurate CIs (up to $\log(n)$ terms) because the orders of E_h and E_l are $O(m^{-1}\log(n) + m^2/n^2)$, where m^{-1} is the order of the variance of $\widehat{Q'}(\tau_j)$ and m^2/n^2 is the order of its bias. Ideally, we could derive more precise expressions of E_h and E_l that could then be minimised over m; for now, we just consider rates.

From (2.5) and (2.6) in Bloch and Gastwirth (1968), up to smaller-order terms,

$$\text{Var}(\widehat{Q'}(\tau)) \doteq m^{-1}\frac{(Q'(\tau))^2}{2}, \quad \text{Bias}(\widehat{Q'}(\tau)) \doteq (m/n)^2\frac{Q'''(\tau)}{6}.$$

One way to select m with the CPE-optimal rate is to minimise the sum of the bias and variance of $\widehat{Q'}(\tau_j)/Q'(\tau_j)$. The variance, bias, and corresponding FOC are

$$\begin{aligned}
\text{Var}(\widehat{Q'}(\tau)/Q'(\tau)) &= \frac{\text{Var}(\widehat{Q'}(\tau))}{(Q'(\tau))^2} \doteq m^{-1}\frac{(Q'(\tau))^2}{2(Q'(\tau))^2} = 1/(2m), \\
\text{Bias}(\widehat{Q'}(\tau)/Q'(\tau)) &= \frac{\text{Bias}(\widehat{Q'}(\tau))}{Q'(\tau)} \doteq (m/n)^2\frac{Q'''(\tau)}{6Q'(\tau)}, \\
0 &= \frac{\partial}{\partial m}((m/n)^2\frac{Q'''(\tau)}{6Q'(\tau)} + (1/2)m^{-1}) = 2m/n^2\frac{Q'''(\tau)}{6Q'(\tau)} - (1/2)m^{-2}, \\
&\implies m^3 = n^2\frac{3Q'(\tau)}{2Q'''(\tau)}.
\end{aligned}$$

From here, we use a "Gaussian plug-in" approach like in the rule-of-thumb bandwidth

of Silverman (1986), i.e., we compute and plug in the quantile derivatives for a $\mathrm{N}(\mu, \sigma^2)$ distribution (shown explicitly in the supplemental appendix). Thanks to having used $\widehat{Q'}(\tau_j)/Q'(\tau_j)$ instead of $\widehat{Q'}(\tau_j)$, the result is invariant to σ (as well as μ):

$$m = n^{2/3}\left(1.5\frac{(\phi(\Phi^{-1}(\tau)))^2}{1+2(\Phi^{-1}(\tau))^2}\right)^{1/3}. \tag{B.1}$$

This is the bandwidth we use in our code, after replacing m with m_j and τ with τ_j.

B.2. Plug-in bandwidth for conditional inference

The following suggestions are all implemented in the code available on the journal's (or latter author's) website.

Since analytic expressions for unconditional CPE do not exist for the methods considered here, we recommend multiplying the single quantile plug-in bandwidth in GK Section 4.3 by the appropriate power of n to achieve the optimal rate from Theorem 4.1. Note that the bandwidth values in GK are only for $d = 1$; only rates are given for $d > 1$.

The one-sided rate for joint inference over multiple quantiles is the same as for a single quantile. For simplicity, we suggest a common bandwidth for all quantiles, using $\tau = \arg\min_{\tau_j} \tau_j(1-\tau_j)$ in the single quantile plug-in bandwidth, and we suggest using α instead of $\tilde{\alpha}$ in the plug-in bandwidth formula; neither choice affects the asymptotic bandwidth rate. For two-sided joint inference, we recommend further multiplying the bandwidth by $n^{-2/((2+d)(4+3d))}$ to get the optimal rate.

For one-sided inference on linear combinations of quantiles, we recommend multiplying the single quantile plug-in bandwidth (in GK Section 4.3) by n to the power of $8/((12+7d)(4+3d))$ to get a bandwidth with the optimal rate. This time $\tilde{\alpha} \geq \alpha$, so we suggest plugging in the calibrated $\tilde{\alpha}$ that would be used if the sample size were n rather than N_n, and again whichever τ_j minimises $\tau_j(1-\tau_j)$. For two-sided inference on linear combinations, we similarly recommend multiplying the single quantile plug-in bandwidth by $n^{-2/((12+7d)(2+d))}$ to get a bandwidth with the optimal rate.

For quantile differences, the adjustment is the same as for linear combinations, but with separate bandwidths for the $T = 0$ and $T = 1$ samples. We again recommend using $\tau = \arg\min_{\tau_j} \tau_j(1-\tau_j)$, and for the one-sided case, the $\tilde{\alpha}$ that would result from sample size n.