Bootstrap Inference for Quantile Treatment Effects in Randomized Experiments with Matched Pairs*

Liang Jiang† Xiaobin Liu ‡ Peter C.B. Phillips § Yichong Zhang¶

*We thank the editor and two referees for extensive comments which have led to many improvements. We mention with thanks Kengo Kato, Yu-Chin Hsu, Shuping Shi, and Jun Yu for useful comments and suggestions. We are grateful to David McKenzie for providing the Stata code of the matching algorithm used in the empirical application, and to Esther Duflo and Cynthia Kinnan for providing the empirical data in the early versions of the paper. Jiang acknowledges support from MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.18YJC790063). Liu acknowledges financial support from the Chinese Ministry of Education Project of Humanities and Social Sciences (No. 18YJC790005) and the National Natural Science Foundation of China (No.72003171). Zhang acknowledges financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169 and a Lee Kong Chian Fellowship. Phillips acknowledges support from NSF Grant No. SES 18-50860, a Kelly Fellowship at the University of Auckland, and a Lee Kong Chian Fellowship. Any and all errors are our own.

†Fanhai International School of Finance, Fudan University. E-mail address: jiangliang@fudan.edu.cn.

‡The corresponding author. School of Economics, Academy of Financial Research, Zhejiang University. E-mail address: liuxiaobin@zju.edu.cn.

§Yale University, University of Auckland, University of Southampton, and Singapore Management University. E-mail address: peter.phillips@yale.edu

¶Singapore Management University. E-mail address: yczhang@smu.edu.sg.

Abstract

This paper examines methods of inference concerning quantile treatment effects (QTEs) in randomized experiments with matched-pairs designs (MPDs). The standard multiplier bootstrap inference fails to capture the negative dependence of observations within each pair, and thus, is conservative. The analytical inference involves estimating multiple functional quantities that requires several tuning parameters. In this paper, we propose two bootstrap methods that can consistently approximate the limit distribution of the original QTE estimator and lessen the burden of tuning parameter choice. In particular, the inverse propensity score weighted multiplier bootstrap can be implemented without knowledge of pair identities.

Keywords: Bootstrap inference, matched pairs, quantile treatment effect, randomized control trials

JEL codes: C14, C21

1 Introduction

Matched-pairs designs (MPDs) have recently seen widespread and increasing use in various randomized experiments conducted by economists. By MPD we mean a randomization scheme that first pairs units based on the closeness of their baseline covariates and then randomly assigns one unit in the pair to be treated. In development economics, researchers routinely pair villages, neighborhoods, microenterprises, or townships in their experiments (Banerjee, Duflo, Glennerster, and Kinnan, 2015; Crepon, Devoto, Duflo, and Pariente, 2015; Glewwe, Park, and Zhao, 2016; Groh and McKenzie, 2016). In labor economics, especially in the field of education, researchers pair schools or students to evaluate the effects of various education interventions (Angrist and Lavy, 2009; Beuermann, Cristia, Cueto, Malamud, and Cruzaguayo, 2015; Fryer, 2017; Fryer, Devi, and Holden, 2017; Bold, Kimenyi, Mwabu, Nganga, and Sandefur, 2018; Fryer, 2018). Bruhn and McKenzie (2009) surveyed leading experts in development field experiments and reported that 56% of them explicitly match pairs of observations on baseline characteristics.

Researchers often use randomized experiments to estimate quantile treatment effects (QTEs) as well as average treatment effects (ATEs). Quantile effects can capture heterogeneity in both the sign and magnitude of treatment effects, which may vary according to position within the distribution of outcomes. A common practice in conducting inference on QTEs is to use bootstrap rather than analytical methods because the latter usually require tuning parameters in implementation. However, the treatment assignment in MPDs introduces negative *dependence* because exactly half of the units are treated. **Neither the standard multiplier bootstrap nor bootstrapping the pairs mimics such dependence. This difficulty raises the question of how to conduct bootstrap inference for QTEs in MPDs in a manner that mitigates these shortcomings.**

To tackle these shortcomings we propose two bootstrap inference methods: the gradient bootstrap and the inverse propensity score weighted (IPW) mul-

tiplier bootstrap. We first show that the gradient bootstrap can consistently approximate the limit distribution of the QTE estimator under MPDs uniformly over a compact set of quantile indexes. Hagemann (2017) proposed using the gradient bootstrap for the cluster-robust inference in linear quantile regression models. Like Hagemann (2017), we rely on the gradient bootstrap to avoid estimating the Hessian matrix that involves the infinite-dimensional nuisance parameters. The gradient bootstrap procedure is therefore free of tuning parameters. On the other hand and differing from Hagemann (2017), we construct a specific perturbation of the score based on pair and adjacent pairs of observations, which can capture the dependence structure in the original data.

To implement our gradient bootstrap method, researchers need to know the identities of pairs. Such information may not be available when they are using an experiment that was run by other investigators in the past and the randomization procedure may not have been fully described. For example, publicly available datasets for papers such as Panagopoulos and Green (2008) and Butler (2010) contain no information on pair identities.[1] Bruhn and McKenzie (2009) also pointed out that many papers in existing experiments do not describe the randomization procedure in detail.

To address these issues, we next propose an IPW multiplier bootstrap, which can be implemented without the knowledge of pair identities. We show that such a bootstrap can consistently approximate the limit distribution of the QTE estimator under MPDs. There is a cost to not using information about pair identities as the method requires one tuning parameter for the nonparametric estimation of the propensity score. In spite of this additional cost, this multiplier bootstrap method still has an advantage over direct analytic inference because practical implementation of the latter requires more than one tuning parameter.

The contributions in the present paper relate to other recent research. Bai, Shaikh, and

[1]Both datasets are available in the data archive of the Institute for Social and Policy Studies at Yale University (https://isps.yale.edu/research/data).

Romano (2021) first pointed out that in MPDs the two-sample t-test for the null hypothesis that the ATE equals a pre-specified value is conservative. They then proposed adjusting the standard error of the estimator and studied the validity of the permutation test. This paper complements those results by considering the QTEs and by developing new methods of bootstrap inference. Unlike the permutation test, our methods of bootstrap inference do not require studentization, which is cumbersome in the QTE context. In addition, our multiplier bootstrap method complements their results by providing a way to perform inference relating to both ATEs and QTEs when pair identities are unknown. In other work, Bai (2019) investigated the optimality of MPDs in randomized experiments. Zhang and Zheng (2020) considered bootstrap inference under covariate-adaptive randomization. A key difference in our contribution is that in MPDs the number of strata is proportional to the sample size, whereas in covariate-adaptive randomization that number is fixed. In consequence, the present work uses fundamentally different asymptotic arguments and bootstrap methods from those employed by Zhang and Zheng (2020). The present paper also fits within a growing literature that studies inference in randomized experiments (e.g., Hahn, Hirano, and Karlan (2011), Athey and Imbens (2017), Abadie, Chingos, and West (2018), Bugni, Canay, and Shaikh (2018), Tabord-Meehan (2018), and Bugni, Canay, and Shaikh (2019), among others).

The remainder of the paper is organized as follows. Section 2 describes the model setup and notation. Section 3 develops the asymptotic properties of our QTE estimator. In Section 4 we study the naive multiplier bootstrap, the naive multiplier bootstrap of the pairs, the gradient bootstrap, and the IPW multiplier bootstrap. Section 5 provides computational details and recommendations for practitioners. Section 6 reports simulation results. Section 7 provides an empirical application of our methods of bootstrap inference to the data in Groh and McKenzie (2016), examining both the ATEs and QTEs of macroinsurance on consumption and profits. Section 8 concludes. Proofs of all results and additional simulations are in the Online Supplement.

2 Setup and Notation

Denote the potential outcomes for treated and control groups as $Y(1)$ and $Y(0)$, respectively. Treatment status is written as A, where $A = 1$ is treated and $A = 0$ is untreated. The researcher only observes $\{Y_i, X_i, A_i\}_{i=1}^{2n}$ where $Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i)$, and $X_i \in \Re^{d_x}$ is a collection of baseline covariates, where d_x is the dimension of X. The parameter of interest is the τth QTE, denoted as

$$q(\tau) = q_1(\tau) - q_0(\tau),$$

where $q_1(\tau)$ and $q_0(\tau)$ are the τth quantiles of $Y(1)$ and $Y(0)$, respectively. The testing problems of interest involve single, multiple, or even a continuum of quantile indexes, as in the following null hypotheses

$$\begin{aligned}
&\mathcal{H}_0 : q(\tau) = \underline{q} \quad \text{versus} \quad q(\tau) \neq \underline{q}, \\
&\mathcal{H}_0 : q(\tau_1) - q(\tau_2) = \underline{q} \quad \text{versus} \quad q(\tau_1) - q(\tau_2) \neq \underline{q}, \text{ and} \\
&\mathcal{H}_0 : q(\tau) = \underline{q}(\tau) \ \forall \tau \in \Upsilon \quad \text{versus} \quad q(\tau) \neq \underline{q}(\tau) \text{ for some } \tau \in \Upsilon,
\end{aligned}$$

for some pre-specified value $\underline{q}$ or function $\underline{q}(\tau)$, where Υ is some compact subset of $(0, 1)$.

The units are grouped into pairs based on the closeness of their baseline covariates, which is now made clear. Pairs of units are denoted

$$(\pi(2j-1), \pi(2j)) \text{ for } j \in [n],$$

where $[n] = \{1, \cdots, n\}$ and π is a permutation of $2n$ units based on $\{X_i\}_{i=1}^{2n}$ as specified in Assumption 1(iv) below. Within a pair, one unit is randomly assigned to treatment and the other to control. Specifically, we make the following assumption on the data generating process (DGP) and the treatment assignment rule.

Assumption 1. *(i)* $\{Y_i(1), Y_i(0), X_i\}_{i=1}^{2n}$ *is i.i.d.*

(ii) $\{Y_i(1), Y_i(0)\}_{i=1}^{2n} \perp\!\!\!\perp \{A_i\}_{i=1}^{2n} | \{X_i\}_{i=1}^{2n}$.

(iii) Conditionally on $\{X_i\}_{i=1}^{2n}$, $\{A_{\pi(2j-1)}, A_{\pi(2j)}\}_{j\in[n]}$ *are i.i.d. and each uniformly distributed over the values in* $\{(1,0),(0,1)\}$.

(iv) $\frac{1}{n}\sum_{j=1}^{n} \left\| X_{\pi(2j)} - X_{\pi(2j-1)} \right\|_2^r \xrightarrow{p} 0$ *for* $r = 1, 2$.

Assumption 1 is used in Bai et al. (2021) to which we refer readers for more discussion. In Assumption 1(iv), $||\cdot||_2$ denotes Euclidean distance. However, all our results hold if $||\cdot||_2$ is replaced by any distance that is equivalent to it, such as L_∞ distance, L_1 distance, and the Mahalanobis distance when all the eigenvalues of the covariance matrix are bounded and bounded away from zero.

3 Estimation

Let $\hat{q}_1(\tau)$ and $\hat{q}_0(\tau)$ be the τth percentiles of outcomes in the treated and control groups, respectively. Then, the τth QTE estimator we consider is just

$$\hat{q}(\tau) = \hat{q}_1(\tau) - \hat{q}_0(\tau).$$

For ease of notation, dependence of $\hat{q}(\tau)$, $\hat{q}_1(\tau)$, $\hat{q}_0(\tau)$ and all the other estimators on n is suppressed throughout the rest of the paper. To facilitate further analysis and motivate our bootstrap procedure, we note that $\hat{q}(\tau)$ can be equivalently computed by direct quantile regression. Let

$$(\hat{\beta}_0(\tau), \hat{\beta}_1(\tau)) = \arg\min_b \sum_{i=1}^{2n} \rho_\tau(Y_i - \dot{A}^\top b),$$

where $\dot{A}_i = (1, A_i)^\top$ and $\rho_\tau(u) = u(\tau - 1\{u \le 0\})$. Then, $\hat{q}(\tau) = \hat{\beta}_1(\tau)$ and $\hat{q}_0(\tau) = \hat{\beta}_0(\tau)$.

Assumption 2. *For $a = 0, 1$, define $F_a(\cdot)$, $F_a(\cdot|x)$, $f_a(\cdot)$, and $f_a(\cdot|x)$ as the CDF of $Y_i(a)$, the conditional CDF of $Y_i(a)$ given $X_i = x$, the PDF of $Y_i(a)$, and the conditional PDF of $Y_i(a)$ given $X_i = x$, respectively.*

(i) $f_a(q_a(\tau))$ is bounded and bounded away from zero uniformly over $\tau \in \Upsilon$, and $f_a(q_a(\tau)|x)$ is uniformly bounded for $(x, \tau) \in Supp(X) \times \Upsilon$.

(ii) There exists a function $C(x)$ such that

$$\sup_{\tau \in \Upsilon} |f_a(q_a(\tau) + v|x) - f_a(q_a(\tau)|x)| \leq C(x)|v| \quad \text{and} \quad \mathbb{E}C(X_i) < \infty.$$

(iii) Let $\mathcal{N}_0$ be a neighborhood of 0. Then, there exists a constant C such that for any $x, x' \in Supp(X)$

$$\sup_{\tau \in \Upsilon, v \in \mathcal{N}_0} |f_a(q_a(\tau) + v|x') - f_a(q_a(\tau) + v|x)| \leq C||x' - x||_2$$

and

$$\sup_{\tau \in \Upsilon, v \in \mathcal{N}_0} |F_a(q_a(\tau) + v|x') - F_a(q_a(\tau) + v|x)| \leq C||x' - x||_2.$$

Assumption 2(i) is a standard regularity condition widely assumed in quantile estimation. The Lipschitz conditions in Assumptions 2(ii) and 2(iii) are similar in spirit to those assumed in Bai et al. (2021, Assumption 2.1) and ensure that units that are "close" in terms of their baseline covariates are suitably comparable. For $a = 0, 1$, let $m_{a,\tau}(x, q) = \mathbb{E}(\tau - 1\{Y(a) \leq q\}|X = x)$ and $m_{a,\tau}(x) = m_{a,\tau}(x, q_a(\tau))$.

Theorem 3.1. *Suppose Assumptions 1 and 2 hold. Then, uniformly over $\tau \in \Upsilon$,*

$$\sqrt{n}(\hat{q}(\tau) - q(\tau)) \rightsquigarrow \mathcal{B}(\tau),$$

where $\mathcal{B}(\tau)$ is a tight Gaussian process with covariance kernel $\Sigma(\cdot,\cdot)$ such that

$$\Sigma(\tau,\tau') = \frac{\min(\tau,\tau') - \tau\tau' - \mathbb{E}m_{1,\tau}(X)m_{1,\tau'}(X)}{f_1(q_1(\tau))f_1(q_1(\tau'))} + \frac{\min(\tau,\tau') - \tau\tau' - \mathbb{E}m_{0,\tau}(X)m_{0,\tau'}(X)}{f_0(q_0(\tau))f_0(q_0(\tau'))} + \frac{1}{2}\mathbb{E}\left(\frac{m_{1,\tau}(X)}{f_1(q_1(\tau))} - \frac{m_{0,\tau}(X)}{f_0(q_0(\tau))}\right)\left(\frac{m_{1,\tau'}(X)}{f_1(q_1(\tau'))} - \frac{m_{0,\tau'}(X)}{f_0(q_0(\tau'))}\right).$$

Several remarks are in order. First, the asymptotic variance of $\hat{q}(\tau)$ under MPDs is

$$\Sigma(\tau,\tau) = \frac{\tau - \tau^2 - \mathbb{E}m_{1,\tau}^2(X)}{f_1^2(q_1(\tau))} + \frac{\tau - \tau^2 - \mathbb{E}m_{0,\tau}^2(X)}{f_0^2(q_0(\tau))} + \frac{1}{2}\mathbb{E}\left(\frac{m_{1,\tau}(X)}{f_1(q_1(\tau))} - \frac{m_{0,\tau}(X)}{f_0(q_0(\tau))}\right)^2. \quad (3.1)$$

Note further that the asymptotic variance of $\hat{q}(\tau)$ under simple random sampling (SRS)[2] is

$$\Sigma^{\dagger}(\tau,\tau) = \frac{\tau - \tau^2}{f_1^2(q_1(\tau))} + \frac{\tau - \tau^2}{f_0^2(q_0(\tau))}. \quad (3.2)$$

It is clear that

$$\Sigma^{\dagger}(\tau,\tau) - \Sigma(\tau,\tau) = \frac{1}{2}\mathbb{E}\left(\frac{m_{1,\tau}(X)}{f_1(q_1(\tau))} + \frac{m_{0,\tau}(X)}{f_0(q_0(\tau))}\right)^2 \geq 0. \quad (3.3)$$

Equality in the last expression holds when both $m_{1,\tau}(X)$ and $m_{0,\tau}(X)$ are zero, which implies that X is irrelevant to the τth quantiles of $Y(0)$ and $Y(1)$.

Second, note that $\hat{q}(\tau)$ has the same asymptotic variance as that for the QTE estimators studied by Firpo (2007) and Donald and Hsu (2014) under SRS.

Third, to provide an analytic estimate of the asymptotic variance $\Sigma(\tau,\tau)$ it is necessary at least to estimate the infinite dimensional nuisance parameters $f_1(q_1(\tau))$ and $f_0(q_0(\tau))$, which requires two tuning parameters. Hence, if a researcher is interested in testing a null hypothesis that involves G quantile indexes, $2G$ tuning parameters are needed to estimate $2G$ densities, a cumbersome task in practical work; and to construct a uniform confidence band for the QTE analytically, two tuning parameters are needed at each grid point of the

[2]By simple random sample, we mean the treatment status is assigned independently with probability 1/2.

quantile indexes. Moreover, if pair identities are unknown, analytic methods of inference potentially require nonparametric estimation of the quantities $m_{a,\tau}(\cdot)$ for $a = 0, 1$ as well. There are other practical difficulties. Nonparametric estimation is sometimes sensitive to the choice of tuning parameters and rule-of-thumb tuning parameter selection may not be appropriate for every data generating process (DGP) or every quantile. Use of cross-validation in selecting the tuning parameters is possible in principle but in practice time-consuming. These practical difficulties of analytic methods of inference provide a strong motivation to investigate bootstrap inference procedures that are much less reliant on tuning parameters.

4 Bootstrap Inference

This section examines four bootstrap inference procedures for the QTEs in MPDs. We first show that both the naive multiplier bootstrap and the naive multiplier bootstrap of the pairs fail to approximate the limit distribution of the QTE estimator derived in Section 3. We then propose two bootstrap methods that can consistently approximate the limit distribution of the QTE estimator.

4.1 Naive Multiplier Bootstrap

Consider first the naive multiplier bootstrap estimators of $\hat{\beta}_0(\tau)$ and $\hat{\beta}_1(\tau)$, defining

$$(\hat{\beta}_0^m(\tau), \hat{\beta}_1^m(\tau)) = \arg\min_b \sum_{i=1}^{2n} \xi_i \rho_\tau(Y_i - \dot{A}_i^\top b),$$

where ξ_i is the bootstrap weight defined in the next assumption.

Assumption 3. *Suppose $\{\xi_i\}_{i=1}^{2n}$ is a sequence of nonnegative i.i.d. random variables with unit expectation and variance and a sub-exponential upper tail.*

In practice, we generate ξ_i independently from the standard exponential distribution. Denote $\hat{q}^m(\tau) = \hat{\beta}_1^m(\tau)$ and recall that $\hat{q}(\tau) = \hat{\beta}_1(\tau)$.

Theorem 4.1. *If Assumptions 1–3 hold, then conditionally on the data and uniformly over* $\tau \in \Upsilon$,

$$\sqrt{n}(\hat{q}^m(\tau) - \hat{q}(\tau)) \rightsquigarrow \mathcal{B}^m(\tau),$$

where $\mathcal{B}^m(\tau)$ *is a tight Gaussian process with covariance kernel* $\Sigma^\dagger(\cdot,\cdot)$ *such that*

$$\Sigma^\dagger(\tau,\tau') = \frac{\min(\tau,\tau') - \tau\tau'}{f_1(q_1(\tau))f_1(q_1(\tau'))} + \frac{\min(\tau,\tau') - \tau\tau'}{f_0(q_0(\tau))f_0(q_0(\tau'))}.$$

Three remarks are in order. First, $\sqrt{n}(\hat{q}^m(\tau) - \hat{q}(\tau))$ and $\mathcal{B}^m(\tau)$ are viewed as processes indexed by $\tau \in \Upsilon$ and denoted by G_n and G, respectively. Then, following van der Vaart and Wellner (1996, Chapter 2.9), we say G_n weakly converges to G conditionally on data and uniformly over $\tau \in \Upsilon$ if

$$\sup_{h \in \mathrm{BL}_1} |\mathbb{E}_\xi h(G_n) - \mathbb{E}h(G)| \xrightarrow{p} 0,$$

where BL_1 is the set of all functions $h : \ell^\infty(\Upsilon) \mapsto [0,1]$ such that $|h(z_1) - h(z_2)| \leq ||z_1 - z_2||_\infty$ for every $z_1, z_2 \in \ell^\infty(\Upsilon)$, and $\mathbb{E}_\xi$ denotes expectation with respect to the bootstrap weights $\{\xi\}_{i=1}^n$.[3] The same remark applies to Theorems 4.2, 4.3, and 4.4 below.

Second, $\Sigma^\dagger(\tau,\tau')$ is just the covariance kernel of the QTE estimator when simple random sampling (instead of the MPD) is used as the treatment assignment rule. It follows that the naive multiplier bootstrap fails to approximate the limit distribution of $\hat{q}(\tau)$ ($\hat{\beta}_1(\tau)$). The intuition is straightforward. Given the data, the bootstrap weights are i.i.d. and thus unable to mimic the cross-sectional dependence in the original sample.

Third, it is possible to consider the conventional nonparametric bootstrap in which the

[3] **The asymptotic measurability holds in our setting due to van der Vaart and Wellner (1996, Lemma 1.5.2), which requires the asymptotic tightness of the bootstrap process. The latter has been established in the proof of Theorem 4.1. For notational simplicity, we ignore the issue of asymptotic measurability.**

bootstrap sample is generated from the empirical distribution of the data. If the observations are i.i.d., van der Vaart and Wellner (1996, Section 3.6) showed that the conventional bootstrap is first-order equivalent to a multiplier bootstrap with Poisson(1) weights. However, in the current setting, $\{A_i\}_{i\in[2n]}$ are dependent. It is technically challenging to show rigorously that the above equivalence still holds and this challenge is left for future research.

4.2 Naive Multiplier Bootstrap of the Pairs

Next consider the naive multiplier bootstrap of the pairs which uses the same bootstrap multiplier for the observations within the pair. Let

$$(\hat{\beta}_0^p(\tau), \hat{\beta}_1^p(\tau)) = \arg\min_b \sum_{i=1}^{2n} \xi_i^p \rho_\tau(Y_i - \dot{A}_i^\top b),$$

where ξ_i^p is the bootstrap weight defined in the next assumption.

Assumption 4. *Suppose $\{\xi^p_{\pi(2j-1)}\}_{j=1}^n$ is a sequence of nonnegative i.i.d. random variables with unit expectation and variance and a sub-exponential upper tail and $\xi^p_{\pi(2j-1)} = \xi^p_{\pi(2j)}$ for $j = 1, \cdots, n$.*

Because the units in the same pair share the same multiplier, we call this the naive multiplier bootstrap of the pairs. Denote $\hat{q}^p(\tau) = \hat{\beta}_1^p(\tau)$.

Theorem 4.2. *If Assumptions 1, 2, and 4 hold, then conditionally on the data and uniformly over $\tau \in \Upsilon$,*

$$\sqrt{n}(\hat{q}^p(\tau) - \hat{q}(\tau)) \rightsquigarrow \mathcal{B}^p(\tau),$$

where $\mathcal{B}^p(\tau)$ is a tight Gaussian process with covariance kernel $\Sigma^p(\cdot,\cdot)$ such that

$$\Sigma^p(\tau,\tau') = \frac{\min(\tau,\tau') - \tau\tau'}{f_1(q_1(\tau))f_1(q_1(\tau'))} + \frac{\min(\tau,\tau') - \tau\tau'}{f_0(q_0(\tau))f_0(q_0(\tau'))} - \frac{\mathbb{E}m_{1,\tau}(X_i)m_{0,\tau'}(X_i)}{f_1(q_1(\tau))f_0(q_0(\tau'))} - \frac{\mathbb{E}m_{1,\tau'}(X_i)m_{0,\tau}(X_i)}{f_0(q_0(\tau))f_1(q_1(\tau'))}.$$

Three remarks are in order. First, Theorem 4.2 implies that bootstrapping pairs of observations alone is unable to mimic the dependence structure in the original sample. The potential outcomes $Y_i(1)$ and $Y_i(0)$ can be written as functions of $(U_i(1), X_i)$ and $(U_i(0), X_i)$, respectively, where $(U_i(1), U_i(0))$ are the unobserved heterogeneities. In MPDs, the units in the same pair share similar covariate X, but their unobserved heterogeneities are independent conditional on X. However, bootstrapping pairs artificially introduces dependence between the unobserved heterogeneities of two observations within the same pair by forcing their bootstrap multipliers to be the same.

Second, the variances of the original estimator $\hat{q}(\tau)$ under MPD and $\hat{q}^p(\tau)$ conditional on data have the following relationship:

$$\Sigma^p(\tau,\tau) - \Sigma(\tau,\tau) = \frac{1}{2}\mathbb{E}\left(\frac{m_{1,\tau}(X_i)}{f_1(q_1(\tau))} - \frac{m_{0,\tau}(X_i)}{f_0(q_0(\tau))}\right)^2 \geq 0. \tag{4.1}$$

Third, the gradient bootstrap procedure proposed below is based on a similar idea and uses the same weight for the observations within the pair to construct the score $S^*_{n,1}$ defined in (4.5). But in order to construct a final score that exactly mimics the dependence in the data, an extra score component, $S^*_{n,2}$ defined in (4.6) below, is needed. This component is constructed based on adjacent pairs of observations.

4.3 Gradient Bootstrap Inference

We develop an approximation for the asymptotic distribution of the QTE estimator via the gradient bootstrap. Let $u = \sqrt{n}(b - \beta(\tau))$ be a localizing estimation error parameter. From

the derivations in Theorem 3.1, we see that

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) = Q^{-1}(\tau) \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} S_n(\tau) + o_p(1), \tag{4.2}$$

where

$$S_n(\tau) = \begin{pmatrix} \sum_{i=1}^{2n} \frac{A_i}{\sqrt{n}} (\tau - 1\{Y_i(1) \leq q_1(\tau)\}) \\ \sum_{i=1}^{2n} \frac{(1-A_i)}{\sqrt{n}} (\tau - 1\{Y_i(0) \leq q_0(\tau)\}) \end{pmatrix},$$

and

$$Q(\tau) = \begin{pmatrix} f_1(q_1(\tau)) + f_0(q_0(\tau)) & f_1(q_1(\tau)) \\ f_1(q_1(\tau)) & f_1(q_1(\tau)) \end{pmatrix}.$$

The gradient bootstrap proposes to perturb the objective function by some random error $S_n^*(\tau)$, which will be specified later. This error in turn perturbs the score function $S_n(\tau)$. The corresponding bootstrap estimator $\hat{\beta}^*(\tau)$ solves the following optimization problem

$$\hat{\beta}^*(\tau) = \arg\min_b \sum_{i=1}^{2n} \rho_\tau(Y_i - \dot{A}_i^\top b) - \sqrt{n} b^\top \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} S_n^*(\tau). \tag{4.3}$$

We can then show that

$$\sqrt{n}(\hat{\beta}^*(\tau) - \beta(\tau)) = Q^{-1}(\tau) \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} [S_n(\tau) + S_n^*(\tau)] + o_p(1). \tag{4.4}$$

Taking the difference between (4.2) and (4.4) gives

$$\sqrt{n}(\hat{\beta}^*(\tau) - \hat{\beta}(\tau)) = Q^{-1}(\tau) \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} S_n^*(\tau) + o_p(1).$$

The second element of $\hat{\beta}^*(\tau)$ in (4.3) is the bootstrap version of the QTE estimator, which is denoted $\hat{q}^*(\tau)$. By solving (4.3) we avoid estimating the Hessian $Q(\tau)$, which involves infinite-dimensional nuisance parameters. Then, for the gradient bootstrap to consistently approximate the limit distribution of the original estimator $\hat{\beta}(\tau)$, we need only construct $S_n^*(\tau)$ in such a way that its weak limit given the data coincides with that of the original score $S_n(\tau)$.

Accordingly, we now show how to specify $S_n^*(\tau)$. Let $\{\eta_j\}_{j=1}^n$ and $\{\hat{\eta}_k\}_{k=1}^{\lfloor n/2 \rfloor}$ be two mutually independent i.i.d. sequences of standard normal random variables. Use the indexes $(j,1),(j,0)$ to denote the indexes in $(\pi(2j-1),\pi(2j))$ with $A=1$ and $A=0$, respectively. For example, if $A_{\pi(2j)}=1$ and $A_{\pi(2j-1)}=0$, then $(j,1)=\pi(2j)$ and $(j,0)=\pi(2j-1)$. Similarly, use indexes $(k,1),\cdots,(k,4)$ to denote the first index in $(\pi(4k-3),\cdots,\pi(4k))$ with $A=1$, the first index with $A=0$, the second index with $A=1$, and the second index with $A=0$, respectively. Now let

$$S_n^*(\tau)=\frac{S_{n,1}^*(\tau)+S_{n,2}^*(\tau)}{\sqrt{2}},$$

where

$$S_{n,1}^*(\tau)=\frac{1}{\sqrt{n}}\begin{pmatrix}\sum_{j=1}^n \eta_j\left(\tau-1\{Y_{(j,1)}\le \hat{q}_1(\tau)\}\right)\\ \sum_{j=1}^n \eta_j\left(\tau-1\{Y_{(j,0)}\le \hat{q}_0(\tau)\}\right)\end{pmatrix}, \tag{4.5}$$

and

$$S_{n,2}^*(\tau)=\frac{1}{\sqrt{n}}\begin{pmatrix}\sum_{k=1}^{\lfloor n/2 \rfloor} \hat{\eta}_k\left[\left(\tau-1\{Y_{(k,1)}\le \hat{q}_1(\tau)\}\right)-\left(\tau-1\{Y_{(k,3)}\le \hat{q}_1(\tau)\}\right)\right]\\ \sum_{k=1}^{\lfloor n/2 \rfloor} \hat{\eta}_k\left[\left(\tau-1\{Y_{(k,2)}\le \hat{q}_0(\tau)\}\right)-\left(\tau-1\{Y_{(k,4)}\le \hat{q}_0(\tau)\}\right)\right]\end{pmatrix}. \tag{4.6}$$

The perturbation S_n^* consists of two parts: $S_{n,1}^*$ and $S_{n,2}^*$. The first part $S_{n,1}^*$ is constructed by pairs of observations, based on the idea of bootstrapping the pairs. But bootstrapping the pairs alone cannot fully capture the dependence structure in the MPD, as shown in Section

4.2. So $S^*_{n,1}$ is adjusted by adding a term capturing the remaining dependence. This second term, $S^*_{n,2}$, is motivated by the idea of using adjacent pairs to adjust the standard error of the ATE estimator under MPDs in Bai et al. (2021).

In Section 5 we show how to compute the bootstrap estimator $\hat{\beta}^*(\tau)$ directly from the sub-gradient condition of (4.3). This method avoids the optimization inherent in (4.3) and computation is fast. The following assumption imposes the condition that baseline covariates in adjacent pairs are also 'close'.

Assumption 5. *Suppose that* $\frac{1}{n}\sum_{k=1}^{\lfloor n/2 \rfloor} \left\| X_{(k,l)} - X_{(k,l')} \right\|_2^r \xrightarrow{p} 0$ *for* $r = 1, 2$ *and* $l, l' \in [4]$.

Assumption 5 and Assumption 1(iv) are jointly equivalent to Bai et al. (2021, Assumption 2.4). We refer readers to Bai et al. (2021) for further discussion of this assumption. In particular, Bai et al. (2021, Theorems 4.1 and 4.2) show that it is possible to implement the matching algorithm to re-order pairs so that both Assumption 5 and Assumption 1(iv) hold automatically. We provide more detail in Section 5.1.

Define $\hat{q}^*(\tau) = \hat{\beta}^*_1(\tau)$ and recall that $\hat{q}(\tau) = \hat{\beta}_1(\tau)$. We have the following result.

Theorem 4.3. *Suppose Assumptions 1, 2, and 5 hold. Then, conditionally on the data and uniformly over* $\tau \in \Upsilon$, $\sqrt{n}(\hat{q}^*(\tau) - \hat{q}(\tau)) \rightsquigarrow \mathcal{B}(\tau)$, *where* $\mathcal{B}(\tau)$ *is the same Gaussian process defined in Theorem 3.1.*

Two remarks on Theorem 4.3 are in order. First, the bootstrap estimator $\hat{q}^*(\tau)$ has the following objectives: (i) to avoid estimating densities; and (ii) to mimic the distribution of the original estimator $\hat{\beta}(\tau)$ under MPDs. Objective (i) relates to the Hessian (Q) and (ii) to the score (S_n) of the quantile regression. The gradient bootstrap provides a flexible approach to achieve both goals.

Second, to implement the gradient bootstrap, researchers need to know identities of pairs. This information may not be available when the experiment was run by others and the randomization procedure was not fully detailed. In such cases, we propose IPW multiplier bootstrap inference for the QTE, whose validity is established in the next section.

4.4 IPW Multiplier Bootstrap Inference

In empirical research, researchers may not know the identities of pairs when they are using an experiment that was run by other investigators in the past and the randomization procedure may not have been fully described. For example, publicly available datasets for papers such as Panagopoulos and Green (2008) and Butler (2010) contain no information on pair identities. Bruhn and McKenzie (2009) also pointed out that many papers in existing experiments do not describe the randomization procedure in detail. In this section we establish the validity of IPW multiplier bootstrap inference for the QTE, which can be implemented without the knowledge of pair identities.

We use the sieve method to nonparametrically estimate the propensity score. Let $b(X)$ be the K-dimensional sieve basis on X and $\hat{A}_i$ be the estimated propensity score for the ith individual. Then,

$$\hat{A}_i = b^\top(X_i)\hat{\theta}, \tag{4.7}$$

where $\hat{\theta} = \arg\min_\theta \sum_{i=1}^{2n} \xi_i (A_i - b^\top(X_i)\theta)^2$ and ξ_i is the bootstrap weight defined in Assumption 3.

Because the true propensity score is $1/2$, by setting the first component of $b(X)$ to unity, we have $1/2 = b^\top(X)\theta_0$ where $\theta_0 = (0.5, 0, \cdots, 0)^\top$. The linear probability model for the propensity score is correctly specified. It is possible to use sieve logistic regression to compute the propensity score, as done by Hirano, Imbens, and Ridder (2003), Firpo (2007), and Donald and Hsu (2014). The main benefit of using logistic regression is to guarantee that the estimated propensity score lies between zero and one. However, in MPDs, the estimated propensity score is always very close to 0.5. Therefore, for simplicity, we use a linear sieve regression here.

The IPW multiplier bootstrap estimator can be computed as

$$\hat{q}^w_{ipw}(\tau) = \hat{q}^w_{ipw,1}(\tau) - \hat{q}^w_{ipw,0}(\tau),$$

where

$$\hat{q}^w_{ipw,1}(\tau) = \arg\min_q \sum_{i=1}^{2n} \frac{\xi_i A_i}{\hat{A}_i} \rho_\tau(Y_i - q) \quad \text{and} \quad \hat{q}^w_{ipw,0}(\tau) = \arg\min_q \sum_{i=1}^{2n} \frac{\xi_i (1-A_i)}{1-\hat{A}_i} \rho_\tau(Y_i - q). \tag{4.8}$$

Assumption 6. *(i) The support of X is compact. The first component of $b(X)$ is 1.*

(ii) $\max_{k\in[K]} \mathbb{E} b_k^2(X_i) \leq \overline{C} < \infty$ *for some constant* $\overline{C} > 0$*, where* $b_k(X_i)$ *is the* k*th coordinate of* $b(X_i)$*.* $\sup_{x \in Supp(X)} ||b(x)||_2 = \zeta(K)$*.*

(iii) $K^2 \zeta(K)^2 \log(n) = o(n)$*.*

(iv) With probability approaching one, there exist constants $\underline{C}$ *and* $\overline{C}$ *such that*

$$0 < \underline{C} \leq \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{2n} \xi_i b(X_i) b^\top(X_i)\right) \leq \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{2n} \xi_i b(X_i) b^\top(X_i)\right) \leq \overline{C} < \infty,$$

where $\lambda_{\min}(\mathcal{M})$ *and* $\lambda_{\max}(\mathcal{M})$ *denote the minimum and maximum eigenvalues of matrix* $\mathcal{M}$*.*

(v) There exist $\gamma_1(\tau) \in \Re^K$ *and* $\gamma_0(\tau) \in \Re^K$ *such that*

$$B_{a,\tau}(x) = m_{a,\tau}(x) - b^\top(x)\gamma_a(\tau), \quad a = 0, 1,$$

and $\sup_{a=0,1,\tau\in\Upsilon, x\in Supp(X)} |B_{a,\tau}(x)| = o(1/\sqrt{n})$*.*

Two remarks are in order. First, requiring X to have a compact support is common in nonparametric sieve estimation. Second, the quantity $\zeta(K)$ depends on the choice of basis

functions. For example, $\zeta(K) = O(K^{1/2})$ for splines and $\zeta(K) = O(K)$ for power series.[4] Taking splines as an example, Assumption 6(iii) requires $K = o(n^{1/3})$. Assumption 6(iv) is standard because $K \ll n$. Assumption 6(v) requires that the approximation error of $m_{a,\tau}(x)$ via a linear sieve function is sufficiently small. For instance, suppose $m_{a,\tau}(x)$ is s-times continuously differentiable in x with all derivatives uniformly bounded by some constant $\overline{C}$, then $\sup_{a=0,1,\tau\in\Upsilon,x\in\text{Supp}(X)} |B_{a,\tau}(x)| = O(K^{-s/d_x})$. Assumptions 6(iii) and 6(v) imply that $K = n^h$ for some $h \in (d_x/(2s), 1/3)$, which implicitly requires $s > 3d_x/2$. The choice of K reflects the usual bias-variance trade-off and is the only tuning parameter that researchers need to specify when implementing this bootstrap method.

Theorem 4.4. *Suppose Assumptions 1–3 and 6 hold, then conditionally on the data and uniformly over* $\gamma \in \Upsilon$, $\sqrt{n}(\hat{q}^w_{ipw}(\tau) - \hat{q}(\tau)) \rightsquigarrow \mathcal{B}(\tau)$, *where* $\mathcal{B}(\tau)$ *is the same Gaussian process as defined in Theorem 3.1.*

To understand the need to nonparametrically estimate the propensity score in the bootstrap sample, note that there are two stages in statistical inference in randomized experiments: the design stage and the analysis stage. In the design stage, researchers can use either simple random sampling (SRS) or matched-pairs design (MPD).[5] In the analysis stage, researchers can choose either the true (1/2 in MPDs) or the nonparametrically estimated propensity score to construct the estimator.[6] The asymptotic variances of the QTE estimator with true and estimated propensity scores under SRS are $\Sigma^\dagger(\tau,\tau)$ defined in (3.2) and $\Sigma(\tau,\tau)$ defined in (3.1), respectively, which are derived by Hahn (1998) and Firpo (2007); the asymptotic variance of the QTE estimator with the true propensity

[4]See Chen (2007) for a full discussion of the sieve method.

[5]Simple random sampling means that treatment status is assigned independently with probability 1/2. Note that SRS and MPD share the same true propensity score 1/2. But MPD achieves the strong balance that exactly half of the units are treated whereas SRS does not.

[6]**Specifically, we have** $\hat{q}_{ipw,1}(\tau) = \arg\min_q \sum_{i\in[2n]} \frac{A_i}{\hat{\pi}(X_i)}\rho_\tau(Y_i - q)$ **and** $\hat{q}_{ipw,0}(\tau) = \arg\min_q \sum_{i\in[2n]} \frac{1-A_i}{1-\hat{\pi}(X_i)}\rho_\tau(Y_i - q)$**, where** $\hat{\pi}(X_i)$ **is an estimator of the propensity score. When we use the true score, i.e.,** $\hat{\pi}(X_i) = 1/2$**, we have** $\hat{q}_{ipw,a} = \hat{q}_a$ **as defined in the paper for** $a = 0, 1$**. However, we can also let** $\hat{\pi}(X_i)$ **be the nonparametrically estimator of the propensity score.**

score under MPD is $\Sigma(\tau,\tau)$, which is shown in Theorem 3.1. These results are summarized in the following table.

Table 1: Asymptotic Variances

	True Score	Nonparametric Score
SRS	$\Sigma^{\dagger}(\tau,\tau)$	$\Sigma(\tau,\tau)$
MPD	$\Sigma(\tau,\tau)$	unknown

Note that the asymptotic variances for the QTE estimator under MPD with the true score and that under SRS with the nonparametrically estimated score are the same. If we conduct multiplier bootstrap inference, conditionally on data, the bootstrap sample of observations is independent. Therefore, in order for the multiplier bootstrap estimator to mimic the asymptotic behavior of the original estimator under MPD with the true score, we need to nonparametrically estimate the propensity score in the bootstrap sample.

The benefit of the IPW multiplier bootstrap is that it does not require knowledge of the pair identities. The cost is that we have to nonparametrically estimate the propensity score, which requires one tuning parameter and is subject to the usual curse of dimensionality. Nonetheless, we still prefer this bootstrap method of inference to the analytic approach. Analytic estimation of the standard error of the QTE estimator without the knowledge of pair identities requires nonparametric estimation of $\{m_{a,\tau}(X), f_a(q_a(\tau))\}_{a=0,1}$, which involves four tuning parameters. The number of tuning parameters further increases with the number of quantile indexes involved in the null hypothesis. To construct uniform confidence bands for QTE over τ, we require $4G$ tuning parameters for grid size G. By contrast, implementation of the IPW multiplier bootstrap requires estimation of the propensity score only once, and thus, the use of a single tuning parameter.

Inference concerning the *ATE* in MPDs can also be accomplished via a similar IPW multiplier bootstrap procedure. We can show that such a bootstrap can consistently approximate the asymptotic distribution of the ATE estimator under MPDs. This result complements

that established by Bai et al. (2021) because it provides a way to make inferences about the ATE in MPDs when information on pair identities is unavailable. That pair identity information is required by Bai et al. (2021) in computing standard errors for their adjusted t-test.

5 Computation and Guidance for Practitioners

5.1 Computation of the Gradient Bootstrap

In practice, the order of pairs in the dataset is usually arbitrary and does not satisfy Assumption 5. To apply the gradient bootstrap, researchers first need to re-order the pairs. For the jth pair with units indexed by $(j,1)$ and $(j,0)$ in the treatment and control groups, let $\overline{X}_j = \frac{1}{2}\{X_{(j,1)} + X_{(j,0)}\}$. Then, let $\overline{\pi}$ be any permutation of n elements that minimizes

$$\frac{1}{n}\sum_{j=1}^{n} ||\overline{X}_{\overline{\pi}(j)} - \overline{X}_{\overline{\pi}(j-1)}||_2.$$

The pairs are re-ordered by indexes $\overline{\pi}(1), \cdots, \overline{\pi}(n)$. With an abuse of notation, we still index the pairs after re-ordering by $1, \cdots, n$. Note that the original QTE estimator $\hat{q}(\tau) = \hat{q}_1(\tau) - \hat{q}_0(\tau)$ is invariant to the re-ordering.

For the bootstrap sample, we directly compute $\hat{\beta}^*(\tau)$ from the sub-gradient condition of (4.3). Specifically, we compute $\hat{\beta}_0^*(\tau)$ as $Y^0_{(h_0)}$ and $\hat{q}^*(\tau) \equiv \hat{\beta}_1^*(\tau)$ as $Y^1_{(h_1)} - Y^0_{(h_0)}$, where $Y^0_{(h_0)}$ and $Y^1_{(h_1)}$ are the h_0th and h_1th order statistics of outcomes in the treatment and control groups, respectively,[7] and h_0 and h_1 are two integers satisfying

$$n\tau + T^*_{n,a}(\tau) + 1 \geq h_a \geq n\tau + T^*_{n,a}(\tau),\ a = 0, 1, \tag{5.1}$$

[7]We assume $Y^a_{(1)} \leq \cdots \leq Y^a_{(n)}$ for $a = 0, 1$.

with

$$\begin{pmatrix} T^*_{n,1}(\tau) \\ T^*_{n,0}(\tau) \end{pmatrix} = \sqrt{n} S^*_n(\tau) = \frac{1}{\sqrt{2}} \Bigg[\begin{pmatrix} \sum_{j=1}^n \eta_j \left(\tau - 1\{Y_{(j,1)} \leq \hat{q}_1(\tau)\} \right) \\ \sum_{j=1}^n \eta_j \left(\tau - 1\{Y_{(j,0)} \leq \hat{q}_0(\tau)\} \right) \end{pmatrix} + \begin{pmatrix} \sum_{k=1}^{\lfloor n/2 \rfloor} \hat{\eta}_k \left[\left(\tau - 1\{Y_{(k,1)} \leq \hat{q}_1(\tau)\} \right) - \left(\tau - 1\{Y_{(k,3)} \leq \hat{q}_1(\tau)\} \right) \right] \\ \sum_{k=1}^{\lfloor n/2 \rfloor} \hat{\eta}_k \left[\left(\tau - 1\{Y_{(k,2)} \leq \hat{q}_0(\tau)\} \right) - \left(\tau - 1\{Y_{(k,4)} \leq \hat{q}_0(\tau)\} \right) \right] \end{pmatrix} \Bigg].$$

As the probability of $n\tau + T^*_{n,a}(\tau)$ being an integer is zero, h_a is uniquely defined with probability one.[8]

We summarize the steps in the bootstrap procedure as follows.

1. Re-order the pairs.

2. Compute the original estimator $\hat{q}(\tau) = \hat{q}_1(\tau) - \hat{q}_0(\tau)$.

3. Let B be the number of bootstrap replications. Let $\mathcal{G}$ be a grid of quantile indexes. For $b \in [B]$, generate $\{\eta_j\}_{j \in [n]}$ and $\{\hat{\eta}_k\}_{k \in \lfloor n/2 \rfloor}$. Compute $\hat{q}^{*b}(\tau) = Y^1_{(h_1)} - Y^0_{(h_0)}$ for $\tau \in \mathcal{G}$, where h_0 and h_1 are computed in (5.1). Obtain $\{\hat{q}^{*b}(\tau)\}_{\tau \in \mathcal{G}}$.

4. Repeat the above step for $b \in [B]$ and obtain B bootstrap estimators of the QTE, denoted as $\{\hat{q}^{*b}(\tau)\}_{b \in [B], \tau \in \mathcal{G}}$.

[8]The sub-gradient condition of (4.3) is $\hat{q}^*(\tau) = Y_{i_1} - Y_{i_0}$ such that $i_i, i_0 \in [2n]$ are two indexes, $A_{i_1} = 1$, $A_{i_0} = 0$,

$$\tau n + T^*_{n,1}(\tau) \geq \sum_{i \in [2n]} A_i 1\{Y_i < Y_{i_1}\} \geq \tau n + T^*_{n,1}(\tau) - 1, \quad \text{and}$$
$$\tau n + T^*_{n,0}(\tau) \geq \sum_{i \in [2n]} (1 - A_i) 1\{Y_i < Y_{i_0}\} \geq \tau n + T^*_{n,0}(\tau) - 1.$$

By letting $h_1 = \sum_{i \in [2n]} A_i 1\{Y_i < Y_{i_1}\} + 1$ and $h_0 = \sum_{i \in [2n]} (1 - A_i) 1\{Y_i < Y_{i_0}\} + 1$, we have $Y_{i_1} = Y^1_{(h_1)}$ and $Y_{i_0} = Y^0_{(h_0)}$.

5.2 Computation of the IPW Multiplier Bootstrap

We first provide more details on the sieve bases. Let $b(x) \equiv (b_1(x), \cdots, b_K(x))^\top$, where $\{b_k(\cdot)\}_{k=1}^K$ are K basis functions of a linear sieve space $\mathcal{B}$. Given that all d_x elements of X are continuously distributed, the sieve space $\mathcal{B}$ can be constructed as follows.

1. For each element $X^{(l)}$ of X, $l = 1, \cdots, d_x$, let $\mathcal{B}_l$ be the univariate sieve space of dimension J_n. One example of $\mathcal{B}_l$ is the linear span of the J_n dimensional polynomials given by

$$\mathcal{B}_l = \left\{ \sum_{k=0}^{J_n} \alpha_k x^k, x \in \text{Supp}(X^{(l)}), \alpha_k \in \Re \right\};$$

   Another is the linear span of r-order splines with J_n nodes given by

$$\mathcal{B}_l = \left\{ \sum_{k=0}^{r-1} \alpha_k x^k + \sum_{j=1}^{J_n} b_j [\max(x - t_j, 0)]^{r-1}, x \in \text{Supp}(X^{(l)}), \alpha_k, b_j \in \Re \right\},$$

   where the grid $-\infty = t_0 \leq t_1 \leq \cdots \leq t_{J_n} \leq t_{J_n+1} = \infty$ partitions $\text{Supp}(X^{(l)})$ into $J_n + 1$ subsets $I_j = [t_j, t_{j+1}) \cap \text{Supp}(X^{(l)})$, $j = 1, \cdots, J_n - 1$, $I_0 = (t_0, t_1) \cap \text{Supp}(X^{(l)})$, and $I_{J_n} = (t_{J_n}, t_{J_n+1}) \cap \text{Supp}(X^{(l)})$.

2. Let $\mathcal{B}$ be the tensor product of $\{\mathcal{B}_l\}_{l=1}^{d_x}$, which is defined as a linear space spanned by the functions $\prod_{l=1}^{d_x} g_l$, where $g_l \in \mathcal{B}_l$. The dimension of $\mathcal{B}$ is then $K \equiv J_n^{d_x}$.

In practice, we suggest not using all the tensor products as otherwise the dimension $J_n^{d_x}$ can be too large for moderate sample size. Instead, with the number of pairs equals 50 or 100, we suggest using splines with one node (usually the median) for each dimension and one interaction term across each pair of dimensions. In the appendix, we also propose a cross-validation method to select the basis functions.

Given the sieve bases, we can estimate the propensity score following (4.7). We then obtain $\hat{q}^w_{ipw,1}(\tau)$ and $\hat{q}^w_{ipw,0}(\tau)$ by solving the sub-gradient conditions for the two optimizations

in (4.8). Specifically, we have $\hat{q}^w_{ipw,1}(\tau) = Y_{h'_1}$ and $\hat{q}^w_{ipw,0}(\tau) = Y_{h'_0}$, where the indexes h'_0 and h'_1 satisfy $A_{h'_a} = a$, $a = 0, 1$,

$$\tau\left(\sum_{i=1}^{2n}\frac{\xi_i A_i}{\hat{A}_i}\right) - \frac{\xi_{h'_1}}{\hat{A}_{h'_1}} \leq \sum_{i=1}^{2n}\frac{\xi_i A_i}{\hat{A}_i}1\{Y_i < Y_{h'_1}\} \leq \tau\left(\sum_{i=1}^{2n}\frac{\xi_i A_i}{\hat{A}_i}\right), \tag{5.2}$$

and

$$\tau\left(\sum_{i=1}^{2n}\frac{\xi_i (1-A_i)}{1-\hat{A}_i}\right) - \frac{\xi_{h'_0}}{1-\hat{A}_{h'_0}} \leq \sum_{i=1}^{2n}\frac{\xi_i (1-A_i)}{1-\hat{A}_i}1\{Y_i < Y_{h'_0}\} \leq \tau\left(\sum_{i=1}^{2n}\frac{\xi_i (1-A_i)}{1-\hat{A}_i}\right). \tag{5.3}$$

In practical implementation we set $\{\xi_i\}_{i\in[2n]}$ as i.i.d. standard exponential random variables. In this case, all the equalities in (5.2) and (5.3) hold with probability zero. Thus, h'_1 and h'_0 are uniquely defined with probability one.

The IPW multiplier bootstrap can also be used to infer the ATE when the pairs identities are unknown. The point estimator of ATE under MPD is just the difference in mean estimator: $\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{2n}(A_i Y_i - (1-A_i)Y_i)$. In order to mimic its limit distribution, we propose to an IPW multiplier bootstrap for the ATE estimator

$$\hat{\Delta}^w_{ipw} = \frac{1}{\sum_{i=1}^{2n}\xi_i A_i/\hat{A}_i}\sum_{i=1}^{2n}\frac{\xi_i A_i Y_i}{\hat{A}_i} - \frac{1}{\sum_{i=1}^{2n}\xi_i (1-A_i)/(1-\hat{A}_i)}\sum_{i=1}^{2n}\frac{\xi_i (1-A_i) Y_i}{1-\hat{A}_i}.$$

We summarize the bootstrap procedure as follows.

1. Compute the original estimator $\hat{q}(\tau) = \hat{q}_1(\tau) - \hat{q}_0(\tau)$ and $\hat{\Delta}$.

2. Let B be the number of bootstrap replications. Let $\mathcal{G}$ be a grid of quantile indexes. For $b \in [B]$, generate $\{\xi_i\}_{i\in[2n]}$ as a sequence of i.i.d. exponential random variables. Estimate the propensity score following (4.7). Compute $\hat{q}^{w,b}_{ipw}(\tau) = Y_{h'_1} - Y_{h'_0}$ for $\tau \in \mathcal{G}$, where h'_0 and h'_1 are computed as in (5.2) and (5.3), respectively, and $\hat{\Delta}^{w,b}_{ipw}$.

3. Repeat the above step for $b \in [B]$ and obtain B bootstrap estimators of the QTE and ATE, denoted as $\{\hat{q}^{w,b}_{ipw}(\tau)\}_{b\in[B],\tau\in\mathcal{G}}$, $\{\hat{\Delta}^{w,b}_{ipw}\}_{b\in[B]}$, respectively.

For comparison, we also consider the naive multiplier bootstrap and the naive multiplier bootstrap of the pairs in our simulations. The computation of the naive multiplier bootstrap follows a procedure similar to the above with only one difference: the nonparametric estimate $\hat{A}_i$ of the propensity score is replaced by the truth, that is, 1/2. The computation of the naive multiplier bootstrap of the pairs follows a similar procedure to that of the naive multiplier bootstrap except that the units in the same pair share the same multiplier.

5.3 Bootstrap Confidence Intervals

Given the bootstrap estimates, we discuss how to conduct bootstrap inference for the null hypotheses with single, multiple, and a continuum of quantile indexes. We take the gradient bootstrap as an example. If the IPW multiplier bootstrap is used, one can just replace $\{\hat{q}^{*b}(\tau)\}_{b\in[B],\tau\in\mathcal{G}}$ by $\{\hat{q}^{w,b}_{ipw}(\tau)\}_{b\in[B],\tau\in\mathcal{G}}$ in the following cases. The same procedure applies to the bootstrap inference of ATE as well.

Case (1). We aim to test the single null hypothesis that $\mathcal{H}_0 : q(\tau) = \underline{q}$ vs. $q(\tau) \neq \underline{q}$. Let $\mathcal{G} = \{\tau\}$ in the procedures described above. Further denote $\mathcal{Q}(\nu)$ as the νth empirical quantile of the sequence $\{\hat{q}^{*b}(\tau)\}_{b\in[B]}$. Let $\alpha \in (0,1)$ be the significance level. We suggest using the bootstrap estimator to construct the standard error of $\hat{q}(\tau)$ as $\hat{\sigma} = \frac{\mathcal{Q}(0.975)-\mathcal{Q}(0.025)}{C_{0.975}-C_{0.025}}$, where C_μ is the μth standard normal critical value. Then a valid confidence interval and Wald test using this standard error are

$$CI(\alpha) = (\hat{q}(\tau) - C_{1-\alpha/2}\hat{\sigma}, \hat{q}(\tau) + C_{1-\alpha/2}\hat{\sigma}),$$

and $1\{\left|\frac{\hat{q}(\tau)-\underline{q}}{\hat{\sigma}}\right| \geq C_{1-\alpha/2}\}$, respectively.[9]

Case (2). We aim to test the null hypothesis that $\mathcal{H}_0 : q(\tau_1)-q(\tau_2) = \underline{q}$ vs. $q(\tau_1)-q(\tau_2) \neq \underline{q}$. In this case, let $\mathcal{G} = \{\tau_1, \tau_2\}$. Further, let $\mathcal{Q}(\nu)$ denote the νth empirical quantile of the

[9]It is asymptotically valid to use standard and percentile bootstrap confidence intervals. In our simulations, we found that the confidence interval proposed in the paper has better finite-sample performance.

sequence $\{\hat{q}^{*b}(\tau_1)-\hat{q}^{*b}(\tau_2)\}_{b\in[B]}$, and let $\alpha \in (0,1)$ be the significance level. We suggest using the bootstrap standard error to construct a valid confidence interval and Wald test as

$$CI(\alpha) = (\hat{q}(\tau_1) - \hat{q}(\tau_2) - C_{1-\alpha/2}\hat{\sigma}, \hat{q}(\tau_1) - \hat{q}(\tau_2) + C_{1-\alpha/2}\hat{\sigma}),$$

and $1\{\left|\frac{\hat{q}(\tau_1)-\hat{q}(\tau_2)-\underline{q}}{\hat{\sigma}}\right| \geq C_{1-\alpha/2}\}$, respectively, where $\hat{\sigma} = \frac{\mathcal{Q}(0.975)-\mathcal{Q}(0.025)}{C_{0.975}-C_{0.025}}$.

Case (3). We aim to test the null hypothesis that

$$\mathcal{H}_0 : q(\tau) = \underline{q}(\tau) \ \forall \tau \in \Upsilon \text{ vs. } q(\tau) \neq \underline{q}(\tau) \ \exists \tau \in \Upsilon.$$

In theory, we should let $\mathcal{G} = \Upsilon$. In practice, we let $\mathcal{G} = \{\tau_1, \cdots, \tau_G\}$ be a fine grid of Υ where G should be as large as computationally possible. Further, let $\mathcal{Q}_\tau(\nu)$ denote the νth empirical quantile of the sequence $\{\hat{q}^{*b}(\tau)\}_{b\in[B]}$ for $\tau \in \mathcal{G}$. Compute the standard error of $\hat{q}(\tau)$ as

$$\hat{\sigma}_\tau = \frac{\mathcal{Q}_\tau(0.975) - \mathcal{Q}_\tau(0.025)}{C_{0.975} - C_{0.025}}.$$

The uniform confidence band with an α significance level is constructed as

$$CB(\alpha) = \{\hat{q}(\tau) - \mathcal{C}_\alpha\hat{\sigma}_\tau, \hat{q}(\tau) + \mathcal{C}_\alpha\hat{\sigma}_\tau : \tau \in \mathcal{G}\},$$

where the critical value $\mathcal{C}_\alpha$ is computed as

$$\mathcal{C}_\alpha = \inf\left\{z : \frac{1}{B}\sum_{b=1}^{B} 1\left\{\sup_{\tau\in\mathcal{G}}\left|\frac{\hat{q}^{*b}(\tau) - \tilde{q}(\tau)}{\hat{\sigma}_\tau}\right| \leq z\right\} \geq 1-\alpha\right\}$$

and $\tilde{q}(\tau)$ is first-order equivalent to $\hat{q}(\tau)$ in the sense that $\sup_{\tau\in\Upsilon}|\tilde{q}(\tau)-\hat{q}(\tau)| = o_p(1/\sqrt{n})$. We suggest choosing $\tilde{q}(\tau) = \frac{1}{2}\{\mathcal{Q}_\tau(0.975) + \mathcal{Q}_\tau(0.025)\}$ over other choices such as $\tilde{q}(\tau) = \mathcal{Q}_\tau(0.5)$ and $\tilde{q}(\tau) = \hat{q}(\tau)$ due to its better finite-sample performance. We reject $\mathcal{H}_0$ at an α significance

level if $\underline{q}(\cdot) \notin CB(\alpha)$.

5.4 Practical Recommendations

Our practical recommendations are straightforward. If pair identities are known, we suggest using the gradient bootstrap for inference. Otherwise, we suggest using the IPW multiplier bootstrap with a nonparametrically estimated propensity score for inference.

6 Simulation

In this section, we assess the finite-sample performance of the methods discussed in Section 4 with a Monte Carlo simulation study. In all cases, potential outcomes for $a \in \{0,1\}$ and $1 \leq i \leq 2n$ are generated as

$$Y_i(a) = \mu_a + m_a(X_i) + \sigma_a(X_i)\,\varepsilon_{a,i},\ a = 0,1, \tag{6.1}$$

where $\mu_a, m_a(X_i), \sigma_a(X_i)$, and $\varepsilon_{a,i}$ are specified as follows. In each of the specifications below, $n \in \{50, 100\}$ and $(X_i, \varepsilon_{0,i}, \varepsilon_{1,i})$ are i.i.d. The number of replications is 10,000. For bootstrap replications we set $B = 5,000$.

Model 1 $X_i \sim \text{Unif}[0,1]$; $m_0(X_i) = 0$; $m_1(X_i) = 10\left(X_i^2 - \frac{1}{3}\right)$; $\varepsilon_{a,i} \sim N(0,1)$ for $a = 0,1$; $\sigma_0(X_i) = \sigma_0 = 1$ and $\sigma_1(X_i) = \sigma_1$.

Model 2 As in Model 1, but with $\sigma_0(X_i) = (1 + X_i^2)$ and $\sigma_1(X_i) = (1 + X_i^2)\,\sigma_1$.

Model 3 $X_i = (\Phi(V_{i1}), \Phi(V_{i2}))^\top$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and

$$V_i \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

$m_0(X_i) = \gamma^\top X_i - 1$; $m_1(X_i) = m_0(X_i) + 10(\Phi^{-1}(X_{i1})\Phi^{-1}(X_{i2}) - \rho)$; $\varepsilon_{a,i} \sim N(0,1)$ for $a = 0, 1$; $\sigma_0(X_i) = \sigma_0 = 1$ and $\sigma_1(X_i) = \sigma_1$. We set $\gamma = (1,1)^\top$, $\sigma_1 = 1$, $\rho = 0.2$.

Model 4 As in Model 3, but with $\gamma = (1,4)^\top$, $\sigma_1 = 2$, $\rho = 0.7$.

Pairs are determined similarly to those in Bai et al. (2021). Specifically, if X_i is a scalar, then pairs are determined by sorting $\{X_i\}_{i\in[2n]}$. If X_i is multi-dimensional, then the pairs are determined by the permutation π computed using the *R* package *nbpMatching*. We refer interested readers to Bai et al. (2021, Section 4) for more detail. After forming the pairs, we assign treatment status within each pair through a random draw from the uniform distribution over $\{(0,1),(1,0)\}$.

We examine the performance of various tests for ATEs and QTEs at the nominal level $\alpha = 5\%$. For the ATE, we consider the hypothesis that

$$\mathbb{E}(Y(1) - Y(0)) = \text{truth} + \Delta \quad vs. \quad \mathbb{E}(Y(1) - Y(0)) \neq \text{truth} + \Delta.$$

For the QTE, we consider the hypotheses that

$$q(\tau) = \text{truth} + \Delta \quad vs. \quad q(\tau) \neq \text{truth} + \Delta,$$

for $\tau = 0.25$, 0.5, and 0.75,

$$q(0.25) - q(0.75) = \text{truth} + \Delta \quad vs. \quad q(0.25) - q(0.75) \neq \text{truth} + \Delta, \tag{6.2}$$

and

$$q(\tau) = \text{truth} + \Delta \quad \forall\tau \in [0.25, 0.75] \quad vs. \quad q(\tau) \neq \text{truth} + \Delta \quad \exists\tau \in [0.25, 0.75]. \tag{6.3}$$

To illustrate size and power of the tests, we set $\mathcal{H}_0 : \Delta = 0$ and $\mathcal{H}_1 : \Delta = 1/2$. The true value for the ATE is 0, whereas the true values for the QTEs are simulated with a $10,000$ sample

size and replications. The computational procedures described in Section 5 are followed to perform the bootstrap and calculate the test statistics. To test the single null hypothesis involving one or two quantile indexes, we use the Wald tests specified in Section 5.3. To test the null hypothesis involving a continuum of quantile indexes, we use the uniform confidence band $CB(\alpha)$ defined in Case (3) in the same section.

Table 2: The Empirical Size and Power of Tests for ATEs

Model	$\mathcal{H}_0$: $\Delta = 0$							
	$n = 50$				$n = 100$			
	Naive	Naive Pair	Adj	IPW	Naive	Naive Pair	Adj	IPW
1	1.32	1.52	5.47	5.44	1.22	1.34	5.75	6.00
2	1.85	2.22	5.35	5.59	1.64	1.83	5.63	5.89
3	1.20	1.48	4.76	4.92	0.77	0.89	4.68	5.16
4	2.32	2.72	6.47	6.01	1.25	1.33	5.33	4.74
Model	$\mathcal{H}_1$: $\Delta = 1/2$							
	$n = 50$				$n = 100$			
	Naive	Naive Pair	Adj	IPW	Naive	Naive Pair	Adj	IPW
1	11.80	13.47	29.10	29.44	27.67	28.57	49.79	50.46
2	10.43	12.06	23.26	24.24	23.72	24.99	40.42	41.68
3	1.31	1.73	5.66	5.91	1.92	2.07	8.13	8.74
4	1.08	1.45	5.16	4.35	0.93	1.05	5.65	4.89

Notes: The table presents the rejection probabilities for tests of ATEs. The columns 'Naive' and 'Adj' correspond to the two-sample t-test and the adjusted t-test in Bai et al. (2021), respectively; the 'Naive pair' column corresponds to the t-test using the standard errors estimated by the naive multiplier bootstrap of the pairs; the column 'IPW' corresponds to the t-test using the standard errors estimated by the IPW multiplier bootstrap ATE estimator.

The results for the ATEs appear in Table 2. Each row presents a different model and each column reports the rejection probabilities for the various methods. The column 'Naive' refers to the two-sample t-test and 'Adj' refers to the adjusted t-test in Bai et al. (2021); the column 'Naive pair' corresponds to the t-test using the standard errors estimated by the naive multiplier bootstrap of the pairs; the column 'IPW' corresponds to the t-test using the standard errors estimated by the IPW multiplier bootstrap ATE estimator.

We make several observations on these findings. First, the two-sample t-test has rejection probability under $\mathcal{H}_0$ far below the nominal level and is the least powerful test among the four. Second, the adjusted t-test has rejection probability under $\mathcal{H}_0$ close to the nominal level and is not conservative. This result is consistent with those in Bai et al. (2021). Third, the t-test using the standard error estimated by bootstrapping the pairs of units alone is conservative under the null and lacks power under the alternative. Fourth, the IPW t-test proposed in this paper has performance similar to the adjusted t-test.[10] Under $\mathcal{H}_0$, the test has rejection probability close to 5%; under $\mathcal{H}_1$, it is more powerful than the 'naive' and 'naive pair' methods and has power similar to the adjusted t-test. These findings indicate that the IPW t-test provides an alternative to the adjusted t-test when pair identities are unknown.

The results for QTEs are summarized in Tables 3 and 4. Each table has four panels (Models 1-4). Each row in the panel displays the rejection probabilities for the tests using the standard errors estimated by various bootstrap methods. Specifically, the rows 'Naive', 'Naive pair', 'Gradient', and 'IPW' respectively correspond to the results of the naive multiplier bootstrap, the naive multiplier bootstrap of the pairs, the gradient bootstrap, and the IPW multiplier bootstrap.

Table 3 reports the empirical size and power of the tests with a single null hypothesis involving one or two quantile indexes. Columns '0.25', '0.50', and '0.75' correspond to tests with quantiles at 25%, 50%, and 75%. Column 'Dif' corresponds to the test with null hypothesis (6.2). As expected given Theorem 4.1, the test with standard errors estimated by two naive methods performs poorly in all cases. It is conservative under $\mathcal{H}_0$ and lacks power

[10] Throughout this section, for both ATE and QTE estimation, we use splines to nonparametrically estimate the propensity score in the IPW multiplier bootstrap. If $\dim(X_i)$=1, we choose the bases $\{1, X, X^2, [\max(X - qx_{0.5}, 0)]^2\}$ where $qx_{0.5}$ is the quantile of X at 50%; if $\dim(X_i)$=2, we choose the bases $\{1, X_1, X_2, \max(X_1 - qx_{1,0.5}, 0), \max(X_2 - qx_{2,0.5}, 0), X_1X_2\}$, where for $j = 1, 2$, $qx_{j,\alpha}$ is the αth sample percentile of X_j. Results are similar using the cross-validation method proposed in the appendix to choose the sieve bases: for details on the cross-validation method and its simulation results, see Section G in the supplement.

under $\mathcal{H}_1$. In contrast, the test using the standard errors estimated by either the gradient bootstrap or the IPW multiplier bootstrap method has a rejection probability under $\mathcal{H}_0$ that is close to the nominal level in almost all specifications. When the number of pairs is 50, the tests in the 'Dif' column constructed based on either the gradient or the IPW multiplier bootstrap method are slightly conservative. Sizes approach the nominal level when n increases to 100.

Table 4 reports empirical size and power of the uniform confidence bands for the hypothesis specified in (6.3) with a grid $\mathcal{G} = \{0.25, 0.27, \cdots, 0.47, 0.49, 0.5, 0.51, 0.53, \cdots, 0.73, 0.75\}$. The test using standard errors estimated by two naive methods has rejection probabilities under $\mathcal{H}_0$ far below the nominal level in all specifications. In Models 1-2, the test using standard errors estimated by either the gradient bootstrap or the IPW multiplier bootstrap yields a rejection probability under $\mathcal{H}_0$ that is very close to the nominal level even when the number of pairs is as small as 50. Nonetheless, in Models 3-4, the tests constructed based on both methods are conservative when the number of pairs equals 50. When the number of pairs increases to 100, both tests perform much better and have rejection probabilities under $\mathcal{H}_0$ that are close to the nominal level. Under $\mathcal{H}_1$, the tests based on both the gradient and IPW methods are more powerful than those based on the naive methods.

In summary, the simulation results in Tables 3 and 4 are consistent with the results in Theorems 4.3 and 4.4: both the gradient bootstrap and the IPW multiplier bootstrap provide valid pointwise and uniform inference for QTEs under MPDs. The findings also show that when the information on pair identities is unavailable, the IPW multiplier bootstrap continues to provide a sound basis for inference.

Table 3: The Empirical Size and Power of Tests for QTEs

	$\mathcal{H}_0$: $\Delta = 0$								$\mathcal{H}_1$: $\Delta = 1/2$							
	$n = 50$				$n = 100$				$n = 50$				$n = 100$			
	0.25	0.50	0.75	Dif	0.25	0.50	0.75	Dif	0.25	0.50	0.75	Dif	0.25	0.50	0.75	Dif
Model 1																
Naive	3.00	2.00	2.22	1.98	3.12	2.06	1.93	1.73	16.67	6.05	5.56	3.96	34.93	11.56	8.11	7.35
Naive pair	2.99	2.29	2.30	2.07	3.37	2.23	2.08	1.81	16.85	6.75	5.64	4.08	35.33	11.41	8.24	7.59
Gradient	5.13	4.82	4.92	3.66	5.07	5.62	5.30	4.04	23.76	13.03	11.27	8.18	42.92	22.91	17.30	14.57
IPW	5.47	5.31	6.17	4.24	5.26	5.83	5.65	3.95	24.81	13.48	12.12	8.40	43.93	23.33	17.21	13.91
Model 2																
Naive	3.08	2.32	2.55	1.96	3.64	2.53	2.08	1.87	14.82	6.54	4.71	3.68	30.29	11.50	7.46	6.88
Naive pair	3.05	2.50	2.65	2.03	3.87	2.77	2.23	1.95	14.67	6.81	4.96	3.65	30.97	11.80	7.85	7.27
Gradient	4.57	4.63	4.39	3.44	5.00	5.42	5.28	3.68	19.51	12.25	8.76	6.57	35.38	20.86	14.79	12.25
IPW	4.93	5.12	5.78	4.45	5.17	5.73	5.88	4.00	20.29	12.90	10.40	7.35	36.38	21.53	15.14	12.53
Model 3																
Naive	2.11	1.03	2.10	0.92	1.56	1.37	1.58	0.86	4.98	2.85	1.92	0.98	6.57	7.14	1.73	1.43
Naive pair	2.21	1.18	2.31	1.13	1.57	1.42	1.55	0.96	4.99	3.29	2.14	1.11	6.80	7.56	1.88	1.44
Gradient	5.24	3.06	3.14	1.76	4.83	4.20	4.27	3.01	9.71	7.43	3.22	2.39	13.80	16.72	5.67	4.40
IPW	4.76	3.19	5.61	2.60	4.77	3.71	4.95	3.02	8.75	7.81	5.35	3.09	13.04	15.42	6.06	4.21
Model 4																
Naive	2.59	1.71	1.98	1.65	2.65	1.66	1.55	1.23	6.09	1.94	1.76	1.28	9.85	2.98	1.19	1.18
Naive pair	2.90	1.79	2.04	1.74	2.74	1.71	1.71	1.35	6.24	2.23	1.93	1.45	10.25	3.17	1.35	1.23
Gradient	4.75	4.00	3.33	2.82	4.70	4.74	5.06	3.88	9.37	5.76	3.35	2.87	14.67	8.88	5.27	4.25
IPW	3.97	3.97	4.91	3.68	4.23	4.51	5.01	3.48	8.08	5.37	4.79	3.26	13.50	8.33	5.17	3.51

Note: The table presents the rejection probabilities for tests of QTEs. The columns '0.25', '0.50', and '0.75' correspond to tests with quantiles at 25%, 50%, and 75%, respectively; the column 'Dif' corresponds to the test with the null hypothesis specified in (6.2). The rows 'Naive', 'Naive pair', 'Gradient', and 'IPW' correspond to the results of the naive multiplier bootstrap, the naive multiplier bootstrap of the pairs, the gradient bootstrap, and IPW multiplier bootstrap, respectively.

Table 4: The Empirical Size and Power of Uniform Inferences for QTEs

	$\mathcal{H}_0$: $\Delta = 0$		$\mathcal{H}_1$: $\Delta = 1/2$	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$
Model 1				
Naive	1.07	1.52	7.50	18.12
Naive pair	1.32	1.63	7.10	18.52
Gradient	4.08	4.64	17.88	33.30
IPW	4.49	4.94	16.30	32.40
Model 2				
Naive	1.37	1.85	6.73	16.50
Naive pair	1.39	1.91	6.63	17.04
Gradient	3.66	4.57	14.30	27.64
IPW	4.25	4.91	14.27	27.47
Model 3				
Naive	0.63	0.63	1.43	3.50
Naive pair	0.60	0.69	1.54	4.02
Gradient	1.90	3.07	5.19	13.33
IPW	2.19	2.99	4.25	11.34
Model 4				
Naive	0.99	1.00	1.40	3.05
Naive pair	0.97	1.00	1.33	3.28
Gradient	2.87	3.72	4.47	8.57
IPW	2.78	3.36	3.18	6.98

Notes: The table presents the rejection probabilities of the uniform confidence bands for the hypothesis specified in (6.3). The rows 'Naive', 'Naive pair', 'Gradient', and 'IPW' correspond to the results of the naive multiplier bootstrap, the naive multiplier bootstrap of the pairs, the gradient bootstrap, and the IPW multiplier bootstrap, respectively.

7 Empirical Application

Policy and macroeconomic uncertainty are considered to be two major constraints to firm growth in developing countries (Bloom, 2014; Bloom, Floetotto, Jaimovich, Saporta-Eksten, and Terry, 2018; World Bank, 2004). Groh and McKenzie (2016) conducted a randomized experiment with a MPD to explore the treatment effect of providing insurance against

macroeconomic and policy shocks to microenterprise owners. In this section, we apply the bootstrap methods developed in this paper to their data and examine both the ATEs and QTEs of macroinsurance on business owners' monthly consumption and their firms' monthly profits.[11]

Table 5: Summary Statistics

	Total	Treatment group	Control group
Outcome variables			
Consumption	1946.9(903.0)	1946.1(900.7)	1947.7(905.7)
Profit	1342.4(1470.7)	1299.9(1444.3)	1384.9(1496.1)
Matching variables			
Owner is female	0.36(0.48)	0.36(0.48)	0.36(0.48)
Expected likelihood of a macro shock	56.5(32.7)	56.1(33.0)	56.8(32.4)
Higher risk aversion	0.48(0.50)	0.47(0.50)	0.49(0.50)
Owner is ambiguity neutral	0.29(0.46)	0.29(0.45)	0.30(0.46)
Sales drop 20% more	0.41(0.49)	0.41(0.49)	0.41(0.49)
Sales drop between 5% and 20%	0.29(0.45)	0.29(0.45)	0.29(0.45)
Considering delaying investments	0.10(0.30)	0.10(0.30)	0.10(0.29)
Expect to renew their loan	0.89(0.31)	0.90(0.31)	0.89(0.31)
Expect to renew a loan of 3000 LE or less	0.27(0.45)	0.28(0.45)	0.27(0.44)
Expect to renew a loan of 3001 to 5000 LE	0.27(0.44)	0.27(0.44)	0.26(0.44)
Profits in Feb 2012	1085.4(1174.5)	1060.4(1149.0)	1110.4(1199.4)
Profits in Jan 2012	1054.0(1161.1)	1023.4(1114.5)	1084.6(1205.5)
Missing Feb 2012 profits	0.04(0.18)	0.04(0.19)	0.03(0.18)
Missing Jan 2012 profits	0.04(0.19)	0.04(0.20)	0.03(0.18)
Observations	2824	1412	1412

Notes: Unit of observation: business owners. The table presents the means and standard deviations (in parentheses) of two outcome variables and all the pair-matching variables.

The sample consists of 2824 business owners, who were the clients of Egypt's largest

[11]Data are available at https://microdata.worldbank.org/index.php/catalog/2063.

microfinance institution – Alexandria Business Association (ABA). After an exact match on gender and microfinance branch code within ABA, the business owners were grouped into pairs by using an optimal greedy algorithm to minimize the Mahalanobis distance between the values of additional 13 matching variables (See Groh and McKenzie, 2016 for the definitions of these 13 variables). This segmentation gives 1412 pairs in the sample; one business owner in each pair was randomly assigned to the treatment group and the other to the control group. In the treatment group, a macroinsurance product was offered. Groh and McKenzie (2016) then examined the impacts of the access to macroinsurance on various outcome variables.

Here we focus on the impacts of macroinsurance on two outcome variables: the business owners' monthly consumption and their firms' monthly profits. Table 5 gives descriptive statistics (means and standard deviations) of these two outcome variables as well as all the matching variables used by Groh and McKenzie (2016) to form the pairs in their experiments.[12]

Table 6: ATEs of Macroinsurance on Consumption and Profits

	Naive	Naive pair	Adj	IPW
Consumption	-1.59(33.98)	-1.59(29.71)	-1.59(29.45)	-1.59(29.45)
Profit	-86.68(56.09)	-86.68(48.41)	-86.68(49.38)	-86.68(45.38)

Notes: The table presents the ATE estimates of the effect of macroinsurance on the monthly consumption and profits. Standard errors are in the parentheses. The columns "Naive" and "Adj" correspond to the two-sample t-test and the adjusted t-test in Bai et al. (2021), respectively. The column 'Naive pair' corresponds to the t-test using standard errors estimated by the naive multiplier bootstrap of the pairs. The column "IPW" corresponds to the t-test using the standard errors estimated by the IPW multiplier bootstrap.

Table 6 reports the ATE estimators of macroinsurance on the consumption and profits with the standard errors (in parentheses) calculated by four methods. Specifically, the

[12]We filter out 137 unbalanced observations (less than 5% of the total observations in Groh and McKenzie, 2016) to keep a balanced data in both the pairs and the pairs of the pairs. The summary statistics in Table 5 are almost exactly the same as those in Table 1 of Groh and McKenzie (2016).

columns 'Naive' and 'Adj' correspond to the two-sample t-test and the adjusted t-test in Bai et al. (2021), respectively; the column 'Naive pair' corresponds to the t-test using standard errors estimated by the naive multiplier bootstrap of the pairs; the column 'IPW' corresponds to the t-test using standard errors estimated by IPW multiplier bootstrap.[13] The results lead to the following observations. First, consistent with the findings in Groh and McKenzie (2016), the naive two-sample t-tests show that expanding access to macroinsurance has no significant average effects on monthly consumption and profits. Second, the standard errors in the adjusted t-test are lower than those in the naive t-test, which is consistent with the finding in Bai et al. (2021). Compared to the standard errors estimated by the naive multiplier bootstrap of the pairs, they are modestly lower for the ATE estimates of macroinsurance on the consumption and slightly larger for those on the profits. More importantly, the standard errors estimated by the IPW multiplier bootstrap are lower than those estimated by two naive methods. These results corroborate our earlier finding that the IPW multiplier bootstrap is an alternative to the approach adopted in Bai et al. (2021), especially when the information on pair identities is unavailable. In addition, the results for the firms' profits also highlight the importance of accounting for the dependence structure within the pairs when estimating the standard errors of the ATE estimates. The ATE of macroinsurance on profits is statistically insignificant based on the naive t-test, but

[13]Throughout this section, to nonparametrically estimate the propensity score in the IPW multiplier bootstrap, we first standardize all the continuous matching variables to have mean zero and variance one. There are only three continuous matching variables; the rest of the matching variables are all dummy variables. We then conduct sieve estimation by choosing the bases $\{1, \max(X_1 - qx_{1,0.3}, 0), \max(X_1 - qx_{1,0.5}, 0), \max(X_2 - qx_{2,0.3}, 0), \max(X_2 - qx_{2,0.5}, 0), \max(X_3 - qx_{3,0.3}, 0), \max(X_3 - qx_{3,0.5}, 0), X_1X_2, X_2X_3, X_1X_3, DV\}$, where (X_1, X_2, X_3) denote three standardized continuous matching variables, $qx_{j,0.3}$ and $qx_{j,0.5}$ are 0.3 and 0.5th quantiles of X_j for $j = 1, 2, 3$, and DV denotes the dummy matching variables except for the variable "missing Feb 2012 profits" as it is collinear with the variable "missing Jan 2012 profits." Results reported in this section are similar to those when the sieve basis functions are selected via cross-validation. For more details on the cross-validation method and the related empirical results, see Section G in the supplement.

is significant at 10% significance level if adjusted or IPW t-test is used instead.

Table 7: QTEs of Macroinsurance on Consumption and Profits

	Naive	Naive pair	Gradient	IPW
Panel A. Consumption				
25%	-14.33(27.40)	-14.33(25.85)	-14.33(25.32)	-14.33(25.89)
50%	-4.50(34.12)	-4.50(32.23)	-4.50(32.50)	-4.50(31.80)
75%	-22.17(61.18)	-22.17(55.78)	-22.17(55.51)	-22.17(56.08)
Panel B. Profit				
25%	-33.33(29.76)	-33.33(29.76)	-33.33(29.12)	-33.33(25.94)
50%	-66.67(59.52)	-66.67(54.42)	-66.67(53.15)	-66.67(51.02)
75%	-200.00(99.92)	-200.00(89.29)	-200.00(93.54)	-200.00(85.04)

Notes: The table presents the QTE estimates of the effect of macroinsurance on the monthly consumption and profits at quantiles 25%, 50%, and 75%. Standard errors are in parentheses. The columns "Naive," "Naive pair," "Gradient," and "IPW" correspond to the results of the naive multiplier bootstrap, the naive multiplier bootstrap of the pairs, the gradient bootstrap, and the IPW multiplier bootstrap, respectively.

Table 7 presents the QTE estimates at quantile indexes 0.25, 0.5, and 0.75 with the standard errors (in parentheses) estimated by four different methods. Specifically, the columns "Naive," "Naive pair," "Gradient," and "IPW" correspond to the results of the naive multiplier bootstrap, the naive multiplier bootstrap of the pairs, the gradient bootstrap,[14] and the IPW multiplier bootstrap, respectively. These results lead to the following three observations.

First, consistent with the theoretical results in Section 4, all the standard errors estimated by the gradient bootstrap or the IPW multiplier bootstrap are lower than those estimated by the naive multiplier bootstrap. For example in Panel A, at the 75th percentile, compared with the naive multiplier bootstrap, the gradient and IPW multiplier bootstraps reduce the standard error by 9.3% and 8.2%, respectively.

Second, the standard errors estimated by the gradient or IPW multiplier bootstrap are

[14]Using the original pair identities and all the matching variables in Groh and McKenzie (2016), we can reorder the pairs according to the procedure described in Section 5.1. We thank David McKenzie for providing us the Stata code to implement the matching algorithm.

mostly lower than those estimated by the naive multiplier bootstrap of the pairs as well. The magnitude of reduction is modest, which may be because the two terms, $\frac{m_{1,\tau}(X_i)}{f_1(q_1(\tau))}$ and $\frac{m_{0,\tau}(X_i)}{f_0(q_0(\tau))}$ in (4.1), are almost equal in the current dataset, and thus, their effects on the standard error estimates are canceled.

Third, there is considerable heterogeneity in the effects of macroinsurance on the firm profits. Specifically, the magnitude of the treatment effects of macroinsurance rises as the quantile indexes increase. For example, in Panel B, the treatment effects on the monthly profits at the 25th percentile and the median are negative but not statistically significantly different from zero, whereas the effects at the 75th percentile are statistically significant. The magnitude of the treatment effect increases by over 100% from the 25th percentile to the median and by about 200% from the median to the 75th percentile. These findings may imply that expanding access to macroinsurance has small but negative effects on the firm profits in the lower tail of the distribution, and that these negative effects become stronger for upper-ranked microenterprises.

The third observation in Table 7 indicates that the heterogeneous effects of macroinsurance on the firms' profits are economically substantial. To assess whether these are statistically significant too, Table 8 reports statistical tests for the heterogeneity of the QTEs. Specifically, we test the null hypotheses that $q(0.50)-q(0.25)=0$, $q(0.75)-q(0.50)=0$, and $q(0.75)-q(0.25)=0$. We find that only the difference between the 75th and 25th QTEs in Panel B is statistically significant at the 10% significance level. This finding implies that the statistical evidence of heterogeneous treatment effects of macroinsurance on the firm profits is strong only in comparisons of the microenterprises between the lower and upper tails of the distribution.

Table 8: Tests for the Difference between Two QTEs of Macroinsurance

	Naive	Naive pair	Gradient	IPW
Panel A. Consumption				
50%-25%	9.83(30.34)	9.83(28.66)	9.83(29.00)	9.83(29.87)
75%-50%	-17.67(51.51)	-17.67(48.30)	-17.67(49.09)	-17.67(48.58)
75%-25%	-7.83(58.42)	-7.83(54.17)	-7.83(55.89)	-7.83(55.87)
Panel B. Profit				
50%-25%	-33.33(51.02)	-33.33(51.02)	-33.33(51.02)	-33.33(49.32)
75%-50%	-133.33(85.04)	-133.33(81.63)	-133.33(82.91)	-133.33(85.04)
75%-25%	-166.67(88.65)	-166.67(85.04)	-166.67(87.16)	-166.67(89.29)

Notes: The table presents tests for the difference between two QTEs of macroinsurance on the monthly consumption and profits. Standard errors are in parentheses. The columns "Naive," "Naive pair," "Gradient," and "IPW" correspond to the results of the naive multiplier bootstrap, the naive multiplier bootstrap of the pairs, the gradient bootstrap, and the IPW multiplier bootstrap, respectively.

8 Conclusion

This paper has studied estimation and inference of QTEs under MPDs and developed new bootstrap methods to improve statistical performance. Derivation of the limit distribution of QTE estimators under MPDs reveals that analytic methods of inference based on asymptotic theory requires estimation of two infinite-dimensional nuisance parameters for every quantile index of interest. A further limitation is that both the naive multiplier bootstrap and the naive multiplier bootstrap of the pairs fail to approximate the limit distribution of the QTE estimator as they do not preserve the dependence structure in the original sample. Instead, we propose a gradient bootstrap approach that can consistently approximate the limit distribution of the original estimator and is free of tuning parameters. Implementation of the gradient bootstrap requires knowledge of pair identities. So when such information is unavailable we propose an IPW multiplier bootstrap and show that it consistently approximates the limit distribution of the original QTE estimator. Simulations provide finite-sample evidence of these procedures that support the asymptotic findings. An empirical application of these bootstrap methods to the real dataset in Groh and McKenzie (2016) shows consid-

erable evidence of heterogeneity in the effects of macroinsurance on firm profits. In both the simulations and the empirical application, the two recommended bootstrap methods of inference perform well in the sense that they usually provide smaller standard errors and greater inferential accuracy than those obtained by naive bootstrap methods.

Two directions for future research are especially evident from the present findings. First, it would be interesting to study inference of the QTEs when data are independent but not identically distributed. Such an assumption is adopted in the linear quantile regression literature to address issues regarding data heteroskedasticity and clustering (see, for example, Chen, Wan, and Zhou, 2015 and Hagemann, 2017). Second, it would be useful to incorporate data-driven methods such as cross-validation to the selection of the sieve basis functions when implementing the IPW multiplier bootstrap and then develop a procedure for inference orthogonal to the model selection bias introduced by data-driven methods.

References

Abadie, A., M. M. Chingos, and M. R. West (2018). Endogenous stratification in randomized experiments. *The Review of Economics and Statistics 100*(4), 567–580.

Angrist, J. D. and V. Lavy (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *The American Economic Review 99*(4), 1384–1414.

Athey, S. and G. Imbens (2017). Chapter 3 - the econometrics of randomized experimentsa. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments*, Volume 1 of *Handbook of Economic Field Experiments*, pp. 73 – 140. North-Holland.

Bai, Y. (2019). Optimality of matched-pair designs in randomized controlled trials. *Available at SSRN 3483834*.

Bai, Y., A. Shaikh, and J. P. Romano (2021). Inference in experiments with matched pairs. *Journal of the American Statistical Association*, 1–37.

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015). The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics 7*(1), 22–53.

Beuermann, D. W., J. Cristia, S. Cueto, O. Malamud, and Y. Cruzaguayo (2015). One laptop per child at home: Short-term impacts from a randomized experiment in peru. *American Economic Journal: Applied Economics 7*(2), 53–80.

Bloom, N. (2014). Fluctuations in uncertainty. *Journal of Economic Perspectives 28*(2), 153–176.

Bloom, N., M. Floetotto, N. Jaimovich, I. Saporta-Eksten, and S. J. Terry (2018). Really Uncertain Business Cycles. *Econometrica 86*(3), 1031–1065.

Bold, T., M. S. Kimenyi, G. Mwabu, A. Nganga, and J. Sandefur (2018). Experimental evidence on scaling up education reforms in kenya. *Journal of Public Economics 168*(12), 1–20.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1741–1768.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics 10*(4), 1747–1785.

Butler, D. (2010). Monitoring Bureaucratic Compliance: Using Field Experiments to Improve Governance. *Public Sector Digest*, 41–45.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics 6*, 5549–5632.

Chen, X., A. T. K. Wan, and Y. Zhou (2015). Efficient Quantile Regression Analysis With Missing Observations. *Journal of the American Statistical Association 110*(510), 723–741.

Crepon, B., F. Devoto, E. Duflo, and W. Pariente (2015). Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco. *American Economic Journal: Applied Economics 7*(1), 123–150.

Donald, S. G. and Y.-C. Hsu (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics 178*(3), 383–397.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), 259–276.

Fryer, R. G. (2017). Management and student achievement: Evidence from a randomized field experiment. *National Bureau of Economic Research*.

Fryer, R. G. (2018). The 'pupil' factory: Specialization and the production of human capital in schools. *The American Economic Review 108*(3), 616–656.

Fryer, R. G., T. Devi, and R. Holden (2017). Vertical versus horizontal incentives in education: Evidence from randomized trials. *National Bureau of Economic Research*.

Glewwe, P., A. F. Park, and M. Zhao (2016). A better vision for development: Eyeglasses and academic performance in rural primary schools in china. *Journal of Development Economics 122*(9), 170–182.

Groh, M. and D. J. McKenzie (2016). Macroinsurance for microenterprises: A randomized experiment in post-revolution egypt. *Journal of Development Economics 118*(1), 1–38.

Hagemann, A. (2017). Cluster-robust bootstrap inference in quantile regression models. *Journal of the American Statistical Association 112*(517), 446–456.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–331.

Hahn, J., K. Hirano, and D. Karlan (2011). Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics 29*(1), 96–108.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Panagopoulos, C. and D. P. Green (2008). Field experiments testing the impact of radio advertisements on electoral competition. *American Journal of Political Science 52*(1), 156–168.

Tabord-Meehan, M. (2018). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv preprint arXiv:1806.05127*.

van der Vaart, A. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

World Bank (2004). *World Development Report 2005: A Better Investment Climate for Everyone*. The World Bank.

Zhang, Y. and X. Zheng (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics 11*(3), 957–982.